The University Of Sheffield.

Ecological epigenetics in *Timema cristinae* stick insects: On the patterns, mechanisms and ecological consequences of DNA methylation in the wild

Clarissa Ferreira de Carvalho

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Science
Department of Animal and Plant Sciences

Submission Date
April 2019

Abstract

Epigenetic factors can contribute to phenotypic diversity and to ecological processes. For instance, DNA methylation can influence gene regulation, and thus phenotypic plasticity. However, little is yet known about how and why methylation varies in the wild. In this dissertation, I build on this knowledge by combining ecological, genetic and DNA methylation data from natural and experimental populations of the stick insect *Timema cristinae*. This species is an important system to ecological genetics studies, which provides good starting point for the investigation of the patterns, drivers, and the possible ecological consequences of natural methylation variation. I obtained methylation data using whole-genome bisulfite sequencing (BS-seq) and genetic data from restriction site associated DNA sequencing (RAD-seq). From a population survey, I found natural methylation variation in *T. cristinae* (1) is characteristic of "Hemimetabola" insects; (2) is structured in geographical space; and (3) is strongly correlated to genetic variation. In addition, an experiment simulating a host shift was carried out to test for the direct effects of host plant species on *T. cristinae* methylation levels. In both the population survey and in the experiment, binomial mixed models were used to perform a methylome scan in search of candidate single methylation polymorphisms (SMPs) associated with host plant use. This analysis is analogous to genome-wide analysis studies, but applied to methylation levels. They use genetic data to estimate random effects arising from relatedness. The results suggest (4) an association between methylation levels and host plant in specific regions and that (5) some of them could be responsive to host shift treatment. Finally, the model suggested (6) significant mean heritability of methylation status, estimated based on the genetic relatedness. My results collectively indicate methylation variation could be ecologically relevant to *T. cristinae*, and adds to the general understanding of the importance of epigenetic variation.

Acknowledgements

First and foremost, I am enormously grateful to my supervisor Patrik Nosil for having trusted in me to do this work and for all the valuable lessons throughout all of it. I have appreciated every single piece of advice he has given me during my PhD. I am also very thankful to my other supervisor Jon Slate, for all his guidance and for always encouraging me. To Royal Society for having granted me this opportunity, and to Mike Siva-Jothy, for having assisted me throughout all my PhD. I am very grateful to my colleagues and friends who have always helped me and gave me strength to keep going. My special thanks to Víctor, who has been a mentor to me and has always given me great support. To Romain, for his partnership and kindred spirit to help me throughout my PhD challenges. To my amazing colleagues, Anamaria, Emma, Gavin, James, Jill, Juan, Kay, Luke, Marisol, Matheus B., Mel, Nicola, Óscar, Pascal-Antoine, Rachel, Roger, Sean, Toni, not only for all the support and fruitful discussions, but for the friendship that we developed with it. My huge gratitude to my friends, for all their love and support. To Anaclara, Céline, Cláudio, Harriet, Mariana, Martha, Matheus A., Richard, Rowan, Yichen, for making my life in Sheffield a real pleasure. To my beautiful friends overseas, Drielle, Felipe, Heloísa, Mário, Rafaela, Rillen, Rodrigo, Vinícius, for always being with me. To Frane, whose passion and enthusiasm has inspired me to give my best. To Angela and Zuzanna, for the lovely friendship that reenergized me every time we met. Finally, I would not have reached this far without the support of my family. I am extremely thankful to my mother and to my father, for all their support and for sparking my curiosity and passion for the natural world, and to my siblings, Carolina and Henrique, my everlasting companions. I could not have done any of that without you.

Table of contents

Abstract III

Acknowledgements IV

Table of contents V

Chapter 1: General introduction 1

1.1. DNA methylation: the most investigated epigenetic mechanism 2

1.2. Ecological epigenetics 5

1.3. Study system: Timema cristinae stick insects 8

1.4. An ecological epigenetics study in T. cristinae 12

1.5. Outline of thesis chapters 16

1.6. A note on contributions made to this thesis 19

1.7. Note on contribution to appendix D 19

Chapter 2: Function and evolution of DNA methylation in *Timema cristinae* stick insects 23

2.1. Summary 23

2.2. Introduction 24

2.3. Material and Methods 27

2.3.1. DNA methyltransferases (DNMTs) 27

2.3.2. Sampling 29

2.3.3. Generating DNA methylation data 30

2.3.4. Annotation 38

2.3.5. Methylation enrichment on genomic features 39

2.3.6. Gene Ontology (GO) enrichments 40

2.3.7. Transposable elements (TEs) 40

2.4. Results 41

2.4.1. Identification of T. cristinae DNA methyltransferases 41

2.4.2. General patterns 43

2.4.3. Distribution of DNA methylation across genome 44

2.4.4. GO terms enriched in methylated and in non-methylated genes 49

2.4.5. Transposable elements 49

2.5. Discussion 51

2.5.1. Two copies of DNMT1 and absence of DNMT3 in T. cristinae 51

2.5.2. Majority of methylated cytosines is in CpG context 52
2.5.3. DNA methylation levels are high in T. cristinae genome and differentially distributed 53
2.5.4. Enriched GO terms are generally similar to those in other insects 55
2.5.5. TEs are normally depleted in methylation 56
2.6. Conclusion **57**
Appendix A: Supplementary Tables and Figures – Chapter 2 **58**
Chapter 3: Patterns and drivers of DNA methylation variation in natural populations of *Timema cristinae* stick insects **75**
3.1. Summary **75**
3.2. Introduction **76**
3.3. Materials and Methods **82**
3.3.1. Study system 82
3.3.2. Sampling design 84
3.3.3. Sampling 90
3.3.4. DNA methylation variation 90
3.3.5. Genetic variation 91
3.3.6. General patterns of geographical structure 94
3.3.7. Binomial Mixed Models 98
3.4. Results **101**
3.4.1. General patterns of methylation in natural populations 101
3.4.2. Heritability of methylation variation 104
3.5. Discussion **106**
3.5.1. DNA methylation is structured in geographical space 106
3.5.2. DNA methylation is strongly associated with genetic background 107
3.5.3. Genome-wide methylation variation is not strongly associated with climate or host plant 109
3.5.4. There is some heritability of DNA methylation patterns 110
3.6. Conclusion **111**
Appendix B: Supplementary Tables and Figures – Chapter 3 **112**
Chapter 4: Differential DNA methylation patterns and host plant use in *Timema cristinae* stick insects **121**
4.1. Summary **121**
4.2. Introduction **122**
4.3. Material and Methods **127**

4.3.1. Study system 127
4.3.2. Sampling 127
4.3.3. Rearing experiment 128
4.3.4. DNA methylation variation 129
4.3.5. Genetic variation 130
4.3.6. Clustering analyses 131
4.3.7. Methylome scan: binomial mixed models 131
4.3.8. Annotation 133
4.3.9. Transcriptome 134
4.4. Results **134**
4.4.1. Clustering analyses 134
4.4.2. Methylome scans 135
4.5. Discussion **140**
4.5.1. Association between methylation variation and host plant 141
4.5.2. Insect allergen gene and ecological context 143
4.5.3. Evolution of insect major allergen genes 146
4.6. Conclusion **148**
Appendix C: Supplementary Tables and Figures – Chapter 4 **150**
Chapter 5: Conclusions and future directions **159**
5.1. General discussion **159**
5.2. Future perspectives in ecological studies in DNA methylation **170**
Appendix D: Ecology helps explain whether genes for cryptic coloration form a supergene or recombine (unpublished manuscript) **175**
References **231**

Chapter 1

General introduction

In his seminal book, *The Origin of Species*, Darwin described the perfect fit between organisms and their environment (Darwin, 1859). When the theory of natural selection and the struggle for life was developed, he did not have an idea of how variation is inherited. With the advent of genetics and the rediscovery of Mendelian laws, biologists identified the genes as a heritable basis of phenotype (Morgan, 1915). Following this, geneticists and statisticians built the foundation for much of the research in evolutionary biology today (*i.e.* the Modern Synthesis), establishing rigorous quantitative methods to understand adaptation and evolutionary processes as changes in gene frequencies in populations over time (Fisher, 1930; Dobzhansky, 1937; Waddington, 1939). Since then, the majority of evolutionary biology studies has used a quantitative genetics approach to understand phenotypic variation, partitioning it into genetic, environmental and genotype-environment variance. More recently, molecular and developmental sciences started to depict the mechanisms behind the relationship between genotype and environment, revealing the mechanisms of epigenetic effects (Richards *et al.*, 2010).

The term "epigenetics" was coined by Conrad Waddington to describe "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being" (Waddington, 1942). Although Waddington's definition is very broad, encompassing all gene activity during the development that causes the phenotype to emerge, it was the first effort to describe events that could not be explained by existing genetic principles (Waddington, 1953; Goldberg *et al.*, 2007). Epigenetics can be better defined as the study of molecular processes that can affect gene expression and its function without a change in the underlying DNA sequence (Richards, 2006; Bird, 2007).

That is, it assumes there are multiple layers of molecular regulation of genetic information before it is expressed as a phenotype (Niederhuth and Schmitz, 2017).

In general, the research labelled epigenetics was marginalised and/or rarely carried out until the 2000s, when an increased number of studies in DNA methylation and histone modifications were published using this label (Deichmann, 2016a). This led to several conceptual and theoretical papers discussing the potential effects that epigenetic modifications could have on adaptation and speciation (Jablonka and Lamb, 1998; Pál and Miklós, 1999), although the proposed mechanisms have always been difficult to empirically test. In addition, the field often faces some suspicion, as the use of the term "epigenetics" is sometimes employed loosely and inconsistently. For example, the "epigenetics hype" in science and in popular culture (Maderspacher, 2010) sometimes falsely claims victory over the genes and defends a shift of paradigm in evolutionary biology to re-habilitate Lamarckian inheritance (Laland *et al.*, 2014; Deichmann, 2016b). These currents of thought are misleading for many reasons, including the fact they ignore epigenetic changes are not directed as depicted by Lamarckian theory (Deichmann, 2016b). Epigenetics is essentially a field in its infancy, which makes it open to misunderstanding and controversy. However, with the next generation sequencing revolution and associated technological advances, the last ten years have seen a growing number of studies providing evidence of epigenetic processes underlying biological patterns. Because there is still a lot of debate and many gaps to be filled, this is a very exciting time to study epigenetics.

1.1. DNA methylation: the most investigated epigenetic mechanism

Epigenetic mechanisms can involve methylation of cytosine residues in the DNA, remodelling of chromatin structure through histone modifications, and regulatory processes mediated by small RNAs (Bird, 2007; Law and Jacobsen, 2010). These epigenetic modifications directly shape the structure of the genome by defining regions of euchromatin and heterochromatin, and by mediating and facilitating gene expression. Among these

mechanisms, DNA methylation is by far the best-studied one. DNA methylation is a covalent base modification, which involves the addition of a methyl group (CH_3) normally at the fifth carbon (C-5) of the cytosine's pyrimidine ring to form 5-methyl-cytosine (Fig. 1). In animals, methylation occurs almost exclusively in cytosine followed by guanine residues in a symmetric conformation in the DNA (*i.e.* CpG context), although it can be found in other contexts (*i.e.* in symmetric CHG or asymmetric CHH, where H stands for non-G nucleotides; Feng *et al.*, 2010). In animals, DNA methylation in CpG context is typically mediated by two enzymes (Goll and Bestor, 2005). The *de novo* DNA methyltransferase (DNMT3) adds methyl groups in specific DNA sites, and has particular importance during embryogenesis (Law and Jacobsen, 2010). The established methylation patterns across the genome are maintained by DMNT1 at mitosis, adding methyl groups at the newly synthesized strand based on the symmetrical information present in the original strand (Goll and Bestor, 2005).

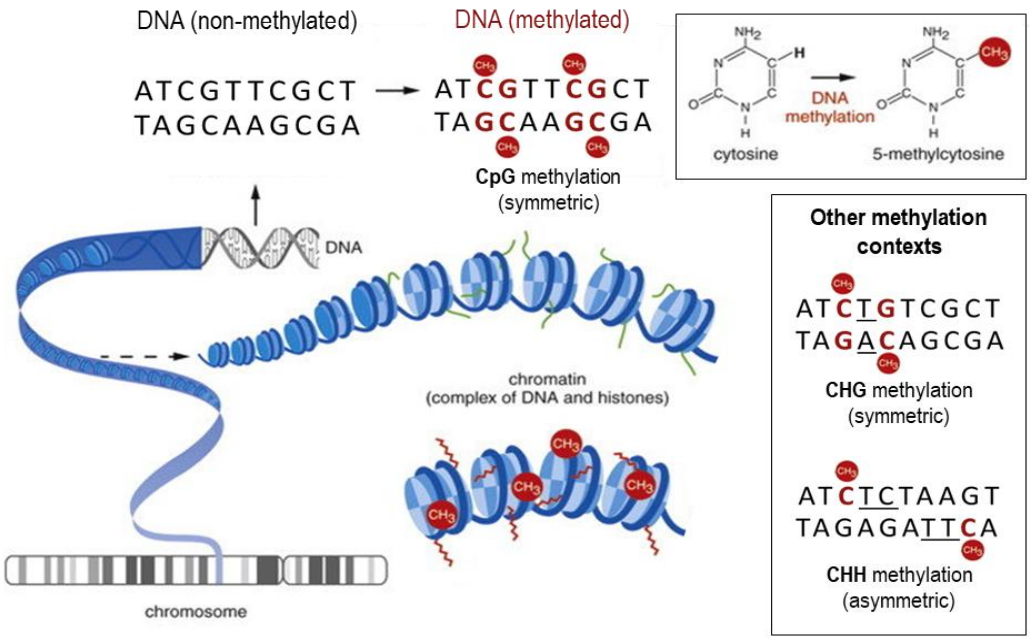


Figure 1: Chromatin modifications mediated by DNA methylation, the addition of a methyl group (CH_3) on the fifth carbon of cytosine residues. It can determine the structure and activity of the genome by defining regions of euchromatin and heterochromatin, and by mediating and facilitating gene expression (Law and Jacobsen, 2010). DNA methylation patterns, functions and molecular pathways vary taxonomically (Suzuki and Bird, 2008; Feng *et al.*, 2010; Zemach *et al.*, 2010). DNA methylation is mainly found on cytosines followed by guanines context (CpG), which is symmetric between strands. Other contexts include CHG and CHH, where H corresponds to non-guanine nucleotides. Figure adapted from Mukherjee *et al.* (2015).

DNA methylation is present in most major eukaryotic groups (Zemach *et al.*, 2010), and it is known to play roles in modulating gene expression, in genomic imprinting, in alternative splicing, and in maintaining genome integrity by supressing transposable element activity (Law and Jacobsen, 2010; Schübeler, 2015). Many studies have demonstrated that these properties of DNA methylation can be translated into phenotypic variation (Cubas *et al.*, 1999; Manning *et al.*, 2006). In addition, it is known DNA methylation can change in response to environmental triggers and ultimately affect the phenotype, making it a possible mechanism behind phenotypic plasticity (Kucharski *et al.*, 2008; Bossdorf *et al.*, 2010; Parrott *et al.*, 2013). DNA methylation is intimately linked with cell differentiation during embryogenesis, which may determine which genes will be transcriptionally active in different tissues (Reik, 2007). For this to occur, extensive demethylation happens in the genome between generations to assure the pluripotency of the embryo and its correct development in plants and mammals (Reik, 2007; Crevillén *et al.*, 2014). This is why DNA methylation patterns tend to be reset during gametogenesis. However, there is accumulating evidence of incomplete erasure of DNA methylation marks, which could be inherited at least for a few generations (Waterland and Jirtle, 2003; Hagmann *et al.*, 2015; van der Graaf *et al.*, 2015). That is, if inherited epigenetic variants can cause phenotypic diversity and possibly lead to fitness differences, there can be a background for natural selection to act upon. At the same time, given DNA methylation variation can be affected by environmental cues, it could provide an additional pathway for evolutionary change (Bossdorf *et al.*, 2008).

1.2. Ecological epigenetics[1]

The ways that DNA methylation can potentially contribute to ecological and evolutionary processes were the focus of several recent literature reviews (Bossdorf *et al.*, 2008; Jablonka and Raz, 2009; Richards *et al.*, 2010; Smith and Ritchie, 2013; Hu and Barrett, 2017; Richards *et al.*, 2017). However, empirical evidence underlying these processes remains scarce. With this in mind, some key topics should be addressed to understand the importance of DNA methylation in an ecological and evolutionary context (Bossdorf *et al.*, 2008; Richards *et al.*, 2017). They concern: (A) the patterns and diversity of natural DNA methylation; (B) the origins and drivers of this variation; and (C) the ecological and evolutionary consequences of natural DNA methylation variation (Fig. 2).

A) Patterns and diversity of natural DNA methylation

How do patterns of DNA methylation vary between species? Even though DNA methylation is widespread among eukaryotes, its patterns and functions vary taxonomically (Feng *et al.*, 2010; Zemach *et al.*, 2010). For example, the type and number of DNA methyltransferases vary between species, which reflects the establishment of methylation during mitosis and meiosis (Goll and Bestor, 2005). While vertebrate genomes are globally methylated (except for regions where methylation status changes dynamically, affecting gene activity; Jones, 2012), many invertebrates lack DNA methylation or it is sparsely distributed in the genome. While transposable elements (TEs) are silenced by being highly methylated in vertebrates and in plants, this is not always the case in invertebrates (Suzuki and Bird, 2008; Cortijo *et al.*, 2014). Differences in the methylation setting of individual sites

[1] *Ecological epigenetics:* study of epigenetic processes in an ecological context, focusing in understanding the epigenetic contributions to ecological and evolutionary processes in nature (Bossdorf *et al.*, 2008).

are likely to be established, maintained and interpreted by different molecular pathways (Niederhuth and Schmitz, 2017). Thus, across taxa, there is great variation in the function and importance of DNA methylation in biological processes. Because the literature is biased towards model organisms, it is important to expand investigations to non-model systems. Ultimately, by comparing the patterns between different clades, one can hypothesize about the different functions of DNA methylation and their evolution and then test the predictions that arise.

In addition, one can explore *how DNA methylation patterns vary in natural populations of the same species.* Although research of DNA methylation under a laboratory setting (*e.g.* Johannes *et al.*, 2009; Verhoeven *et al.*, 2010; Foret *et al.*, 2012) is valuable when trying to find the mechanisms underlying DNA methylation changes and the molecular pathways leading to them, it is also desirable to place the studies into a natural context, in the complex environments where organisms live and evolve. By studying the extent and spatial structure of natural DNA methylation, one can capture the effects of forces that are possibly acting cumulatively over many generations (Herrera *et al.*, 2016). With this, one can obtain a more comprehensive understanding of the role DNA methylation variation might play in ecological and evolutionary processes (Richards 2008, 2011; Richards *et al.* 2010; Herrera et al. 2014).

B) Origins and drivers of natural DNA methylation variation

Natural DNA methylation variation can result from many factors. It can arise by stochastic changes, by response to environment, and by genetic control, and possibly be further shaped by forces of natural selection and drift (Bossdorf *et al.*, 2008; Richards *et al.*, 2017). Some studies have assessed the effect of these factors in laboratory conditions, suggesting: (i) that epimutations rates are elevated compared to genetic mutations (Becker *et al.*, 2011; van der Graaf *et al.*, 2015); (ii) that there is genetic control over methylation

variation (see Taudt *et al.*, 2016); (iii) that a responsiveness to environmental variation can trigger changes in ecologically relevant traits (see Feil and Fraga, 2012; Duncan *et al.*, 2014). These questions can be extended to natural environments, to ask specifically about *the extent to which genetic variation can explain natural DNA methylation variation*, and about *whether environmental change can influence epigenetic variation.* As part of this investigation, it is useful to try to understand *the heritability of DNA methylation variation* (Richards *et al.*, 2010). As such, one can determine how DNA methylation patterns vary in space, which factors influence that variation, and whether adaptive changes in methylation are likely to accumulate over time (*i.e.* estimates of its heritability can be used to determine if it will evolve).

C) Ecological and evolutionary consequences of natural DNA methylation variation

The importance of DNA methylation variation on ecological and evolutionary processes might be considered from its influence on phenotypes. Investigating *how much DNA methylation affects ecologically relevant traits* and *whether they respond to environmental change* will shed light on its contribution to phenotypic plasticity, and from there to numerous ecological processes. For example, if DNA methylation mediates phenotypic plasticity in response to environmental shifts, it could be an important facilitator for responding to new threats (Herrera and Bazaga, 2011) and to climate change (Dimond and Roberts, 2016), at colonization of new habitats (Herrera *et al.*, 2012), and at biological invasions (Richards *et al.*, 2012; Xie *et al.*, 2015; Ardura *et al.*, 2017; Huang *et al.*, 2017). Over a longer timescale, the evolutionary consequences of DNA methylation variation will depend on the *extent of inheritance and stability of the epigenetic variation across generations*. Addressing this point requires the long term evaluation of ecological processes outlined above (Richards *et al.*, 2017).

In this thesis, I have developed a body of work addressing some of the outstanding questions described above. I considered natural DNA methylation variation, its patterns,

drivers, and its ecological and evolutionary consequences. To this end, I have combined ecological, genetic and DNA methylation data from natural populations to investigate outstanding issues regarding: (A1) DNA methylation patterns at the between-species level; (A2) DNA methylation variation and spatial structure in different populations of the same species; (B1) the extent to which within-species variation is associated with genetic variation, and (B2) with environmental variation; (B3) the heritability of DNA methylation variation; and (C) the ecological consequences of DNA methylation variation (Fig. 2). Below I introduce the organism used to conduct this study, an overview of the methods applied, and a brief summary of the issues explored in each chapter.

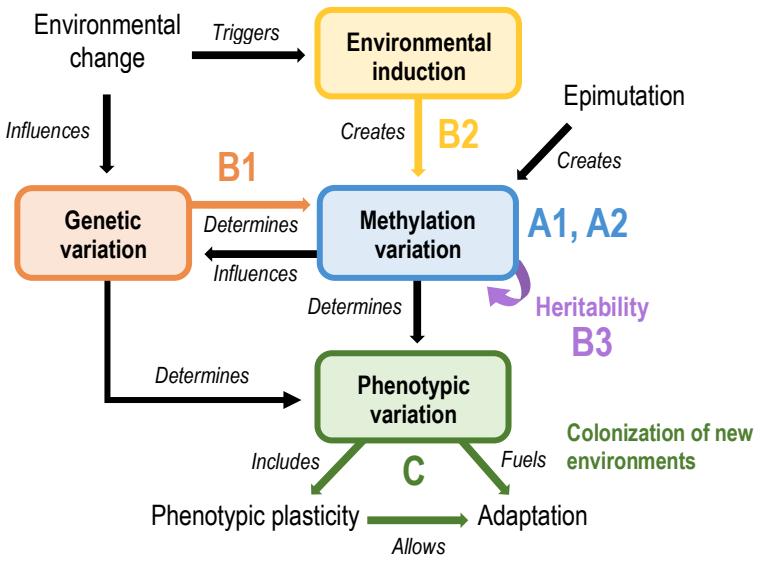


Figure 2: Outstanding questions in the study of DNA methylation from an ecological perspective. This thesis is focused on the following issues: (A1) DNA methylation patterns at species level; (A2) DNA methylation variation and spatial structure in different populations of the same species; (B1) the extent the within-species variation is associated with genetic variation, and (B2) with environmental variation; (B3) the heritability of DNA methylation variation; and (C) the ecological consequences of DNA methylation variation. Figure adapted from Richards *et al.* (2017).

1.3. Study system: Timema cristinae *stick insects*

To address the questions outlined above (Fig. 2), I have used *Timema cristinae* stick insects (Phasmatodea: Timematodea; Vickery, 1993) as a model. *Timema* are plant-feeding insects native to the chaparral in Santa Ynez Mountains, in Southern California (Sandoval,

1994a). Like other stick insects, *T. cristinae* is hemimetabolous and does not undergo metamorphosis. After hatching, individuals go through a series of moults during their nymphal instars until they reach adulthood, lacking a pupal stage. They are univoltine and their life cycle lasts for approximately 16 weeks (Sandoval, 2000), hatching in February and reaching adulthood in April (personal observation). Both nymphs and adults are wingless, and they rest on their host plants during the day, and feed on leaves at night. They can disperse very little, with a mean distance of travel of 2 metres per week and maximum of 8 metres per week (Sandoval, 2000). This suggests that individuals can travel to up to 128 metres during their 16-week lifetime (assuming a constant linear travel rate; Sandoval, 2000).

Although *T. cristinae* can feed on a variety of plant species, it is primarily found on two species of host plant: *Ceanothus spinosus*: Rhamnaceae and *Adenostoma fasciculatum*: Rosaceae, which define the *Timema*'s ecotypes. These two host plant species differ considerably in their leaf morphology, with *Ceanothus* plants presenting broad leaves and *Adenostoma* plants exhibiting thin needle-like leaves (Fig. 3). *Timema* rely on crypsis to escape detection by visual predators, having evolved body colouration that matches the leaves and stems of the host plants they rest on (Sandoval, 1994a). A green morph bearing a dorsal white stripe is more frequently found on *Adenostoma* plants, and a green and unstriped morph on *Ceanothus* plants (Fig. 3; Sandoval, 1994a). Manipulative field experiments have shown predation is a key factor determining differential survival rates of these two morphs depending on the host plant species they are resting on (*i.e.* survival rates do not differ when predators are precluded to access the experimental sites; Nosil, 2004). The striped morph is more cryptic and suffers less predation on the needle-like leaves of *Adenostoma*, whereas the green unstriped one is more cryptic and suffers less predation on the broad leaves of *Ceanothus* plants (Sandoval, 1994a; Nosil and Crespi, 2006). In other words, divergent selection promoted by differential predation between the two host plant

species contributes to ecological isolation between the two *Timema* ecotypes (Sandoval, 1994a; Nosil and Crespi, 2006).

Figure 3: The two main *T. cristinae* ecotypes, characterized by the host plants (**A**) *Adenostoma fasciculatum* and (**B**) *Ceanothus spinosus*. Individuals from the *Adenostoma* ecotype typically have a longitudinal white dorsal stripe and dark green body colouration, which makes them cryptic on their host plant needle-like leaves. Individuals from the *Ceanothus* ecotype normally have a plain light green body, matching the broad leaves of their host plant (Nosil and Crespi, 2006). Photo on the left by Marc Kummel, and on the right by Aaron Comeault.

Individuals with dark body colouration (*i.e.* melanistic morph) are often found on both host plants, but at much lower frequencies (~10% frequency; Sandoval, 1994a,b). They are cryptic at stems of both hosts, but are conspicuous in leaves (Sandoval, 1994a; Comeault *et al.*, 2015). The three morphs segregate as a highly heritable polymorphism with strong genetic dominance: melanistic body coloration is recessive to green (either striped or unstriped), and stripe pattern is recessive to unstriped (Comeault *et al.*, 2015). Green versus brown morphs of *T. cristinae* are distinguished by a major locus on linkage group eight, named *Mel-Stripe* (Nosil *et al.* 2018). This locus exhibits two major features. First, it spans ~10 mega-bases of sequence and exhibits suppressed recombination (putatively due to an inversion, Lindtke *et al.*, 2017). Second, one edge of the locus exhibits a large-scale (~1

mega base pair) insertion/deletion (indel) polymorphism. Whether one, few, or many loci within *Mel-Stripe* affect colour is still unknown. *Mel-Stripe* exhibits three core haplotypes (*i.e.*, alleles), one corresponding to each morph, designated *s*, *u*, and *m* for green-striped, green-unstriped, and melanistic, respectively (Lindtke *et al.*, 2017). That is, in terms of diploid genotypes, *uu*, *us*, and *um* are green-unstriped; *ss* and *sm* are green-striped, and *mm* is melanistic.

The *Adenostoma* and *Ceanothus* ecotypes differ not only in morph frequencies, but also in a suite of other traits. For example, there are significant differences in body size (individuals form *Ceanothus* ecotype tend to be larger; Nosil and Crespi, 2006) and in host plant preference, as individuals from different ecotypes exhibit greater differences in host preference compared to individuals from the same ecotype, independently of geographical distances (Nosil, 2007). In addition, the ecotypes exhibit mate choice and partial sexual isolation (Nosil, 2007; Nosil and Sandoval, 2008), which is associated with differences in cuticular hydrocarbons (CHCs, molecules with roles in anti-desiccation and in insect communication; Chung *et al.*, 2014; Riesch *et al.*, 2017). Previous studies have shown the *Adenostoma* environment presents some physiological challenges to *T. cristinae* individuals compared to *Ceanothus*, as lifetime fecundity is significantly reduced when they are reared on this host species (Sandoval and Nosil, 2005; Nosil and Sandoval, 2008). However, *Timema* seem to have a good molecular machinery for coping with different plant chemical defences, given that they can feed on a variety of host plants species (Larose *et al.*, 2019).

The landscape where *T. cristinae* is found is characterized by a mosaic distribution of patches of the two different hosts, varying in patch size and abundance of each plant species. Previous studies have shown that *T. cristinae* has probably gone through many episodes of colonization and local extinction of different patches in a metapopulation dynamic (Sandoval, 1994b; Farkas *et al.*, 2013). Gene flow between patches of different selection regimes occurs despite the risk of maladaptation. Allele frequencies (including at *Mel-Stripe* locus) in this species are thus determined by a balance between selection and gene flow

(Sandoval, 1994b). In addition, limited dispersal between non-adjacent patches (*i.e.* allopatric populations) contributes to low gene flow and to the accumulation of genetic differentiation by neutral processes, resulting in patterns of isolation by distance (Sandoval, 1994b). The clear understanding about evolutionary ecology in *T. cristinae* in terms of interplay between genotype, phenotype and the surrounding environment (Nosil and Crespi, 2006; Gompert *et al.*, 2014; Comeault *et al.*, 2015) provides a good opportunity to test some of the aims of the thesis (Fig. 2).

The estimate genome size in *T. cristinae* is 1.3 gigabases, comprising 13 linkage groups (*i.e.* chromosomes; Soria-Carrasco *et al.*, 2014). The most recent genome assembly (version 1.3c2) presents a total length of 953.3 megabases (73.3% of the estimate genome size; Nosil *et al.*, 2018). It was generated using Chicago libraries, which are produced by reconstructing the chromatin *in vitro* followed by chemical stabilization with histones, digestion with restriction enzymes, and ligation. As a result, it generated longer scaffolds (4,068 scaffolds, N50=16.4 megabases, N90=1.1 megabases, L50=16, L90=135). Analyses performed in this dissertation showed the quality of the current version of *T. cristinae* genome assembly is reasonably good in terms of gene completeness (using Benchmarking Universal Single-Copy Orthologs tools; Waterhouse *et al.*, 2017; Appendix A, Chapter 2).

1.4. An ecological epigenetics study in T. cristinae

This thesis focuses on investigating natural DNA methylation variation in an ecological context. Some of the issues addressed here are analogous to those addressed in genetics, which means there is an initial framework, which can be modified to study epigenetics (Bossdorf *et al.*, 2008; Herrera *et al.*, 2016). At the same time, DNA methylation has specific attributes that mean the data must be processed and interpreted accordingly, such as being more prone to spontaneous changes (*i.e.* it is less stable than DNA), its sensitivity to environmental conditions, and the general reprogramming between

generations (Richards *et al.*, 2017). As *Timema* is a non-model organism, a good first step is to depict and to describe the methylation patterns, and to compare them to patterns observed in other species in order to understand its typical genomic context and molecular functions (A1, Fig. 2). Then, one can move to a within-species approach to investigate the DNA methylation variation in nature (A2, Fig. 2), and the drivers influencing it (B1-B3, Fig. 2). As multiple factors can underlie this variation, it is of key importance to identify and isolate them in order to understand their individual effects and ecological consequences (C, Fig. 2).

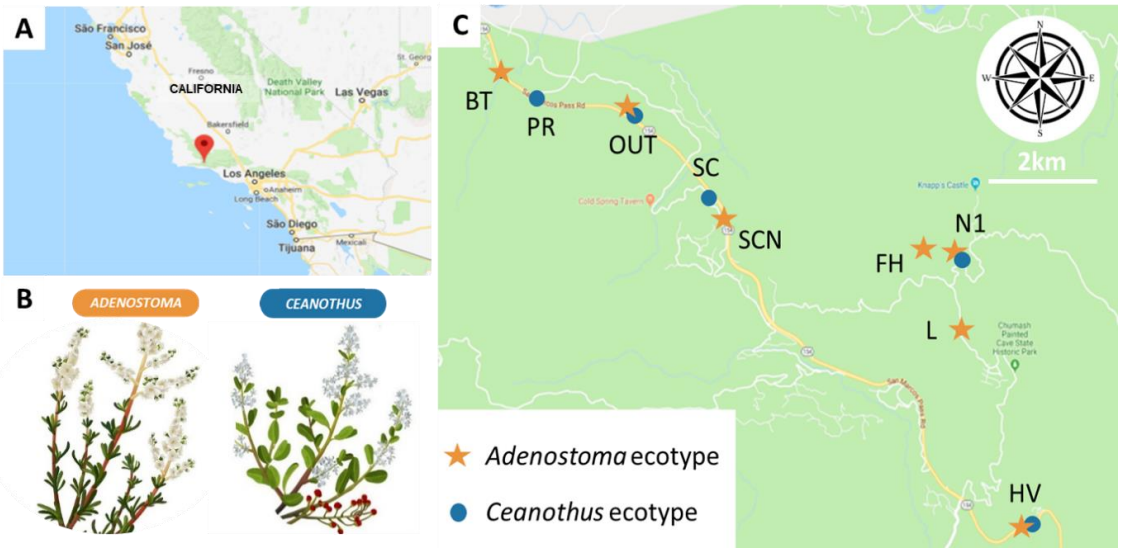


Figure 4: Map detailing geographic position of *T. cristinae* populations included in the sampling plan. (**A**) Location in Southern California where the species is found (Santa Ynez Mountains, Los Padres National Forest). (**B**) Representation of the two main host plants characterizing *T. cristinae* ecotypes (*Adenostoma fasciculatum* and *Ceanothus spinosus*). (**C**) Populations selected for the survey.

With this in mind, a sampling strategy was designed aiming to capture substantial variation in DNA methylation in wild *T. cristinae* and to disentangle some of the factors that could be shaping this variation. Throughout this thesis, a 'population' was defined as all insects collected within a homogeneous patch of a single host species (*i.e.* a locality), as has been done in previous *Timema* studies (*e.g.* Sandoval 1994a,b; Nosil *et al.*, 2002; Sandoval and Nosil, 2005). Some key factors were selected to be studied in this population survey:

abundance of host plants, elevation, climatic variables and geographical distance. Considering all these factors, 12 localities were chosen (Fig. 4): four with only *Adenostoma*, two with only *Ceanothus* (*i.e*. pure populations), and six with a mixed landscape where patches of the two host plants are side-by-side (Table 1). Each selected locality presents a different combination of the factors cited above[2]. In addition, given methylation variation can be under genetic control (Dubin *et al.*, 2015; Taudt *et al.*, 2016), genetic variation was assessed in each population by using previously published sequencing data, and by performing new additional genotyping by sequencing.

To obtain genome-wide information on methylation levels, two similarly sized and large females from each selected population had their whole-genome sequenced after bisulfite treatment (BS-seq). This treatment consists of a reaction between sodium bisulfite and DNA, which leaves methylated cytosines unaffected, and converts non-methylated cytosines into uracil residues (subsequently amplified as thymines, Cokus *et al.*, 2008). Ultimately, estimates of DNA methylation levels can be obtained at individual cytosines by comparing the number of non-converted cytosines (*i.e.* methylated bases) and the number of thymines (i.e. non-methylated bases) at a specific position. Thus, these datasets have properties that differ in fundamental ways from other high-throughput sequencing genomic data (Lea *et al.*, 2017); these properties are addressed throughout the research described in this thesis. Differences in methylation status can be obtained by analysing single loci (*i.e.* single methylation polymorphisms; SMPs), or by investigating larger genomic regions, the so called differently methylated regions (DMRs). The DMRs span close sites that have different methylation patterns between samples and are regarded as possible functional regions (Lea *et al.* 2017). Studies in plants and mammals have extensively worked with

[2] Complete information about the sampling design and about each factor is described in full details in Chapter 3 of this dissertation.

Table 1: Localities selected in the population survey and details of the 24 individuals used in the studies presented throughout the thesis.

Locality	*Host*	*Latitude*	*Longitude*	*Description*	*Ind.*	*Morph*	*BL*	*BW*	*HW*
N1	A	34.517	-119.797	Network 1	17_0003	G	2.1	0.4	0.2
					17_0005	G	2.0	0.4	0.2
N1	C	34.517	-119.797	Network 1	17_0006	G	2.1	0.4	0.2
					17_0009	M	1.9	0.4	0.2
FH	A	34.518	-119.801	Far Hill	17_0012	S	1.9	0.4	0.2
					17_0015	S	1.9	0.4	0.2
L	A	34.509	-119.796	Laurel Springs	17_0018	S	1.8	0.4	0.2
					17_0019	S	1.8	0.4	0.2
HV	A	34.488	-119.787	Hidden Valley	17_0043	G	2.1	0.4	0.2
					17_0045	S	2.1	0.4	0.2
HV	C	34.488	-119.786	Hidden Valley	17_0049	S	2.2	0.4	0.2
					17_0051	M	2.0	0.4	0.2
SCN	A	34.521	-119.83	Stagecoach North	17_0057	S	1.9	0.4	0.2
					17_0058	S	2.1	0.4	0.2
SC	C	34.523	-119.832	Stagecoach	17_0062	G	2.1	0.5	0.2
					17_0065	G	2.0	0.4	0.2
OUT	A	34.532	-119.843	Outlook	17_0067	G	2.1	0.4	0.2
					17_0070	G	2.1	0.4	0.2
OUT	C	34.532	-119.844	Outlook	17_0074	G	1.9	0.4	0.2
					17_0075	S	1.8	0.4	0.2
PR	C	34.533	-119.857	Paradise road	17_0077	G	2.1	0.4	0.2
					17_0081	G	2.0	0.4	0.2
BT	A	34.536	-119.862	Bottom	17_0082	G	2.0	0.4	0.2
					17_0086	G	1.9	0.4	0.2

Morph abbreviations: G=green, S=striped, and M=melanistic. BL= body length, BW=body width, HW= head width. Morphometric measurements were performed in ImageJ 1.4.882 (Abràmoff *et al.*, 2004), following previous works on *T. cristinae* (Comeault *et al.*, 2014; Riesch *et al.*, 2017). All individuals used in this work were female.

DMRs, as DNA methylation levels are spatially correlated in promoters, transposable elements and regulatory regions (Suzuki and Bird, 2008; Lea *et al.*, 2017). In insects, the studies are usually conducted based on SMPs (*e.g.* Bonasio *et al.*, 2012; Glastad *et al.*, 2016; Libbrecht *et al.*, 2016). Considering the gaps in the knowledge of the genomic distribution of DNA methylation in insects, and the limitations in the use of DMRs (*e.g.* it tends to extrapolate or ignore the variation between the samples; it faces statistical limitations dealing with the binomial nature of methylation data; Gaspar and Hart, 2017), I focused at

investigating SMPs in this dissertation. Although one single methylation polymorphism might not be enough to affect genomic activity, identifying these sites can be a good first step to understand the patterns and variation of DNA methylation between samples. Limitations using this approach are discussed throughout the thesis when appropriate.

Previously published sequencing data (Soria-Carrasco *et al.*, 2014; Comeault *et al.*, 2015; Lindtke *et al.*, 2017; Riesch *et al.*, 2017) were used to identify the regions where some single nucleotide polymorphisms (SNPs) could have been confounded with single methylation polymorphisms (SMPs). Finally, to estimate genetic variation among the samples, each individual with methylation information had its genome sequenced partially using RAD-seq.

1.5. Outline of thesis chapters

Chapter 2 of this thesis addresses the status of DNA methylation in *T. cristinae* stick insects, characterizing its genomic methylation profile for the first time and comparing the emerging patterns to the state-of-the-art in other insect species (A1, Fig. 2). To this end, the population survey BS-seq dataset was used to estimate variation in the species' methylation profile. Chapter 2 thoroughly describes the details of the BS-sequencing steps and generation of this dataset (Fig. 5). The results revealed a highly methylated genome compared to other insects, targeted mostly to gene bodies (*i.e.* exons and introns). Overall, DNA methylation patterns in *T. cristinae* resemble the ones found in other hemimetabolous insects, and show some similarities and differences to vertebrate methylome profiles (Glastad *et al.*, 2016).

Chapter 3 addresses DNA methylation patterns at the population level to capture the within-species variation, and to investigate the possible factors underlying it (Fig. 5). The population survey data were used to test the hypotheses that natural DNA methylation variation: is structured in geographical space (A2, Fig. 2); is associated with genetic

variation (B1, Fig. 2); and is correlated with environmental factors such as climate and host plant species (B2, Fig. 2). This chapter details how the sampling strategy was designed, aiming to disentangle some variables that could be underlying methylation variation. Here, multiple datasets were used, including methylation variation (using the population survey BS-seq data), genetic variation (using newly acquired genetic data and reanalysis of previously published data), and ecological information about the population localities (*e.g.* abundance of host plants, elevation, climatic variables and geographical distance). This chapter revealed that genome-wide DNA methylation variation in *T. cristinae* tends to cluster following the geographical distribution of populations. Multivariate analyses revealed this trend in DNA methylation variation was better explained by its association with genetic variation than by its association with geographical distance. Although there was not a noticeable correlation between general DNA methylation variation and climate or host plants, the results do not exclude the possibility of such associations in specific regions of the genome. Binomial mixed models (Lea *et al.*, 2015) revealed that some similarity in DNA methylation variation was correlated with similarity in genetic kinship, suggesting some heritability of methylation status (*i.e.* specific sites show the same methylation levels over generations). Taken together, these results indicate genetic differences explain DNA methylation variation and suggest that differentiation between populations can accumulate given limited dispersal in space.

Chapter 4 focuses on the interactions between *T. cristinae* and their host plants. Here, studies of natural variation, combined with a rearing experiment, were used to test for a host plant effect on DNA methylation variation (B2, Fig. 2; Fig. 5). The work considers how DNA methylation variation can be implicated in host shifts and colonization of new environments (C, Fig. 2). The experiment involved rearing adult *T. cristinae* on different host plants in controlled conditions to test for a response to host shift in methylation levels. BS-seq from the population survey and from six individuals used in the experiment were used to obtain information about methylation variation. Methylome scans using binomial mixed

models performed independently on the different datasets suggested some potential single methylation polymorphisms (SMPs) associated with host plant. In particular, SMPs located in the coding region of an insect allergen gene were identified in the outputs of both analyses of natural and experimental populations. In other insects this gene is related to digestion and nutrient uptake (Randall *et al.*, 2013). This region was differently methylated between the ecotypes in natural populations. The rearing experiment suggested it responded to the host shift in the opposite direction as expected from the natural population survey, so that the response to the environmental change could be interpreted as 'non-adaptive'. Although there was not any measurement of gene expression or of fitness more analyses are needed to support these results, this study suggested not only that methylation can respond to an environmental change, but also that it does not necessarily happen towards the 'optimum state'. Copies of this gene domain are found in other regions of the genome, but they showed no traces of methylation. In summary, these first results suggest some good candidate genes for investigating the role of methylation in the interaction between *T. cristinae* and their host plants.

Finally, Chapter 5 concludes the findings of this dissertation and outlines some unresolved issues and directions for future work in the field of ecological epigenetics. Collectively, these studies represent analyses of *T. cristinae* DNA methylation patterns at different scales: from a broad species level, through to the factors underlying within-species variation, and closing with a focused study of whether DNA methylation status can affect insect-plant interactions. Overall, this thesis highlights the importance of studying this molecular feature to better understand complex organismal life.

1.6. A note on contributions made to this thesis

In the proceeding chapters of this dissertation, I have used 'I' instead of 'we' when describing the research. Although I was the one to lead the work, I had support from my colleagues and mentors. For example, Dr. Romain Villoutreix initiated the work on DNA methylation in *T. cristinae* and developed some of the first scripts to process BS-seq data. In addition, together we designed and performed the rearing experiment (Chapter 4). Dr. Víctor Soria-Carrasco developed the pipelines to process genetic data, and was constantly advising me not only on bioinformatics jobs and analyses, but also during the entire thesis writing process. Finally, my two supervisors Dr. Patrik Nosil and Dr. Jon Slate have always provided great support in the interpretation of results and in critical discussions. Although these four researchers were not represented at the beginning of each chapter, they will be acknowledged in the manuscripts for publication, following the conventions of authorships.

1.7. Note on contribution to appendix D

Appendix D comprises a manuscript submitted for publication, titled *Ecology helps explain whether genes for cryptic coloration form a supergene or recombine*, on which I am a co-author. I included this piece of work as part of my thesis because I have significantly contributed to it, having led some of the experiments and analyses detailed in the manuscript. This manuscript describes the ecological aspects involved in polygenic adaptation in *Timema* stick insects. By combining studies in ecological, experimental and genomic datasets, we identified multiple, recombining loci affecting cryptic colouration in different *Timema* species. Briefly, we found high recombination among colour genes is associated with use of host plants that favour continuous colour variation. Conversely, the use of host plants with discrete colours (*i.e.* green leaves versus brown stems) is associated with strong disruptive selection and suppressed recombination – involving a structural variation that formed a supergene. In this manuscript, I led the analyses about crypsis,

investigating the colour variation on the insects and on their host plants: the phenotypic measurements from photographs, the differentiation, overlap, and correlation analyses between morphs. In addition, Patrik Nosil and I designed the manipulative field experiment to test the hypothesis regarding colour variation in *Timema* (*i.e.* continuous versus discrete colour polymorphisms) and adaptation to their host plants. I presented the preliminary results about genome-wide analysis studies (GWAS) and camouflage in *Timema chumash* from this study in my confirmation evaluation at University of Sheffield.

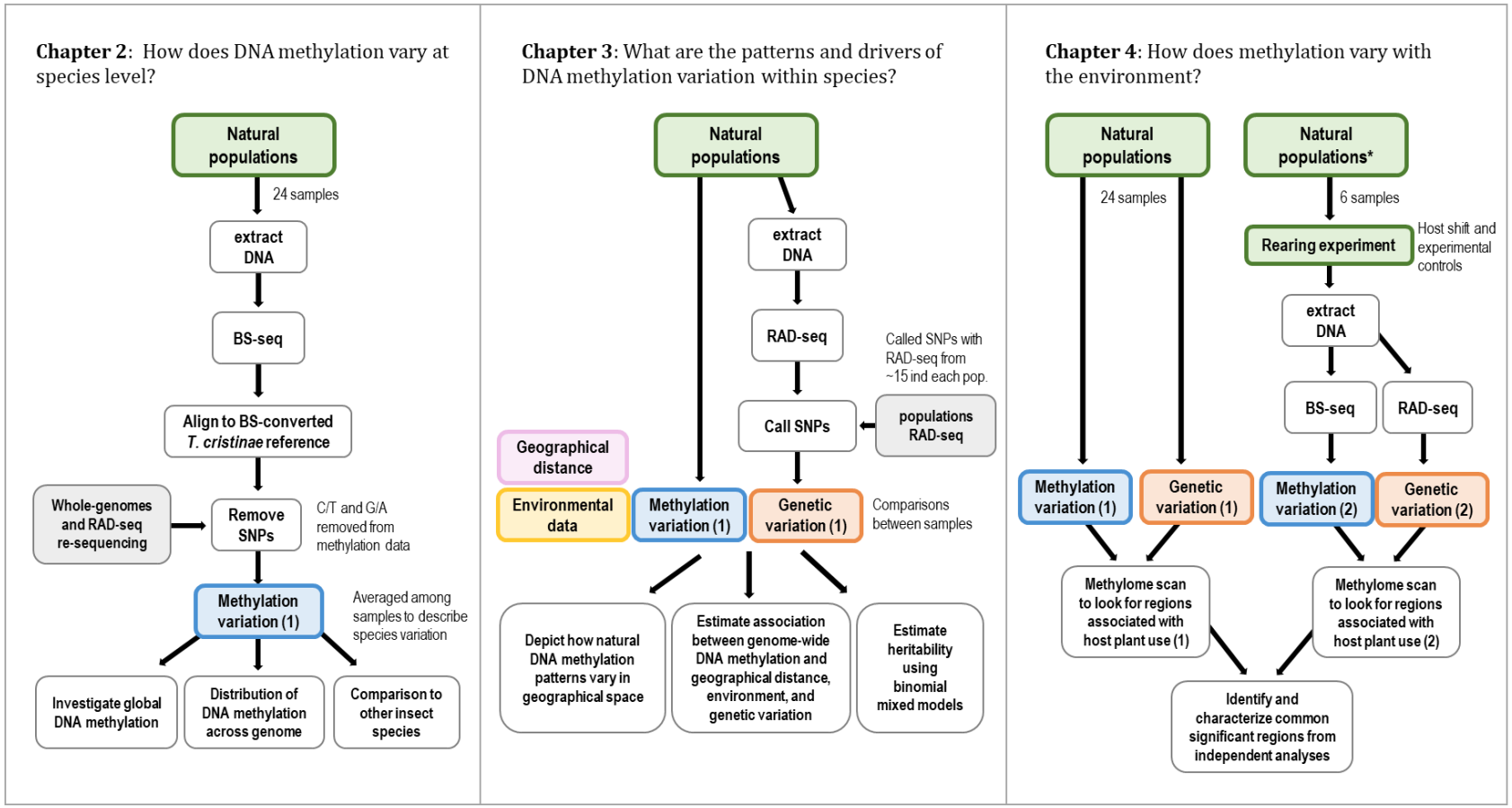


Figure 5: Flow chart describing the contents of each chapter in this thesis. Diagrams depict the steps used to manipulate the data acquired from natural populations. Details about the BS-sequencing steps and generation of methylation variation are thoroughly described in Chapter 2, and about acquiring and processing genetic data in Chapter 3. 'Methylation variation (1)' and 'Genetic variation (1)' correspond to data generated from the population survey, while 'Methylation variation (2)' and 'Genetic variation (2)' correspond to the rearing experiment data (*i.e.* these datasets were obtained independently). Asterisk in 'Natural populations*' represent different population sampling events from the initial population survey of 24 individuals.

Chapter 2

Function and evolution of DNA methylation in *Timema cristinae* stick insects

2.1. Summary

DNA methylation is involved in gene expression, genomic imprinting, alternative splicing, and in silencing transposable elements. Although DNA methylation is widespread among eukaryotes, it differs considerably among taxa, which can affect its role and importance. The variable patterns among different insect species denotes an evolutionarily flexible role of DNA methylation. However, little is known about its function and evolution in the group as most studies are focused in eusocial insects. Thus, investigating different clades can expand the knowledge of patterns and functions of DNA methylation in insects. Here, whole-genome bisulfite sequencing was used to describe the DNA methylation profile in *Timema cristinae* stick insects. The results suggest the DNA methyltransferase 3 (DNMT3) has been lost in this species, relying only on DNMT1. *Timema* presents elevated global DNA methylation levels compared to other insects (14% mCpG), targeting mainly the gene body (*i.e.* both exons and introns) increasing towards the 3' end, in patterns that resemble other "Hemimetabola" insects. Methylated genes generally play housekeeping functions, while non-methylated genes play signalling and transduction functions. Similar to other insects, transposable elements were impoverished in methylation. With this work, I highlight the importance of investigating different insect taxa to obtain a better understanding of some specific and general roles of DNA methylation.

2.2. Introduction

DNA methylation is a covalent base modification, which normally happens at the fifth carbon (C-5) of the base cytosine's pyrimidine ring to form 5-methyl-cytosine. It is ubiquitous in eukaryotes (Feng *et al.*, 2010), and it is known to affect gene expression (switching genes on and off by influencing transcription factor binding, Schübeler, 2015), alternative splicing (Foret *et al.*, 2012; Sati *et al.*, 2012), and transcriptional elongation (Lorincz *et al.*, 2004). In animals, methylation occurs almost exclusively in cytosines (C) followed by guanines (G) in a symmetric conformation in the DNA (*i.e.* CpG context), although it can be found in other contexts (*i.e.* in symmetric CHG or asymmetric CHH, where H stands for non-G nucleotides; Feng *et al.*, 2010). The DNA methylation activity in CpG context is evolutionarily well conserved, and it is typically mediated by two enzymes in animals (Goll and Bestor, 2005). The *de novo* DNA methyltransferase (DNMT3) adds methyl groups in specific DNA sites, and has particular importance during embryogenesis and tissue differentiation (Law and Jacobsen, 2010). The established methylation patterns across the genome are then maintained by DMNT1 at mitosis, adding methyl groups at the newly synthesized strand based on the symmetrical information present in the original strand (Goll and Bestor, 2005). The DNMT2 is not involved in DNA methylation, but it is rather a tRNA methyltransferase (Goll *et al.*, 2006). A number of enzymes is responsible for the demethylation activity, and it varies in different taxa (Law and Jacobsen, 2010). Thus, the balance between methyltransferase and demethylation enzymes activities culminates with a characteristic DNA methylation pattern in tissues, individuals, populations and species.

Although CpG methylation is widespread across the tree of life, the proportion of methylated cytosines in the genome, its distribution, and genomic targets vary across different taxa (Suzuki and Bird, 2008). In vertebrates, the genome is globally methylated, except promoter regions which are generally non-methylated (Jones, 2012). These regions are the so called CpG islands, and their methylation state is known to be dynamic, which

influences whether a transcription factor will bind to the region and initiate the transcription or not (*i.e.* respectively the non-methylated and methylated state; Yin *et al.* 2017; Onuchic *et al.* 2018). This role of methylation is well described in vertebrates (Jones, 2012; Schübeler, 2015), and a number of studies has shown associations between environmental variables and changes in methylation state and in gene expression (Duncan *et al.*, 2014). In invertebrates, DNA methylation levels are much reduced and distributed in sparse patterns across the genome (Suzuki and Bird, 2008). In addition, in invertebrates DNA methylation is mainly found in gene bodies, which suggests that methylation may have different properties in invertebrates and vertebrates (Zemach *et al.*, 2010).

For many years, it was speculated that insects underwent very little or no DNA methylation in their genomes, as the model organism *Drosophila melanogaster* exhibits insignificant methylation levels (Goll and Bestor, 2005). In fact, DNA methylation levels in insects generally follow the mosaic distribution found in other invertebrates (Xiang *et al.*, 2010). However, it seems to vary widely across the group, being absent in many clades. Of the six investigated insect orders in a large phylogenetic comparison, each one exhibits at least one loss of DNA methylation, with no evidence to date of it in dipterans (Bewick *et al.*, 2017). On the other hand, it has been proposed DNA methylation is involved in developmental plasticity and social behaviour. For instance, in honeybees (*Apis mellifera*), the development of larva into queens or workers depends on differential feeding with royal jelly, a process that involves variation in DNA methylation modifications (Kucharski *et al.*, 2008; Foret *et al.*, 2012). These observations gathered great attention, and led to a large focus on studies about DNA methylation in Hymenoptera, which include many examples of social insects, such as ants, bees and wasps (*e.g.* Bonasio *et al.*, 2012; Wang *et al.*, 2013; Patalano *et al.*, 2015). As a result, studies about DNA methylation in other orders are generally underrepresented. Thus, studies should be conducted in different clades for a better understanding of patterns, function and evolutionary importance of DNA methylation in insects. For example, DNA methylation tends to happen more extensively in

insects that undergo incomplete metamorphosis (*i.e.* passing through egg, nymph and adult stages; "Hemimetabola") and is reduced in those that face complete metamorphosis (*i.e.* passing through egg, larva, pupa, and adult stages; Holometabola), where it is occasionally absent (Bewick *et al.,* 2017; Provataris *et al.* 2018). When it is present, DNA methylation normally occurs in genes that are broadly expressed across tissues (Glastad *et al.* 2016; Glastad *et al.* 2017), it tends to be depleted in transposable elements (Wang *et al.*, 2013; Glastad *et al.*, 2017), and it does not necessarily depend on DNMT3 activity to be established, as this enzyme is lacking in some taxa (Bewick *et al.* 2017).

These circumstances challenge the understanding of the role of DNA methylation in insects, which is still widely debated. Even the hypothesis that it plays a general function in the development plasticity of social insect caste systems has been questioned recently (Bewick *et al.*, 2017). In fact, recent empirical evidence indicates there is not a clear relationship between methylation levels and eusociality or reproductive division of labour (Libbrecht *et al.*, 2016; Standage *et al.*, 2016; Glastad *et al.*, 2017). Moreover, the functional role of DNA methylation in gene bodies, as is commonly observed in insects and other invertebrates (Zemach *et al.*, 2010), is not well understood (Hunt *et al.*, 2013). A recent study revealed changes in methylation levels in response to presence or absence of maternal care in a subsocial bee (*Ceratina calcarata*), but those changes were not linked to the changes observed in gene expression or to alternative splicing (Arsenault *et al.*, 2018). In addition, a knockdown of DNMT1 (and its consequential depletion for DNA methylation) in the hemipteran *Oncopeltus fasciatus* resulted in inviable eggs and reproductive failure. However, it did not result in changes in genes or transposable element expression, suggesting DNMT1 and DNA methylation present biological functions that are independent of gene expression (Bewick *et al.*, 2019).

In this work, the levels and patterns of DNA methylation in the *Timema cristinae* stick insect were investigated (Phasmatodea: Timematodea; Vickery, 1993). *T. cristinae* are plant-feeding wingless insects native to the chaparral in Santa Ynez Mountains, in Southern

California (Sandoval, 1994). Although this species is likely to be found in a broad range of host plants, it is typically found in two ecotypes, characterized by the host plants *Adenostoma fasciculatum* and *Ceanothus spinosus* (Fig. 3 in Chapter 1). Like other stick insects, this species is hemimetabolous and develops gradually. That is, it goes through the egg stage, proceeding through a series of nymphal instars, and reaches adulthood after many molts. Although "Hemimetabola" is a paraphyletic group, here this classification is applied to emphasize different metamorphosis processes, contrasting this group with those with complete metamorphosis (*i.e.* Holometabola or Endopterygota). The objectives of this study were: (1) to first describe the general methylation profile in *T. cristinae* and (2) to compare the patterns with what is known in other insect species. For this, DNA methylation information was obtained from samples collected in different populations in the wild by using bisulfite sequencing (BS-seq). To my knowledge, this work used the largest sample size in the study of DNA methylation in insects, which enabled to obtain statistically reliable results. The analyses focused on the CpG context as it is the main context targeted by methylation in animals, but some additional results are reported in CHG and CHH contexts. The results show the genome is highly methylated in *T. cristinae* compared to other insects, and that methylation targets both exons and introns in the gene body. Similar to other insects, the methylated genes were enriched in functions related to fundamental cellular processes, and transposable elements were mostly depleted in methylation. By studying this non-Holometabola species, this work aimed to contribute to the knowledge of different forms, functions and evolution of DNA methylation levels in insects.

2.3. Material and Methods

2.3.1. DNA methyltransferases (DNMTs)

To identify and characterize the DMNT genes, I first used the *T. cristinae* functional annotation (version 1.3c2; Villoutreix *et al. in prep*). Briefly, this functional annotation was obtained using *T. cristinae* RNA sequencing data (Comeault *et al.*, 2012; Misof *et al.*, 2014),

used to generate gene predictions in the reference genome and its respective putative proteins. The functions from the putative proteins were estimated by aligning their sequences to multiple databases, and stored in the *T. cristinae* functional annotation dataset (methods at Villoutreix *et al. in prep*). In this study, putative proteins with C-5 cytosine methyltransferase function were selected from the functional annotation dataset, as this activity is characteristic of eukaryotic DNMT enzymes (Goll and Bestor, 2005). The amino acid sequence from all selected *T. cristinae* putative proteins were input into the BLASTp tool (Altschul *et al.*, 1997) at National Center for Biotechnology Information (NCBI) platform, and aligned to NCBI's non-redundant protein sequence database. The BLASTp analyses returned the most similar protein sequences in the database along with its described function. This similarity is estimated by a matching score and by the Expect value (*E-value*). The E-value describes the number of hits that are expected to be retrieved by chance when searching a database of a particular size. It decreases exponentially as the score of the match increases, so that the lower the E-value the more significant the match is (Altschul et al., 1997). The best hits were selected, and all the accompanying information reported here.

In addition, I aligned DNMT proteins from a few species of insects to the *T. cristinae* reference genome assembly (see Table A1 for estimate of quality of genome assembly). To this end, I downloaded the protein sequences of DNMTs 1 and 3 and its different isoforms from: *Apis melifera* (Hymenoptera: Apidae; Elsik *et al.*, 2014), *Nasonia vitripensis* (Hymenoptera: Pteromalidae; Werren *et al.*, 2010), and *Zootermopsis nevadendis* (Isoptera: Termopsidae; Terrapon *et al.*, 2014) using GenBank (National Center for Biotechnology Information, NCBI). *N. vitripensis* and *A. melifera* were used because their DNMTs toolkit is well characterized (Werren *et al.*, 2010; Provataris *et al.*, 2018). I generated a database on *T. cristinae* reference genome (version 1.3c2; Nosil *et al.*, 2018) using the 'makeblastdb' tool in BLAST+ (Camacho *et al.*, 2009). The DNMTs amino acid sequence were then aligned to this database using tBLASTn (2.8.1+; Altschul *et al.*, 1997).

2.3.2. Sampling

I collected individuals from 12 different locations around Santa Ynez Mountains (Table 1 in Chapter 1). These localities (*i.e.* populations) were chosen based on the different factors, including different genus of host plant (*i.e.* either *Adenostoma* or *Ceanothus*), and varying in geographical distance and climatic variables. Given methylation is expected to vary according to different ecological factors and genetic background, this design aimed to obtain a comprehensive dataset of DNA methylation information in *T. cristinae*. For this, individuals from the different localities were all sampled on the same day (25 April 2017) in the Californian spring. Specimens were collected using sweep nets and kept in plastic containers at room temperature overnight and fed with leaves from the same plant on which they were collected.

Photographs of every specimen were recorded using a Canon EOS 70D digital camera equipped with a macro lens (Canon EF 100 mm f/2.8 L Macro IS USM) and two external flashes (Yongnuo YN560-II speedlights). The pictures were taken with the camera set on manual, an aperture of f/14, a shutter speed of 1/250 s and flashes adjusted to 1/4 power in S2 mode. The selected individuals were then flash frozen using liquid nitrogen one day after sampling (26 April 2017) and preserved at -80^{0}C temperature. Thus, the sampling was immediately followed by preservation. All procedures were performed to assure the methylation status was minimally affected by changing conditions after sampling. This way, one can assume the methylation levels likely match the patterns present in the wild. Individuals were measured using ImageJ 1.4.882 (Abràmoff *et al.*, 2004) following previous work on *T. cristinae* (Comeault *et al.*, 2014; Riesch *et al.*, 2017). Two similar-sized adult females were chosen for BS-seq of each population (Table 1 in Chapter 1).

2.3.3. Generating DNA methylation data

2.3.3.1. Whole genome bisulfite sequencing (BS-seq)

I used whole-genome bisulfite sequencing to obtain genome-wide information on methylation levels in *T. cristinae*. This technique consists of Illumina sequencing preceded by a bisulfite (BS) treatment of the DNA, which involves a reaction between sodium bisulfite and DNA. This treatment converts non-methylated cytosine residues into uracil (subsequently amplified as thymines [T] following polymerase chain reaction [PCR]), but leaves 5-methyl-cytosines unaffected (Fig. 1; Cokus *et al.*, 2008). Thus, only methylated cytosines are retained after this treatment. For every genomic locus, BS treatment and subsequent PCR amplification give rise to four possible different DNA strands, which all have the potential to be sequenced (Fig. 1). Ultimately, estimates of DNA methylation levels can be obtained by comparing the number of methylated bases and the number of non-methylated bases at a specific position. This provides genome-wide information to be processed *in silico* to assess methylation levels at base-pair resolution. This technique has been widely used to estimate DNA methylation information (Lea *et al.*, 2017; Richards *et al.*, 2017) .

I used half of the whole body (cut longitudinally) of each specimen to isolate its genomic DNA using DNeasy Blood and Tissue Kits (Qiagen). Although this implies a mix of DNA from different tissues, this procedure has been used in a number of other studies of insects (*e.g.* Bonasio *et al.*, 2012; Wang *et al.*, 2013; Patalano *et al.*, 2015; Glastad *et al.*, 2016, 2017). In addition, methylation seems to be preferentially targeted to genes that are broadly expressed across tissues (Glastad *et al.*, 2018). A small amount of non-methylated cl857 Sam7 Lambda phage DNA (Promega Corporation) was added to all samples, equivalent to 1% of the final volume to be processed. This strain lacks methylase activity (Arraj and Marinus, 1983), thus all cytosines are non-methylated and are expected to be converted into thymines after BS-treatment (*i.e.* 0% methylated cytosines in the phage sample). Hence,

here the efficiency of BS-conversion was used in the phage as a proxy to determine the conversion efficiency in each sample. In addition, genomic DNA of one *T. cristinae* sample (individual 17_0015) was submitted not only for BS-seq, but also as a control for the BS-treatment (*i.e.* was sequenced without sodium bisulfite treatment).

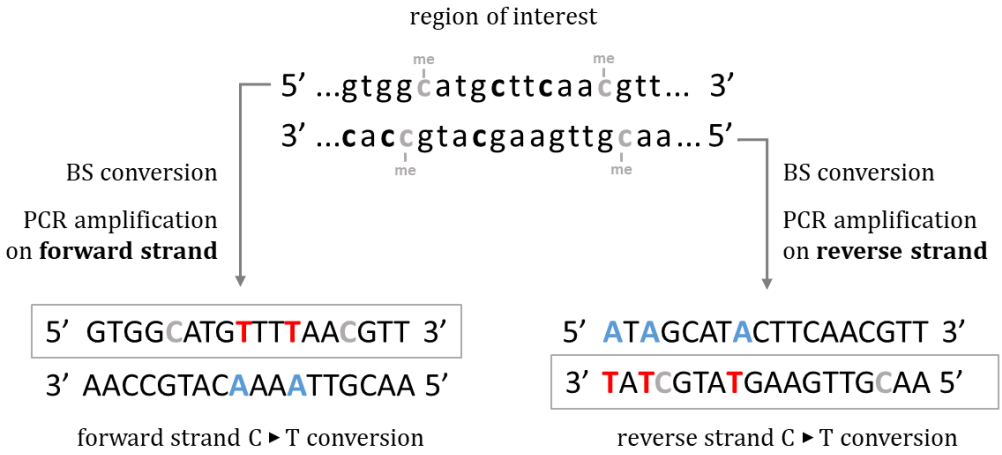


Figure 1: DNA strands generated by bisulfite treatment. Methylated cytosines (grey) remain unaffected after BS treatment, while non-methylated cytosines are converted into uracil and amplified as thymines (red). Adenines (blue) are amplified when new thymines (originally non-methylated cytosines) are used as template. Thus, bisulfite conversion followed by PCR amplification can result in four different DNA strands, and consequently four different states at a specific locus. This figure was adapted from Krueger & Andrews (2011).

The BS-treatment and sequencing were performed by Biomedicum Functional Genomics Unit (FuGU, Helsinki). To perform the sodium bisulfite reaction the DNA was treated with Zymo EZ DNA Methylation-GoldTM kit (Zymo Research). The converted single-stranded DNA generated in this process was used for the non-directional library preparation using TruSeq DNA Methylation Kits. The converted DNA was synthesized using random primers, selectively tagging the 3' end with unique indexes for each sample. After amplification and following purification using AMPure XP beads (0.7x ratio), the libraries were sequenced using the Illumina NextSeq 500 system, with High Output 2 x 150 bp runs. In total, three flow cells with four lanes were run (*i.e.* total of 12 lanes).

2.3.3.2. Filtering reads

Filtering was done using Trimmomatic (0.36; Bolger, Lohse, & Usadel, 2014), and read quality assessed using FASTQC v0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Briefly, reads with the minimum length of 36bp (MINLEN=36) were selected. Given all reads had maximum length of 150bp, I performed the step of removing reads longer than 150bp (CROP=150) only as a convention, as this step did not effectively filter any reads. The first 10bp from each read's start always presented increased cytosine content, and were thus removed (HEADCROP=10). The reads were scanned with 4bp sliding windows, and cut when the mean Phred quality per base dropped below 20 (SLIDINGWINDOW=20). In addition, bases at the start and end of a read were removed if below Phred 20 (LEADING=20, TRAILING=20). Finally, the Illumina specific sequences were all removed from the reads using a custom file (ILLUMINACLIP='2:30:8:1:true'). Phred +33 quality score was used during the filtering. The filtering removed around 1-1.5 million reads in each sample's lane; leaving 8.5-9 million reads to be used for mapping. The reads from different lanes were then merged by sample id. The mean number of reads was 33,634,576 [31,569,755 – 35,699,397] (mean [95% confidence interval] across all 24 samples; Table A2). There were some differences in coverage and in read counts between the flow cells, which were evidenced in downstream studies such as the clustering analyses used in Chapter 3 (data not shown in this dissertation). These batch effects were minimized after standardizing the number of reads between all the samples. After subsampling the reads of each individual to a maximum of 24 million reads (the minimum read count), randomly sampled from the *.fastq* files before mapping (Table A3).

2.3.3.3. Read mapping and methylation calls

The reads were mapped using Bismark (0.16.1; Krueger & Andrews, 2011). This software processes the four reads from the sequencing libraries by converting them *in silico*

into C-to-T and G-to-A versions (*i.e.* the reverse strand of C-to-T; A stands for adenine; Fig. 2). Then, Bismark aligns each one of them to the BS-transformed version of a genome of interest using bowtie2 (Langmead and Salzberg, 2012). This allows the software to determine the strand origin of a BS-seq read and the unique best alignment (Krueger and Andrews, 2011). Finally, Bismark uses the best alignment to do the methylation calls, determining the methylation state of each cytosine on the read (Fig. 3). Here, Bismark's '*bismark_genome_preparation*' tool was used to convert the Lambda phage DNA (GenBank – EMBL Accession Number: J02459) and the most recent *T. cristinae*'s reference genome (1.3c2; Nosil *et al.*, 2018) into their BS-transformed version. The mapping was performed using the '*bismark*' tool, using the paired-end and non-directional options to use all four different strands generated at PCR amplification. The ambiguously mapped reads were always discarded.

I first mapped the good quality reads to the BS-transformed Lambda phage genome to isolate the data from this strain, and to obtain estimates of BS conversion. The mapping yielded a mean of 737,086 [626,125 – 848,047] reads across samples mapped uniquely to Lambda phage (mapping efficiency of 3.1% [2.6% – 3.6%]). The proportion of methylated cytosines in the phage was 0.3% in CpG context, 0.4% in CHG, and 0.3% in CHH (Table A2). This means 0.3% of non-methylated cytosines in CpG context, for example, were not properly converted by the bisulfite treatment, as this strain does not contain any methylated cytosine. Thus, the mean conversion efficiency was 99.7%, in CpG context across all samples. The reads that were not mapped to the phage (23,262,914 [23,151,953 – 23,373,875]) were then aligned to *T. cristinae* BS-converted reference genome, yielding a mean of 10,232,740 [9,803,341 – 10,662,139] reads uniquely mapped (mapping efficiency of 44.0% [43.3% – 44.7%]; Table A3).

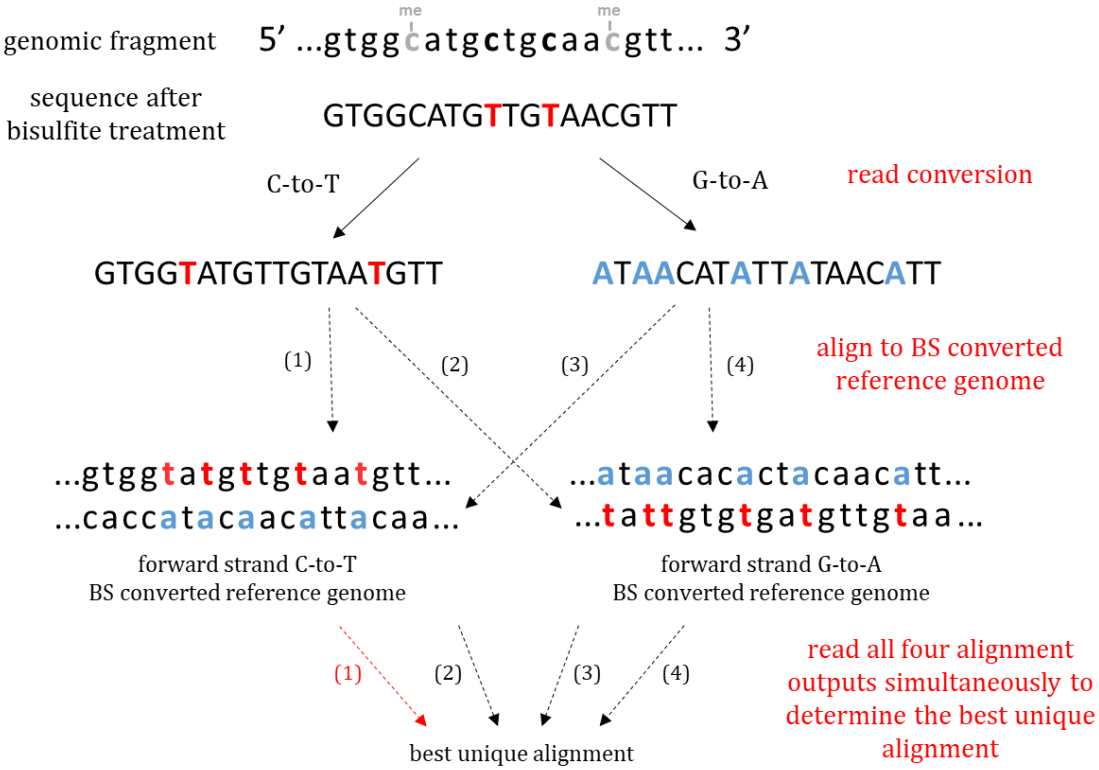


Figure 2: Methylation call used by Bismark (Krueger and Andrews, 2011). Reads from BS-seq are first converted *in silico* to its C-to-T and G-to-A versions to obtain the four possible strand versions. The four outputs are simultaneously aligned to the BS-converted reference genome to determine the best unique alignment (here, it corresponds to alignment 1). Note that the reference genome sequence illustrated here does not fully mirror the genomic fragment of interest (see Fig. 3). This figure was adapted from Krueger & Andrews (2011).

The BS-control sample showed 98.0% methylation for CpG context, 99.4% on CHG and 99.2% on CHH. That is, Bismark detected a high percentage of the cytosines as methylated in different contexts because they were not modified by the BS-treatment, implying the software was sensitive to detect cytosines in methylated state. The '*bismark_methylation_extractor*' tool and its '*--cytosine_report*' option from Bismark were used to extract the methylation call for every single cytosine in each context from the mapped files, generating a table with methylated and non-methylated counts for every genomic site for each individual. The values are reported in the Results section.

2.3.3.4. Controlling for genetic polymorphisms (SNPs) in methylation calls

The nature of BS-seq data offers some difficulties to its processing, as misleading methylation calls can arise due to the transformation from C/T (or G/A in the reverse strand) and subsequent alignment to reference genome. That is, a natural single nucleotide polymorphism (SNP) could be assigned as a differently methylated position and confound the results of this work (Fig. 3). To investigate this effect, the BS-treatment control sample (*i.e.* individual 17_0015) was first used to identify C/T and G/A SNPs, and then compared this list of sites to its BS-treated equivalent. The genetic data was processed following a pipeline often used in previous *Timema* sp. studies (*e.g.* Comeault *et al.*, 2016; Riesch *et al.*, 2017). Briefly, the good quality reads were aligned to the *T. cristinae* reference genome (1.3c2; Nosil et al. 2018) using bowtie 2.3.4.1 (Langmead & Salzberg, 2012), applying the paired-end argument. The mapped reads were sorted and indexed using SAMTOOLS 1.8 (Li *et al.*, 2009). SNPs were called following a custom *Perl* script (Comeault *et al.*, 2014), which uses SAMTOOLS *mpileup* and BCFTOOLS using the full prior, calling a variant only if the probability of the data was less than 0.5 under the null hypothesis that all samples were homozygous for the reference allele. Every SNP that presented Phred quality score equal to or greater than 20 (*i.e.* QUAL≥20) was retained. The filtering retained 3,487,275 SNPs, of which 12.1% were C/T SNPs and 12.2% corresponded to G/A SNPs, giving 846,998 potential genetic polymorphisms that could be confounders to the methylation counts. However, only 0.5% of all CpG sites called in the 17_0015 BS-treated sample overlapped with the list of C/T and G/A polymorphisms obtained from the BS-control.

Following this, to obtain a list of SNPs for the other individuals, newly acquired restriction site associated DNA sequencing (RAD-seq) data were used, as well as similar RAD-seq data from previous studies (Comeault *et al.*, 2014; Lindtke *et al.*, 2017; Riesch *et al.*, 2017). Complete information about samples and about the data processing will be detailed in Chapter 3 (section 3.3.5). After calling the SNPs using the same custom *Perl* script

cited above (Comeault *et al.*, 2014), only the genetic variants with Phred quality score equal to or higher than 20 were retained, where 460,757 C/T and 459,480 G/A transitions were listed. In addition, whole-genome accessions from 20 individuals from five populations included in this study (Soria-Carrasco *et al.*, 2014; Riesch *et al.*, 2017, Table A4) were retrieved from NCBI database (https://www.ncbi.nlm.nih.gov/). The data were processed similarly to the proceedings described above for the BS-treatment control sample, listing 7,547,750 C/T and 7,549,895 G/A SNPs. Merging the lists from RAD-seq and from whole-genome sequencing, 15,534,254 sites identified as C/T or G/A SNPs were obtained. From this list, 10.5% [10.4% – 10.6%] SNPs overlapped with CpG sites in the BS-treated samples, including the sample 17_0015, 4.8% in CHG, and 4.4% in CHH. These values are considerably higher than the proportion of SNPs overlapping in the comparison between BS and non-BS treatment sample (0.5%). As such, there was likely an overestimation of SNPs present in the methylation tables, but this conservative approach ensures most methylation polymorphisms (SMPs) were not genetic polymorphisms (SNPs). All the SNPs overlapping with methylation sites were removed aiming to reduce these confounding effects.

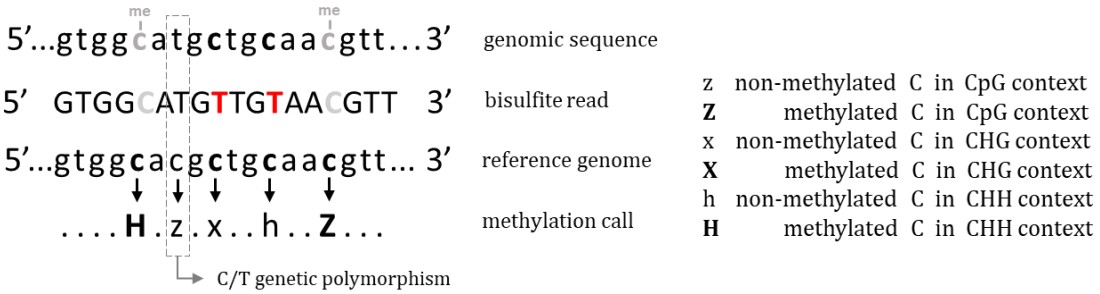


Figure 3: After determining the unique best alignment (Fig. 2), Bismark (Krueger and Andrews, 2011) calls the methylation variants for each locus. In this study, *T. cristinae* reference genome was used to do the alignments. Hence, misleading non-methylation calls might arise when C/T or G/A genetic polymorphisms are present in the individual's genomic sequence compared to the reference genome. In the example given here, a C/T SNP was interpreted as a non-methylated cytosine in CpG context. This figure was adapted from Krueger & Andrews (2011) and contains some alterations compared to the original one.

2.3.3.5. Final methylation tables

The final tables without the potential SNPs had mean coverage of 2.7 [2.5 – 2.9] reads per site, where 60.0% [57.9% – 62.1%] of the sites had coverage greater or equal to 2x; dropping to 36.1% [33.0% – 39.2%] for greater or equal to 3x ; and then 13.2% [10.3% – 16.2%] for coverage greater or equal to 5x per site. That is, the final read coverage was much reduced after all the processing steps: filtering low quality reads; subsampling the reads to a maximum of 24 million reads each sample to minimize batch effects (see section 2.3.3.2); mapping to the reference genome (mapping efficiency of 44.0%); and removing the potential C/T and G/A SNPs. The low coverage could compromise some of the analyses using methylation data, as they depend on comparisons between methylated and non-methylated cytosine counts in a specific locus (Lea *et al.*, 2017). On the other hand, despite the low coverage and the reduced number of covered sites (and the errors arising from these numbers), the processing steps cited above circumvented some potential confounders to the interpretation of the data patterns (*e.g.* low-quality reads, sequencing batch effects, SNPs, etc.). The data was analysed throughout this dissertation considering the limitations in the data coverage, and the best possible approaches to handle these data. The individual tables with methylation information were filtered using a minimum threshold of 5 reads covering the sites, which is higher than the threshold used in some other studies (*e.g.* Cunningham *et al.*, 2015; Glastad *et al.*, 2016). Sites with coverage outliers above the 99.9th percentile were removed to avoid PCR bias (*i.e.* above 60 reads). After all the filtering, the mean number of sites with cytosines in CpG context was 2,193,306 [2,128,581 – 2,258,031], 2,839,901 [2,761,571 – 2,918,231] in CHG context, and 12,801,094 [12,518,520 – 13,083,668] in CHH context, averaged across all 24 individuals.

For each sample, the methylation levels were calculated for each site as the total number of unconverted C (*i.e.* methylated cytosines) divided by the total number of reads mapped to the site. The methylation levels were estimated separately for each cytosine context (*i.e.* for CpG, CHG, and CHH independently). The methylation status (*i.e.* methylated

versus non-methylated) was estimated at each site by comparing the proportion of methylated reads (*i.e.* unconverted cytosines) to a binomial distribution. For this, the number of unconverted cytosines at each site was used as successes and the coverage as trials. The non-conversion rates of unmodified cytosines obtained from the non-methylated lambda phage were used as probability of success. In other words, a site would be considered methylated if the proportion of unconverted cytosines could not be expected by chance (*p-value* < 0.01, using a Benjamini–Hochberg FDR correction at 1%). If the proportion could be expected by chance (*i.e.* similar to the proportions found at the non-methylated phage), it would be considered non-methylated (Glastad *et al.*, 2016; Libbrecht *et al.*, 2016; Standage *et al.*, 2016). The results found using this method were very similar to those obtained when a threshold was used to determine methylation status. A site was considered as significantly methylated if the percentage of methylated cytosine was higher than 20% (mC > 20%). This definition requires at least two unconverted C containing reads to call a site methylated in the minimum coverage of five reads. This way, a single T → C Illumina sequence error would not result in a spurious methylated site. This approach has been used in previous studies (*e.g.* Wang *et al.*, 2013), and in this work it was consistent with the binomial approach. All the reported statistics were performed using R (3.3.1; R Core Team 2016).

2.3.4. Annotation

I used the *T. cristinae* genomic annotation table (Villoutreix *et al. in prep*) to obtain information about DNA methylation levels in different genomic features. Only the genes with InterPro or GO accessions (InterPro EMBL-EBI; Gene Onthology, UniProt) were selected, retaining 19,383 genes. In the annotation tables, the genes begin at the start codon and finish at the stop codon. Thus, information about untranslated regions (UTRs) is not represented in the present data. The upstream and downstream regions around the gene were defined as 1kb 5' and 3' from the gene, following rules that are widely used in the

literature (Wang *et al.*, 2013; Cunningham *et al.*, 2015). The remaining regions were considered intergenic. For some analyses, the mean methylation percentage across sites and across samples were considered for each annotated element.

To identify "methylated" and "non-methylated" genes, first the probability of a mCpG occurring within a gene was calculated. This was done by dividing the total number of mCpG sites within all genes by the sum of all reads covering sites within genes (*i.e.* including non-converted and converted cytosines). Then, a binomial test was performed using the number of mCpGs at a specific site and its coverage, using the probability estimated above. These results were then corrected for multiple testing using a Benjamini–Hochberg FDR correction at 1%. Only genes with at least five mapped cytosine sites were reported (Cunningham *et al.*, 2015; Glastad *et al.*, 2016).

2.3.5. Methylation enrichment on genomic features

Enrichment analyses were performed to estimate the likelihood a genomic feature (*e.g.* exons, introns, etc.) presents higher or lower methylation levels compared to background genome-wide levels. In summary, the analyses used the mean levels of methylation in single CpGs sites found in at least 12 samples to estimate (1) the number of methylated sites found in a certain genomic feature and (2) its enrichment in methylated sites, calculated using:

$$\frac{(N_{rand}/mCpG)}{(N_{bg}/nCpG)} \quad (1)$$

For example, to calculate the methylation enrichment in exons, one estimate the proportion of methylated sites within exons using the number of methylated CpGs site within exons (N_{rand}) and the total number of methylated CpGs in the genome ($mCpG$). Then, one divides it over the proportion of CpG sites in exons, using the total number of CpGs sites found within exons (N_{bg}) and the total number of CpG sites across the genome ($nCpG$). A *p-value* for the enrichment can then be estimated using the Fisher's Exact Test to statistically

compare the number of methylated CpGs in each genomic feature with the background values in the genome. For comparison, I estimated a null distribution of the expected number of mCpGs in a genomic feature by randomizing their position on the genome, then computing how many of those were found in each genomic feature at each iteration. The null distribution was generated after 10,000 iterations of randomization. All analyses were performed using R (3.3.1; R Core Team 2016).

2.3.6. Gene Ontology (GO) enrichments

I generated a list of GO terms that were over-represented in genes with methylation information using the R package *TopGO* (v 2.34.0). This analysis used 8,472 genes, as they presented information about methylation across all 24 samples. The analysis was performed using genes that were consistently methylated across all individuals compared to the remaining genes, and using genes that were consistently non-methylated. Fisher's Exact Test was used to calculate the significance of the enrichment, coupled with a weight algorithm. This algorithm uses a hierarchical approach to compute the *p-value* of a GO term, conditioning the process based on the neighbouring terms (*i.e.* it accounts for GO topology). Hence, the tests are not independent from each other, which means the multiple testing theory does not apply. Given this, the authors of the R package attest the *p-values* are internally corrected and do not need further correction for multiple testing (Alexa and Rahnenfuhrer, 2019).

2.3.7. Transposable elements (TEs)

I used the *T. cristinae* RepeatMasker database (Villoutreix *et al. in prep*) to extract information about transposable elements. The analyses were focused on families of transposable elements that contained more than 400 copies across the *T. cristinae* genome (following the method in Libbrecht *et al.*, 2016), including DNA transposons; long terminal

repeats (LTR) retrotransposons; non-LTR retrotransposons; and Penelope-like elements (PLE; Table A5). The mean methylation across TEs was estimated in sites that were present in at least 12 samples and its enrichment, following the same procedures described at section 2.3.5. These estimates were performed using all transposons, but also in those found within genes and in the intergenic regions separately. In addition, the analyses were repeated for each TE family considered here. Some of the families cited in Table A5 were not used in the analyses because they were not represented in the table with CpG sites present in at least 12 samples, or they presented a very low number of CpG sites (*e.g.* SINE presented only 26 CpG sites). All analyses were performed using R (3.3.1; R Core Team 2016).

2.4. Results

2.4.1. Identification of T. cristinae DNA methyltransferases

The *T. cristinae* genomic annotation presented a DNMT1 replication foci domain, and three genes with predicted proteins characterized by C-5 cytosine methyltransferase (GO:0008168; IPR001525). Two of these queries were identified as DNMT1-like proteins (Table 1). One of them existing in LG3 (gene g34132.t1) encodes a protein with 465 amino acids, and it is composed by the Dcm domain, which is a site-specific DNA-cytosine methylase. It shared strong similarity with the DNMT1 of a termite species *Zootermopsis nevadensis* [Isoptera: Termopsidae], which is also a hemimetabolous insect (Table 1). The second one (g25566.t1) encodes a protein with 200 amino acids, presenting the Dcm and the BAH domains (bromo-adjacent-domain). It had a lower BLASTp score, with the best matches corresponding to DNMT1-like proteins in more distantly related organisms, such as a wasp species (*Trichogramma pretiosum* [Hymenoptera: Trichogrammatidae]) or to a spider (*Parasteatoda tepidariorum* [Aranae: Theridiidae]). The third query had the best match to the DNMT2 enzyme, which does not present DNA methyltransferase activity (Goll

et al., 2006). Thus, none of the candidate queries was matched to the *de novo* methyltransferase (*i.e.* DNMT3) in the *T. cristinae* annotation.

Following these results, the tBLASTn analyses using DNMTs of a few insect species and *T. cristinae* database on its reference genome revealed consistent results for the DNMT1 enzyme (Table 2). The three analyses pointed the same genomic region in linkage group 3 (LG3) with very significant BLAST scores (*e.g.* low *E-value*). On the other hand, the analyses using the *de novo* DNA methyltransferase (DNMT3) from other insects had very high low matching scores, and output different genomic regions. In other words, there was not a significant match between the DNMT3 and a specific region in the *T. cristinae* genome. These results, added to the finding regarding the reasonably good quality of the *T. cristinae* genome assembly, suggest *T. cristinae* does not have this enzyme.

Table 1: Best results from BLASTp using the putative proteins related to C-5 methyltransferase activity in *T. cristinae* genome annotation. None of the putative proteins had a result related to the *de novo* DNMT3.

Gene	Genomic region	Description	Max score	*E-value*	Ident	Organism
g34132.t1	LG3_scaf715	DNMT1	778	0	79%	*Zootermopsis nevadensis* [Isoptera: Termopsidae]
g25566.t1	LGNA_scaf2537	DNMT1-like	176	7e-49	42-47%	*Parasteatoda tepidariorum* [Aranae: Theridiidae]; *Trichogramma pretiosum* [Hymenoptera: Trichogrammatidae]
g3428.t1	LG7_scaf763	DNMT2	436	1e-149	58%	*Zootermopsis nevadensis* [Isoptera: Termopsidae]

The *E-value* describes the number of hits that are expected to be retrieved by chance when searching a database of a particular size. It decreases exponentially as the score of the match increases, so that the lower the *E-value* the more significant the match is (Altschul *et al.*, 1997).

Table 2: Best results from tBLASTn using the described DNA methyltransferase proteins in a few representatives of Insecta clade: *Apis melifera*, *Nasonia vitripensis*, and *Zootermopsis nevadendis*. The lower *E-value* score the higher is the match between the query protein and the genomic database (*T. cristinae* reference genome).

Protein	***Organism***	***Score***	***E-value***	***Ident***	***Genomic region***
DNMT1	*A. melifera*	231	1e-87	58%	lg3_scaf715
	N. vitripensis	224	3e-57	69%	
	Z. nevadensis	778	0	79%	
DNMT3	*A. melifera*	35	6.6	26%	lg9_scaf527
	N. vitripensis	37	1.8	34%	lg12_scaf2191
	Z. nevadensis	37	1.5	34%	lgNA_scaf1772

2.4.2. General patterns

The mean proportion of methylated cytosines across the 24 samples was 2.1% [2.0% – 2.2%] (mean [95% CI]). Methylation was found primarily on CpG dinucleotides, as 80.2% [79.0% – 81.4%] of methylated cytosines were on CpG context, 3.8% [3.6% – 4.0%] on CHG, and 16.0% [15.0% – 17.0%] on CHH context (Fig. 4). Considering each context separately, the mean proportion of cytosines that were methylated across the genome was 14.0% [13.3% – 14.7%] in CpG, 0.5% on CHG and 0.5% on CHH (Table A3). The predominance of methylation in CpG context was expected, as this is the most prevalent DNA methylation context found among animals (Suzuki and Bird, 2008). Thus, the main results reported here refer to CpG context, unless the other contexts are mentioned. Among the methylated CpGs, 82.3% [81.8% – 82.8%] had methylation levels above or equal to 50%. The numbers cited above were estimated independently for each individual, and then the mean was obtained across samples with its corresponding 95% confidence interval to estimate the general pattern.

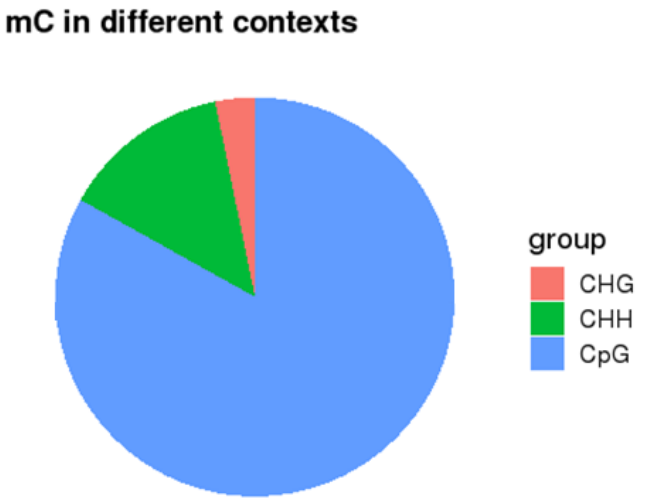


Figure 4: Mean proportion of methylated cytosines in each context across the 24 samples. Methylation is found primarily in CpG context, corresponding to 80% of all methylated cytosines.

2.4.3. Distribution of DNA methylation across genome

As mentioned in the previous section, a mean of 14.0% of genomic CpGs were methylated across the samples. However, these methylated sites were not evenly distributed throughout the genome. Among the mCpGs, 19.3% [18.8% – 19.8%] were located in exonic sites, and 32.7% [32.4% – 33.0%] were in intronic sites (whereas 7.0% [6.8% –7.2%] and 12.9% [12.7% – 13.1%] of all CpGs are exonic and intronic, respectively), which shows a preferential target of DNA methylation in gene bodies (Fig. 5, Table A6). Half of the genes in *T. cristinae* were methylated, where 50.2% were methylated in at least half of the samples and 45.6% were methylated in all samples (Table 3).

Both exons and introns had a significant enrichment of methylation levels compared to genomic background levels (*p-value* < 2.2e-16, Fisher's exact test; Table 4, Fig. 5-6). In particular, exons showed considerable proportion of methylated sites, with 49.5% [48.0% – 51.0%] of CpGs being methylated (Fig. 6A), and had mean methylation levels of 39.4% [38.0% – 40.8%] (Fig. 6B). In comparison to exons, introns showed a marginally lower proportion of CpGs that were methylated (44.6% – 47.0%), with mean methylation levels of 35.8% [34.7% – 36.9%]. In fact, the difference in methylation levels between exons and their surrounding introns was not very pronounced (Fig. 7). Although a higher proportion

of CpGs are methylated in exons, introns generally presented more mCpGs than exons, around 1.7x as many (129,319 [121,319 – 137,319] mCpGs in introns and 76,675 [70,417 – 82,933] in exons). This is possibly due to the fact introns are normally larger than exons in *T. cristinae* (mean 2,305bp [2,285bp – 2,325bp] and 231bp [229bp – 233bp], respectively, considering all genes used in this study). The regions flanking the genes (*i.e.* up to 1kbp at 5' and at 3' of the genes, respectively the upstream and downstream regions) also tended to be enriched in DNA methylation, although in lower levels compared to the gene body values (*p-value* < 2.2e-16, Fisher's exact test; Table 4, Fig. 6).

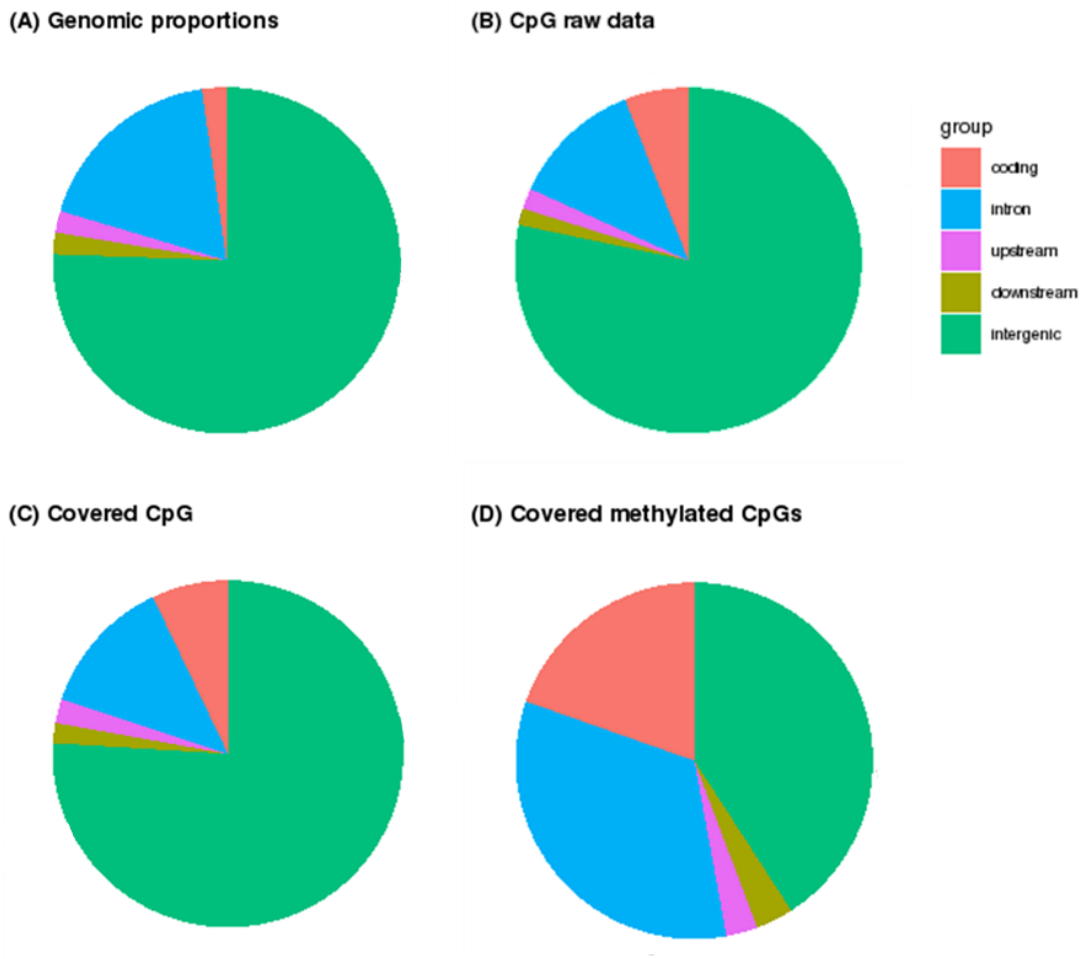


Figure 5: Proportion of sites in genomic features. Numbers of sites were estimated independently for each individual, and then averaged (see Table A6). **(A)** Proportion of features across all genome; **(B)** proportion of CpGs present in the genomic features using the raw data; **(C)** after selecting for minimum of 5 and maximum of 60 reads per site; **(D)** and proportion of CpGs in each feature considering only methylated cytosines.

Table 3: Number of genes and their methylation status covered by a minimum of 5 reads and maximum of 60 reads per site. Methylation status (*i.e.* methylated or non-methylated) was estimated averaging the percentage of methylated CpG in the gene body (*i.e.* using both exons and introns), using sites found in at least one sample, in at least in 12 samples, or in all samples. The total number of annotated genes used in this study was 19,383.

	1 sample	**12 samples**	**24 samples**
Total number of genes	17,929	14,656	8,554
Methylated genes	9,997 (55.8%)	7,364 (50.2%)	3,897 (45.6%)

Table 4: Enrichment in methylation status in the genomic features in CpG sites found in at least 12 samples. All genomic features studied here were found to be more frequently methylated than background levels (*p-values* estimated using Fisher's exact test). Levels of methylation were averaged between individuals for each site (895,343 sites analysed). Similar results are found when the individuals are analysed separately (Tables A7-A10).

Genetic feature	**Number of sites**	**95% Quantiles (null dist.)**	**Enrichment**	***p-value***	
exon	34,291	[14,446–14,846]	2.3	<2.2e-16	***
intron	52,292	[23,081–23,568]	2.2	<2.2e-16	***
upstream	6,731	[5,385–5,639]	1.2	<2.2e-16	***
downstream	7,031	[4,706–4,944]	1.5	<2.2e-16	***

There was a noticeable trend of increasing methylation levels from 5'→ 3', especially within the gene body (Fig. 7). That is, when all genes are considered, exons and introns tend to have higher methylation levels at features that are more distant from the initiation site. The same pattern is found in the 1kbp around the genes, as the genic downstream region was generally more methylated than the upstream region (Fig. 6-7). However, this trend is not as substantial when the genes are separated according to their number of exons (Fig. 8). Genes with more than five exons (*i.e.* long genes) were more methylated those with up to four exons (*p-value*< 2.2e-16; unpaired t-test; Table A11). Thus, the considerable increase in methylation from 5'→ 3' is confounded with the general high methylation levels in genes with more exons. Finally, outside the genes, the DNA methylation levels were generally lower the more distant the regions are from genes (Fig. 9).

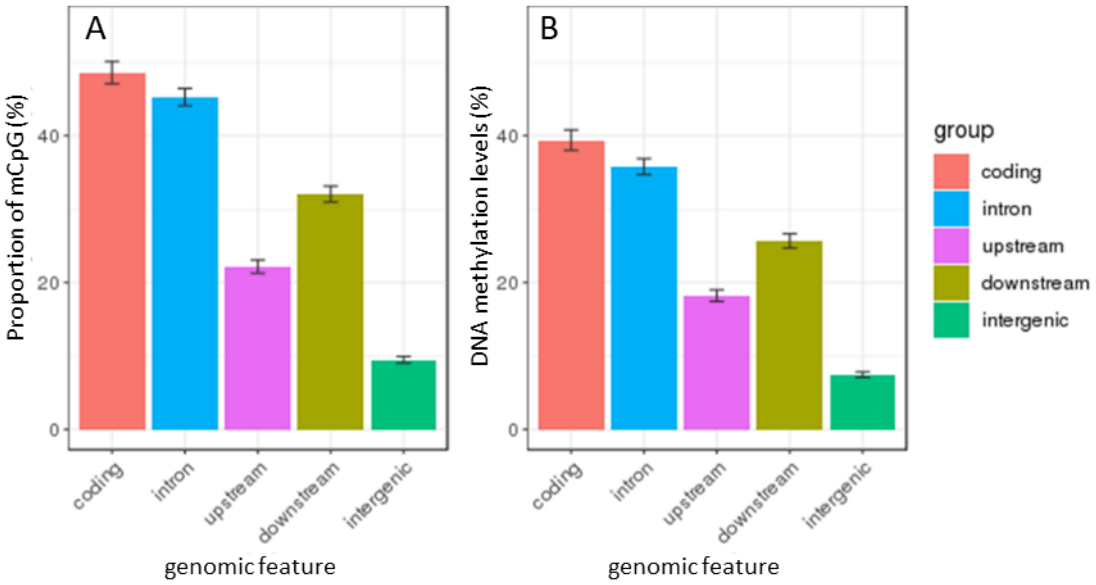


Figure 6: CpG methylation in each genomic feature across the 24 samples. **(A)** Proportion of methylated CpGs in covered sites in different genomic features, and **(B)** mean methylation levels at covered CpGs. Error bars represent 95% confidence interval.

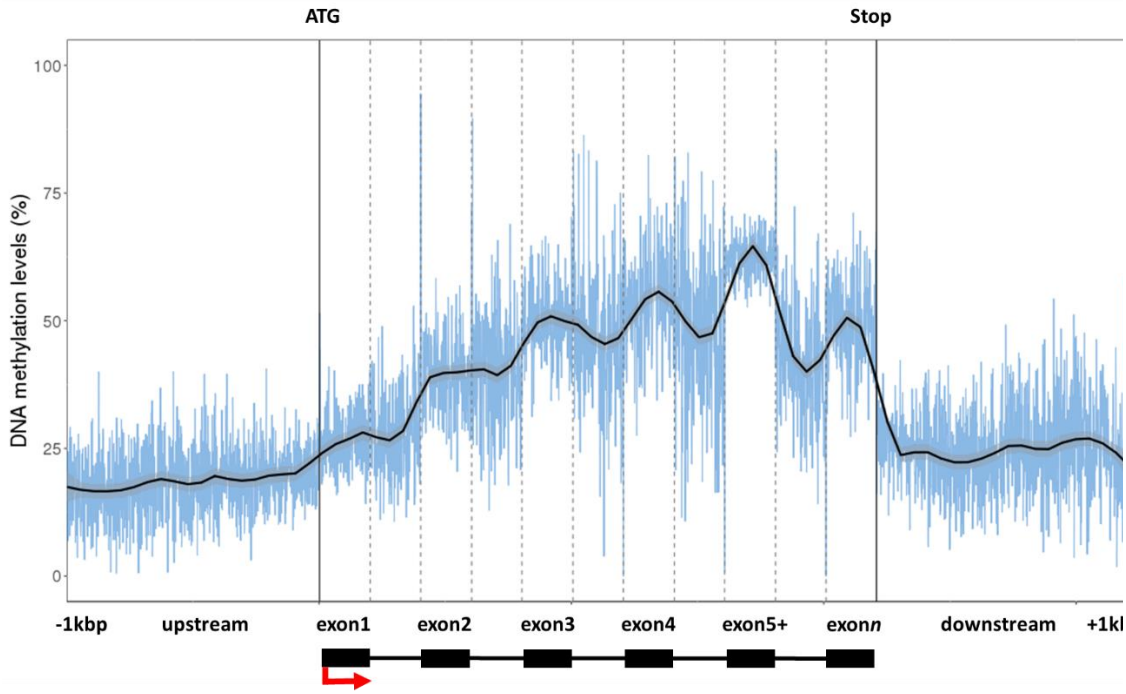


Figure 7: DNA methylation levels in genes and their flanking regions. The graph represents the 5' downstream flanking region, the multiple exons and introns, and the 3' downstream region. The graph shows mean methylation levels estimated at CpG sites found in at least 12 samples, in both methylated and non-methylated genes (n=14,656 genes, see Table 3). The x-axis represents nucleotide position from the beginning or from the end of the genomic feature. To compare exons and introns of different genes, I used the mean methylation in the first 100bp at 5' and the last 100bp 3' of each exon and each intron (following Hunt *et al.*, 2013; Glastad *et al.*, 2016) A blue line was drawn to connect the means in each position, and a black smooth line was plotted to represent the overall trend using the method 'loess' (standard error shown in grey).

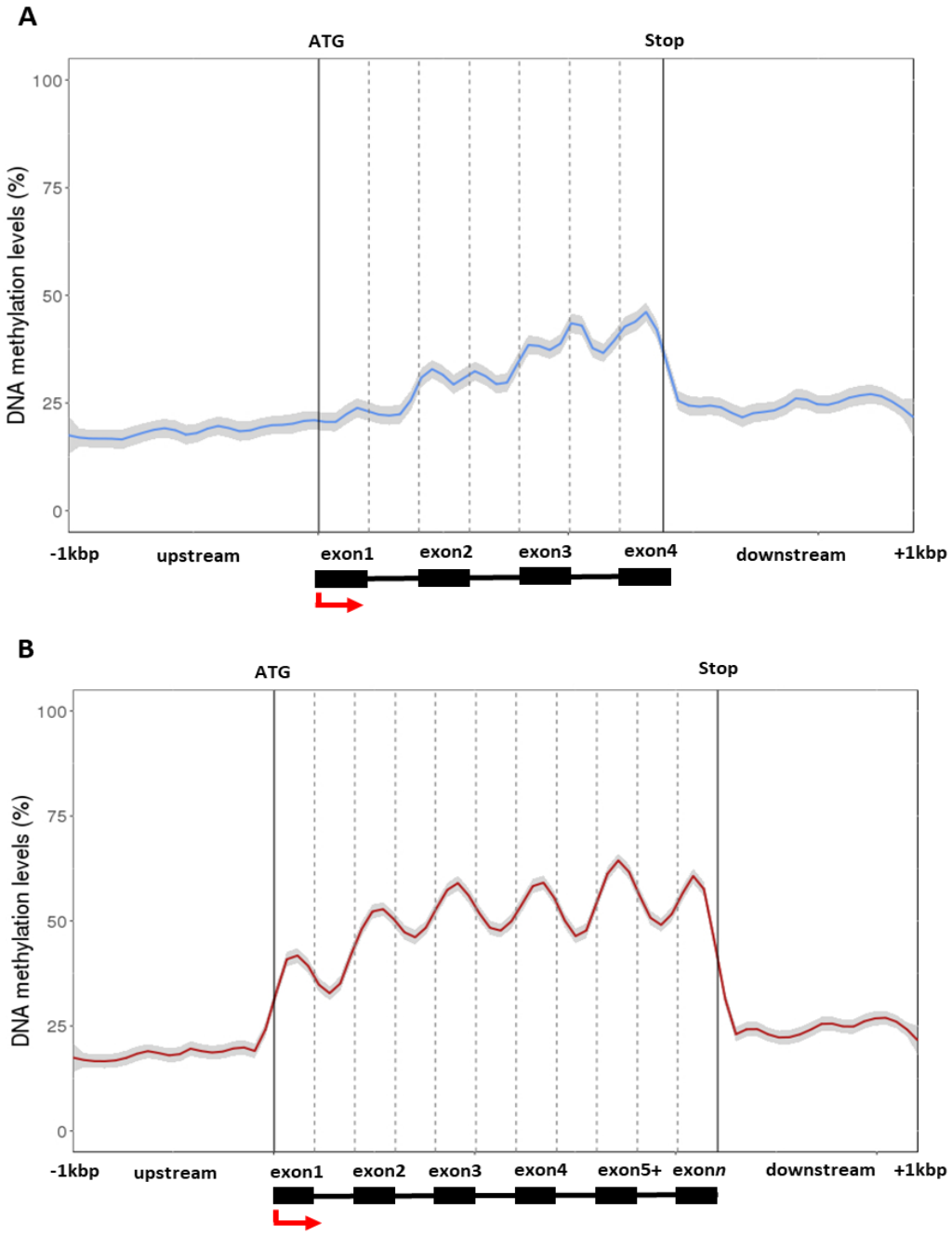


Figure 8: DNA methylation levels in the 5' downstream flanking region, in the multiple exons and introns, and in the 3' downstream region, using both methylated and non-methylated genes. The graph shows mean methylation levels estimated at CpG sites found in at least 12 samples (n=14,656 genes, Table 3), with a smooth line plotted to represent the overall trend using the method 'loess' (standard error shown in grey). The x-axis represents nucleotide position from the beginning or from end of each feature. To be able to compare exons and introns of different genes, I used the mean methylation in the first 100bp at 5' and the last 100bp 3' of each exon and each intron (following Hunt *et al.*, 2013; Glastad *et al.*, 2016). Differently sized genes were represented in separated graphs, divided in (**A**) smaller genes with up to four exons (n=7,529), and (**B**) longer genes, with five or more exons (n=7,127). Longer genes present higher methylation levels, and more contrasting difference between exons and introns (see Table A11 for unpaired t-tests on the data used to generate this graph).

2.4.4. GO terms enriched in methylated and in non-methylated genes

The enrichment analyses in consistently methylated genes across all samples (n=3,866) showed an over-representation of functional terms related to fundamental cellular processes (*e.g.* involved in protein metabolic processes, RNA binding and RNA metabolic processes, biosynthesis of nucleotides, and protein folding; Tables A12-A13). Some GO terms were related to methylation activity, including the one responsible for methyltransferase (GO:0008168). In addition, some functional terms were related to gene expression and transcription. Meanwhile, non-methylated genes (n=3,176) were generally associated with terms related to dynamic functions, especially related to signalling and reception pathways. For example, there were some GO related to olfaction and perception of smell. In fact, many of those GO terms are related to nervous system processes and the components associated to it, such as neurotransmitters' receptor activity and processes related to ions channel transport (*i.e.* possibly related to membrane potential difference). In addition, some GO functions were related to metabolism of the chitin and development of the cuticle.

2.4.5. Transposable elements

The mean methylation levels of all transposable elements used in this study was 15.0% [14.9% – 15.1%] (Table 5). TEs were found to be less methylated than the genomic background levels, especially when only regions within the genes are analysed (Table 5). That is, although transposons tend to be more methylated within genes (24.5% [24.2% – 24.8%]), they are impoverished in methylation compared to other genic CpGs. On the other hand, TEs seem to be relatively enriched in methylation in intergenic regions (Table 5). Thus, transposons seem to have basal methylation levels, although it is significantly higher in intergenic regions and lower compared to those located within genes, which tend to be enriched in DNA methylation.

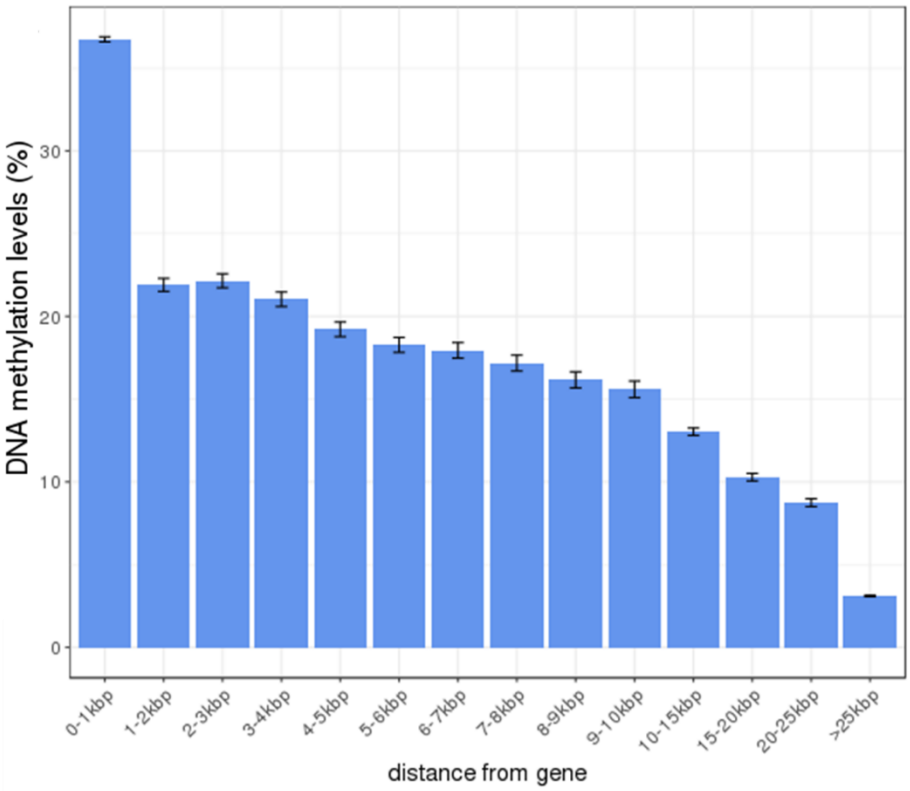


Figure 9: Mean methylation levels in intergenic regions according to distance to the closest gene across 24 samples. Error bars represent 95% confidence intervals. Methylation levels tend to be lowered the more distant the sites are from the gene (R^2=0.07, *p-value* < 2.2e-16, linear models).

I analysed each transposon family separately to investigate the differences in methylation patterns between them, both within genes and in intergenic regions. Some DNA transposons were impoverished in methylation independently of the genomic context, such as *Helitron*, *MuDr*, *Polinton* (Tables A14-A15). However, some of them seemed to be preferentially targeted for methylation in both genic and intergenic regions, such as *PiggyBac*, *Mariner*, and *Sola* (Tables A14- A15). A Penelope-like element (PLE) was also enriched in methylation in *T. cristinae* compared to its background levels in any genomic context. Curiously, retrotransposons were found to either always be impoverished in methylation (*e.g. HERV*, an LTR retrotransposon; *Jockey* a non-LTR retrotransposon), or were enriched only when compared to the intergenic baseline levels.

Table 5: Transposable elements and methylation enrichment in all genomic contexts, or only in genic or intergenic regions using at least 12 samples. Enrichment was estimated comparing the number of methylated CpG sites compared to the expected null distribution estimated based on randomizations (section 2.3.5). *P-values* were calculated using Fisher's exact test.

	CpG sites	**mean**	**mCpG sites**	**95%Quantiles (null dist.)**	**Enrich.**	***p-value***	
All TEs	132,089	15.0% [0.1%]	27,402	[27,967–28,506]	0.97	1.2e-09	***
TEs genic	30,465	24.5% [0.3%]	10,001	[14,573–14,890]	0.68	<2.2e-16	***
TEs intergenic	101,624	12.1% [0.1%]	17,401	[13,253–13,635]	1.29	<2.2e-16	***

2.5. Discussion

2.5.1. Two copies of DNMT1 and absence of DNMT3 in T. cristinae

Among the enzymes with C-5 cytosine methyltransferase activity in *T. cristinae* functional annotation, two copies of the maintenance DNMT1 gene were found: one with high identity to the same enzyme in *Z. nevadensis*; and the other with a lower identity score to more distantly related individuals. However, the latter presented one domain, the bromo-adjacent homology domain, that is characteristic of DNMT1 enzymes, which is absent in the other form. Gene duplication can be a source of gene novelty in genomic evolution (Lynch, 2002). Thus, it is possible these DNMT1 paralogs exert different activities, maintaining methylation differently in space and time. Duplicates of DNMT1 were also found in pea aphid (Walsh *et al.*, 2010), and in the Hymenoptera clade, and the different forms seem to face weak divergent selection (Bewick *et al.*, 2017). Alternatively, it is possible the DNMT1-like enzyme was degenerated and lost its function in *T. cristinae*, although more studies are required to depict the functional role of those two DNMT1 enzymes.

Curiously, an enzyme matching the *de novo* DNMT3 in *T. cristinae* was not found. This pattern was also found in other insects, with comparative analyses showing DNMT3 was possibly lost numerous times during the evolutionary history of insects (Bewick *et al.*, 2017;

Provataris *et al.*, 2018). These evidences suggest that DNMT3 might be dispensable for DNA methylation in some clades, including *Timema*. This implies that either DNMT1 has acquired some *de novo* methyltransferase functionality in insects, or that the loss of DNMT3 enzyme activity is compensated by DNMT1 or by other molecular pathways. However, DNMT1 lacks protein domains associated with *de novo* methyltransferase activity, such as the PWWD domain (Qiu *et al.*, 2002; Bewick *et al.*, 2017). Indeed, Mitsudome *et al.* (2015) showed DNMT1 in *Bombyx mori* preferentially methylated the hemimethylated DNA, suggesting it functions primarily as a maintenance methyltransferase. Alternatively, there is the speculation about the maintenance of DNA methylation status being sufficient for it to persist across generations (Glastad *et al.*, 2018). The latter hypothesis contradicts the overall demethylation and reset of parental methylation levels during gametogenesis observed in vertebrates (Law and Jacobsen, 2010). Whether DNA methylation reprogramming exists in insects is largely unknown, but some evidence suggests there is stable inheritance of methylation status between generations in *Nasonia* wasps (Wang *et al.*, 2016). The enzymatic functions of methyltransferases are not known in insects, but rather inferred from studies in vertebrates and plants. The fact that vertebrates and insects differ in the number and types of methyltransferases points to the possibility of novel roles in the insects' DNMTs (Wang *et al.*, 2016). In any case, further investigations in insects should be carried to elucidate the mechanisms by which some insects can establish DNA methylation in the absence of DNMT3.

2.5.2. Majority of methylated cytosines is in CpG context

A great proportion of the methylated cytosines in *T. cristinae* was found in the symmetric context of cytosines followed by guanines (*i.e.* CpG context). This pattern is ubiquitous in animals, which is generated by the action of DNMT1 (Suzuki and Bird, 2008; Feng *et al.*, 2010). It has been suggested that non-CpG methylation in animals is mainly found in embryonic cells, but not somatic ones, and is generated as a by-product of the *de*

novo DNA methyltransferase 3a (DNMT3a) activity (Ramsahoye *et al.*, 2000). Thus, it is curious there was some methylation found in CHG and CHH contexts in adult *T. cristinae*, given this species does not produce the *de novo* DNMT3 enzyme. This suggests a different molecular pathway culminating in non-CpG methylation. Very few studies in insects report non-CpG methylated sites because either they were not detected or they were not given attention. When they are reported, they tend to be much reduced in numbers compared to CpG methylation (*e.g.* Bonasio *et al.*, 2012; Cunningham *et al.*, 2015). However, the distribution patterns and particular functions for methylation in these contexts remain largely unexplored.

2.5.3. DNA methylation levels are high in T. cristinae *genome and differentially distributed*

On average, 14.0% of the CpGs were methylated across the *T. cristinae* samples. This proportion is relatively high when compared to the great majority of insect species described in the literature (Bewick *et al.*, 2017). In fact, the majority of DNA methylation studies in insects have focused on understanding the role of this epigenetic mechanism in the development of castes and division of labour in eusocial insects (*e.g.* Kucharski *et al.*, 2008; Standage *et al.*, 2016; Bewick *et al.*, 2017; Glastad *et al.*, 2017). This way, the literature tends to be biased towards the Holometabola superorder of insects, which includes the Hymenoptera clade. When representatives of this superorder exhibit any trace of DNA methylation, it is normally at very low levels (Bewick *et al.*, 2017). Here, I showed *T. cristinae* mean methylation levels in CpG were much higher than has been reported in Holometabola insects (Table 6). On the other hand, the results presented here are consistent with studies of other "Hemimetabola", which tend to describe higher methylation levels (Provataris *et al.*, 2018).

Following the same trend as in other insects (Zemach *et al.*, 2010; Bewick *et al.*, 2017), DNA methylation in *T. cristinae* seems to target the gene body, where methylation is considerably enriched compared to intragenic levels. Both exons and introns are highly

methylated in *T. cristinae*, contrasting the pattern found in Holometabola insects where exons are the main genomic target for DNA methylation (Wang *et al.*, 2013; Libbrecht *et al.*, 2016; Standage *et al.*, 2016; Glastad *et al.*, 2017). Thus, although exons are generally more methylated than introns, this trend is less pronounced in *T. cristinae* compared to Holometabola insects (Fig. A1). Another difference is the increased methylation from 5'→ 3' in the gene body, with more elevated DNA methylation levels the longer the gene is (*i.e.* genes with more exons are more methylated). A classic explanation of gene body methylation is that it reduces transcriptional noise by preventing initiation of transcription outside transcription start sites (Bird, 1995). As longer genes are likely more prone to this noise, it is possible their high methylation levels are acting to supress spurious transcription in *Timema*, assuring the integrity of the genes' function. A similar trend to increase methylation levels from 5'→ 3' was found in termites, suggesting a generality among "Hemimetabola" insects (Fig. A2, Glastad *et al.*, 2016). However, the steep increase shown in the study might be due to the high general methylation levels in longer genes, as observed in this present study. Differences between "Hemi" and Holometabola were shown to be consistent by a study that used normalized CpG content on 53 arthropod species (Provataris *et al.*, 2018). These results collectively suggest that the *T. cristinae* methylome profile, as well as that of other hemimetabolous insects, is relatively underived during the evolution of insects and could point to an ancestral loss of DNA methylation occurring in Holometabola (Provataris *et al.*, 2018). In general, *Timema*'s methylation is more similar to the tunicate *Ciona intestinalis* methylation patterns (31.1% CpG methylated; Feng *et al.*, 2010; Zemach *et al.*, 2010) than to the reduced and restricted to exons DNA methylation shown in holometabolous insects. In some aspects, such as the generalized methylation in both exons and introns and the increased methylation towards the 3' end of the gene, where the gene is methylated *Timema* methylation patterns resemble those of vertebrates (Table 6, Fig. A2 B; Glastad *et al.*, 2016).

Table 6: List of animal organisms and their respective proportion of methylated cytosines in CpG context (mCpG). *Timema cristinae* shows mCpGs at 14%. The clade Polyneoptera have incomplete metamorphosis, previously called "Hemimetabola".

Clade		Organism	mCpG	Reference
Insecta (Holometabola)	*Coleoptera*	*Tribolium castaneum* (flour beetle)	0.0%	Schulz *et al.*, 2018
	Diptera	*Drosophila melanogaster* (fruit fly)	0.0%	Zemach *et al.*, 2010
	Hymenoptera	*Apis melifera* (honeybee)	0.9%	Feng *et al.*, 2010
		Camponotus floridanus (carpenter ant)	0.3%	Bonasio *et al.*, 2012
		Cerapachys biroi (clonal raider ant)	2.1%	Libbrecht *et al.*, 2016
		Harpegnathos saltator (jumping ant)	0.2%	Bonasio *et al.*, 2012
		Nasonia vitripensis (parasitoid wasp)	1.6%	Wang *et al.*, 2013
	Lepidoptera	*Bombyx mori* (silkworm)	0.1%	Xiang *et al.,* 2010
Insecta (Polyneoptera)	*Orthoptera*	*Locusta migratoria* (migratory locust)	11%	Wang *et al.*, 2014
	Phasmatodea	*Medauroidea extradentata* (Annam stick insect)	12%	Krauss *et al.*, 2009
	Isoptera	*Zootermopsis nevadensis* (Nevada termite)	12%	Glastad et al., 2016
Ascidiaceae		*Ciona intestinalis* (sea squirt)	31%	Feng *et al.*, 2010
Actinopterygii		*Danio rerio* (zebrafish)	80%	Feng *et al.*, 2010
Mammalia		*Mus musculus* (house mouse)	74%	Feng *et al.*, 2010

2.5.4. Enriched GO terms are generally similar to those in other insects

Although the individuals studied here came from different populations and different environmental contexts, some genes were consistently methylated or consistently non-methylated across all samples. While methylated genes are normally associated with housekeeping functions and activities inside the cell, the non-methylated genes were related to dynamic functions, generally involving the cells' external environment (especially receptor and signalling pathways). A similar pattern was discovered in other insects, with

methylation being preferentially targeted to genes broadly expressed across tissues (Glastad *et al.*, 2016; Glastad *et al.*, 2018). Thus, despite the differences in methylation patterns across insects, the methylation targets in functional terms seem to be conserved. The non-methylated genes in the termite *Z. nevadensis* were also enriched in GO terms related to signalling receptor activities, although not necessarily related to neurotransmission or olfactory functions, but with circadian behaviour (Glastad *et al.*, 2016). However, few studies reported the GO terms that are over-represented in non-methylated genes.

2.5.5. TEs are normally depleted in methylation

Our results revealed a general impoverishment of methylation in transposable elements in *T. cristinae*, close to baseline levels (15.0% and 17.0% respectively, in sites present in at least 12 samples), as previously shown in other insects (*e.g.* Bonasio *et al.*, 2012; Wang *et al.*, 2013; Cunningham *et al.*, 2015; Libbrecht *et al.*, 2016; Glastad *et al.*, 2018). However, this study showed different methylation patterns depending on the TE family. Interestingly, some families exhibited a higher percentage of methylated CpGs when only the intergenic regions were considered. That is, these transposons seem to display a basal methylation level that is higher than the intergenic background levels. In addition, some DNA transposons (*i.e.* sequences that do not require an RNA intermediate) are enriched in methylation in both intergenic and genic regions. One of them, the *Mariner* element, was hyper-methylated in the ant *C. floridanus*. This transposable element is widespread in insects and it is commonly used to mutate genes and transfer foreign DNA sequences into the genome (Lidholm *et al.*, 1993). In this species and in some other Hymenoptera, the rare hyper-methylated TE are positively correlated with their expression levels (Bonasio *et al.*, 2012; Wang *et al.*, 2013). These results suggest higher methylation is associated with active TEs. This hypothesis is yet to be tested in *T. cristinae*, and future work could estimate the relationship between hyper-methylated TEs and their expression. Overall, this evidence and

the extensive methylation depletion in insect TEs contrasts markedly with the typical patterns in plants and in mammals, where methylation is linked with suppression of transposons and plays a role in maintaining genomic stability (Yoder *et al.*, 1997; Suzuki and Bird, 2008; Jones, 2012).

2.6. Conclusion

This study contributed to the understanding of DNA methylation patterns in insects. It showed there are many similarities between *T. cristinae* stick insects' methylation profile and other hemimetabolous insects in the literature. With this, it is possible to highlight the particularities of insects and differences in their methylomes: a group with such diverse forms and functions and disparate patterns in DNA methylation. Given the potential roles DNA methylation performs during development and in adaptation to natural environments, it is possible such differences could have had a role in the diversification of insects. The role of DNA methylation in insects remains controversial and largely unexplored. Although there are many marked differences in the methylation patterns between groups of insects, little is known about its molecular role and its functional consequences. In this context, the investigation of different taxa will help shed light on the understanding of some specific and general roles of these epigenetic mechanisms in insects.

Appendix A: Supplementary Tables and Figures – Chapter 2

Quality of T. cristinae *genome assembly*

To assess the quality of the current *Timema* genome assembly (version 1.3c2), I used the **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs (BUSCO; Waterhouse *et al.*, 2017) tools. Genes that compose the BUSCO datasets for each major lineage are selected from orthologous groups with genes present as single-copy orthologs in at least 90% of the species. The BUSCO software provides quantitative measures to assess the completeness of the genome based on evolutionary-informed expectations of gene content from OrthoDB *v9* (Zdobnov *et al.*, 2017). It identifies BUSCO gene ortoholog group ('BUSCOs') matches using Hidden Markov Models (HMMER; Johnson *et al.*, 2010) and *de novo* gene predictions using Augustus (Stanke *et al.*, 2008). The matches are then classified according to the orthologs database. The recovered matches are classified as: 'complete' if their lengths correspond to BUSCO profile match lengths; 'duplicated' if are found more than once; 'fragmented' if are partially recovered; and 'missing' if no matches are recovered. I used the Insecta lineage from OrthoDB *v9* as database, and the default species *Drosophila melanogaster* gene finding parameters to be used by Augustus. The analyses output 95.3% of the gene ortholog groups to be complete in length in *T. cristinae* genome assembly v1.3c2 (Table A1). This result suggests a good level of completeness in terms of the expected gene content, and therefore on the quality of the genome assembly, especially comparing to other insect models. For example, *D. serrata* presents 94.1% of BUSCO completeness, and *Heliconius melpomene* presents 81.6% (Waterhouse *et al.*, 2019). Thus, it is possible to conclude the current version of *T. cristinae* genome assembly is reasonably good.

Table A1: Output table from BUSCO analyses, using the Insecta lineage and *D. melanogaster* gene ortholog groups, which presented 1,658 BUSCOs.

	N BUSCOs	***Perc.***
Complete	1579	95.3%
Complete and single-copy	1573	94.9%
Fragmented	40	2.4%
Missing	39	2.3%

Table A2: Details about bisulfite sequencing data from the 24 individuals used in the population survey when mapped to Lambda phage.

Ind.	*Pop. Code*	*Flow cell*	*Reads parsed**	*Reads mapped*	*Mapping efficiency*	*Number mCpG*	*mCpG*	*mCHG*	*mCHH*
17_0003	N1.A	1	34168604	855943	3.6%	45261	0.3%	0.3%	0.3%
17_0005	N1.A	2	26972768	773449	3.2%	42749	0.3%	0.4%	0.3%
17_0006	N1.C	1	39084659	715644	3.0%	32586	0.3%	0.3%	0.3%
17_0009	N1.C	1	41538867	789980	3.3%	48231	0.3%	0.4%	0.4%
17_0012	FH.A	1	39058299	639604	2.7%	29695	0.3%	0.3%	0.3%
17_0015	FH.A	2	28605951	701599	2.9%	43430	0.4%	0.4%	0.4%
17_0018	L.A	1	28164953	741842	3.1%	44757	0.4%	0.4%	0.4%
17_0019	L.A	1	38153090	557688	2.3%	28155	0.3%	0.3%	0.3%
17_0043	HV.A	2	27926464	879418	3.7%	56195	0.4%	0.4%	0.4%
17_0045	HV.A	1	40963899	857479	3.6%	51307	0.3%	0.4%	0.4%
17_0049	HV.C	2	31688742	884972	3.7%	46816	0.3%	0.4%	0.3%
17_0051	HV.C	2	26277649	568375	2.4%	31528	0.3%	0.4%	0.3%
17_0057	SCN.A	2	33803298	644683	2.7%	35239	0.3%	0.4%	0.3%
17_0058	SCN.A	3	30211263	602447	2.5%	28159	0.3%	0.3%	0.3%
17_0062	SC.C	2	30774916	527677	2.2%	28750	0.3%	0.4%	0.4%
17_0065	SC.C	2	27550463	892349	3.7%	50861	0.3%	0.4%	0.4%
17_0067	OUT.A	3	33570295	669363	2.8%	30939	0.3%	0.3%	0.3%
17_0070	OUT.A	2	26509336	746317	3.1%	42062	0.3%	0.4%	0.4%
17_0074	OUT.C	3	38641548	751502	3.1%	39647	0.3%	0.3%	0.3%
17_0075	OUT.C	3	34971845	862776	3.6%	41707	0.3%	0.3%	0.3%
17_0077	PR.C	3	35172047	803637	3.3%	36049	0.3%	0.3%	0.3%
17_0081	PR.C	3	35313802	671131	2.8%	30435	0.3%	0.3%	0.3%
17_0082	BT.A	3	37679020	844279	3.5%	41296	0.3%	0.3%	0.3%
17_0086	BT.A	3	40428034	707912	2.9%	35245	0.3%	0.3%	0.3%

Pop. code= Population where the individual was collected. Locality and host are separated by a dot. Flow cell= Information about the flow cell that each individual was sequenced. Reads parsed= Represents the total number of reads retained after filtering. This step was followed by a random subsampling of 24 million reads in each sample before mapping. Reads mapped= Number of reads uniquely mapped to the unmethylated Lambda phage BS-converted genome, starting from the 24 million reads. Mapping efficiency= Percentage of reads uniquely mapped to Lambda phage. Number mCpG= number of methylated cytosines in CpG context. mCpG, mCHG, and mCHH correspond to the proportion of methylated cytosines in each one of those contexts.

Table A3: Details about BS-seq data from the 24 individuals used in the population survey when mapped to *T. cristinae* BS-converted reference genome 1.3c2. Mapping was performed using the reads that were not mapped to the phage.

Ind.	*Pop. code*	*Flow cell*	*Non-map. reads*	*Reads mapped*	*Mapping efficiency*	*Number mCpG*	*mCpG*	*mCHG*	*mCHH*
17_0003	N1.A	1	23144057	10224621	44.2%	8651643	14.2%	0.4%	0.4%
17_0005	N1.A	2	23226551	9849254	42.4%	7371433	12.6%	0.5%	0.5%
17_0006	N1.C	1	23284356	10244423	44.0%	8460007	14.0%	0.5%	0.5%
17_0009	N1.C	1	23210020	9768192	42.1%	8226048	14.5%	0.4%	0.4%
17_0012	FH.A	1	23360396	10618466	45.5%	8848849	14.3%	0.4%	0.4%
17_0015	FH.A	2	23298401	10523150	45.2%	7967329	13.0%	0.4%	0.4%
17_0018	L.A	1	23258158	10362239	44.6%	8759187	15.0%	0.4%	0.4%
17_0019	L.A	1	23442312	10306326	44.0%	8826738	14.4%	0.4%	0.4%
17_0043	HV.A	2	23120582	10430206	45.1%	7254779	12.3%	0.5%	0.5%
17_0045	HV.A	1	23142521	10007169	43.2%	8729881	14.9%	0.7%	0.6%
17_0049	HV.C	2	23115028	10349562	44.8%	7077714	11.9%	0.5%	0.5%
17_0051	HV.C	2	23431625	10750822	45.9%	7602182	12.3%	0.6%	0.7%
17_0057	SCN.A	2	23355317	10608031	45.4%	7465945	12.3%	0.9%	0.9%
17_0058	SCN.A	3	23397553	9095218	38.9%	7742320	14.1%	0.7%	0.7%
17_0062	SC.C	2	23472323	11068668	47.2%	7788747	12.7%	0.5%	0.5%
17_0065	SC.C	2	23107651	10191423	44.1%	6932999	11.8%	0.9%	0.9%
17_0067	OUT.A	3	23330637	10479511	44.9%	8004357	13.5%	0.4%	0.4%
17_0070	OUT.A	2	23253683	10717584	46.1%	7632506	12.8%	0.6%	0.6%
17_0074	OUT.C	3	23248498	10172910	43.8%	7737917	13.6%	0.5%	0.5%
17_0075	OUT.C	3	23137224	10460090	45.2%	7828798	13.3%	0.5%	0.5%
17_0077	PR.C	3	23196363	9814177	42.3%	7010376	12.4%	0.7%	0.6%
17_0081	PR.C	3	23328869	9829352	42.1%	7584060	12.9%	1.0%	1.0%
17_0082	BT.A	3	23155721	10125500	43.7%	7299680	12.9%	0.4%	0.4%
17_0086	BT.A	3	23292088	9588872	41.2%	6831541	12.4%	0.5%	0.5%

Non-map. reads= Number of reads that were not uniquely mapped to the Lambda phage. Reads mapped= Number of reads uniquely mapped to *T. cristinae* BS-converted reference genome starting from the reads that were not mapped to the Lambda phage. Mapping efficiency= Percentage of reads uniquely mapped to *T. cristinae.* Number mCpG= number of methylated cytosines in CpG context. mCpG, mCHG, and mCHH correspond to the proportion of methylated cytosines in each one of those contexts.

Table A4: Details about whole-genome sequencing data re-analysed to estimate a list of single nucleotide polymorphism (SNPs) in *T. cristinae*. Accessions were downloaded from NCBI database (https://www.ncbi.nlm.nih.gov/) and a subsample of 20 accessions was randomly selected for downstream analysis. Sites identified as C/T and G/A polymorphisms were selected and added to the list of SNPs to be removed from methylation datasets.

Location	Host	N	Publication
HV	A	20	Soria-Carrasco *et al.* 2014
HV	C	20	Soria-Carrasco *et al.* 2014
L	A	19	Soria-Carrasco *et al.* 2014
PR	C	19	Soria-Carrasco *et al.* 2014
FH	A	20	Riesch *et al.* 2017

Table A5: Frequency of transposable elements (TEs) found in the *T. cristinae* genome. Only TEs found in frequency higher than 400 copies across *T. cristinae* RepeatMasker database (Villoutreix *et al. in prep*) were considered in this study, following the procedures from Libbrecht et al. (2016). I tested whether TEs were enriched in methylation levels, and the results are reported in Tables A8-A9.

TE class	TE family	Frequency
DNA Transposon	*Academ*	7,088
	Chapaev	1,073
	EnSpm	3,066
	Harbinger	10,340
	HAT	64,998
	Helitron	8,809
	Mariner	15,728
	MuDr	3,238
	PiggyBac	987
	Polinton	22,029
	Sola	1,762
LTR Retrotransposon	*BEL*	3,127
	Copia	14,016
	Ginger	540
	Gypsy	14,225
	HERV	3,683
Non-LTR retrotransposon	*CR1*	4,543
	Crack	7,070
	Jockey	4,062
	Kiri/L2	1,632
	R1	2,727
	RTE	12,354
	SINE	971
PLE	*Penelope*	4,888

Table A6: Number of sites distributed in different genomic features across *T. cristinae*. Numbers regarding the genomic distribution **(A)** were estimated based on the reference genome. Number of CpG sites **(B-D)** were estimated for each individual, then averaged (95% CI in brackets). **(B)** Average number of CpG sites present in the datasets prior filtering for minimum coverage, and after removal of potential genetic polymorphisms. **(C)** Number of CpGs after filtering for minimum 5 reads and maximum of 60 reads covering each site. **(D)** Number of methylated CpGs at the covered datasets.

Genomic feature	(A) Genome	(B) CpG raw data	(C) Covered CpGs	(D) Methylated CpGs
exon	21,771,317	759,261 [8,129]	154,476 [10,869]	76,675 [6,258]
intron	172,540,652	1,560,284 [31,695]	282,149 [15,830]	129,319 [8,000]
upstream	19,315,665	246,537 [4,456]	49,149 [2,462]	11,333 [691]
downstream	19,298,682	198,583 [3,879]	40,833 [1,866]	13,459 [797]
Intergenic	720,405,671	9,951,002 [193,945]	1,666,514 [103,158]	164,880 [9,642]

Table A7: Enrichment in exons across all individuals (low5_high60)

Ind.	Number sites	95% Quantiles	Enrich.	*p-value*	
17_0003	88285	[31730 - 32339]	2.76	<2.2e-16	***
17_0005	91358	[31009 - 31628]	2.92	<2.2e-16	***
17_0006	81109	[29675 - 30262]	2.71	<2.2e-16	***
17_0009	68859	[25968 - 26513]	2.62	<2.2e-16	***
17_0012	88447	[32712 - 33330]	2.68	<2.2e-16	***
17_0015	98252	[33509 - 34136]	2.91	<2.2e-16	***
17_0018	79613	[29739 - 30317]	2.65	<2.2e-16	***
17_0019	90496	[33967 - 34594]	2.64	<2.2e-16	***
17_0043	54873	[19282 - 19761]	2.81	<2.2e-16	***
17_0045	79072	[30109 - 30689]	2.6	<2.2e-16	***
17_0049	57223	[19851 - 20347]	2.85	<2.2e-16	***
17_0051	66187	[22622 - 23149]	2.89	<2.2e-16	***
17_0057	65663	[22488 - 23019]	2.89	<2.2e-16	***
17_0058	106481	[37739 - 38400]	2.8	<2.2e-16	***
17_0062	66796	[22751 - 23281]	2.9	<2.2e-16	***
17_0065	59590	[21085 - 21591]	2.79	<2.2e-16	***
17_0067	74702	[26621 - 27181]	2.78	<2.2e-16	***
17_0070	59531	[21083 - 21587]	2.79	<2.2e-16	***
17_0074	63333	[23268 - 23794]	2.69	<2.2e-16	***
17_0075	69341	[25020 - 25567]	2.74	<2.2e-16	***
17_0077	68155	[23965 - 24487]	2.81	<2.2e-16	***
17_0081	103284	[35813 - 36463]	2.86	<2.2e-16	***
17_0082	62736	[22564 - 23079]	2.75	<2.2e-16	***
17_0086	62310	[22426 - 22948]	2.75	<2.2e-16	***

Table A8: Enrichment in introns across all individuals (low5_high60)

Ind.	Number sites	95% Quantiles	Enrich.	*p-value*	
17_0003	137463	[55075 - 55850]	2.48	<2.2e-16	***
17_0005	145285	[54005 - 54782]	2.67	<2.2e-16	***
17_0006	132033	[53175 - 53930]	2.47	<2.2e-16	***
17_0009	115104	[47326 - 48043]	2.41	<2.2e-16	***
17_0012	150671	[61223 - 62042]	2.44	<2.2e-16	***
17_0015	166105	[61749 - 62581]	2.67	<2.2e-16	***
17_0018	153612	[62468 - 63287]	2.44	<2.2e-16	***
17_0019	140884	[57792 - 58574]	2.42	<2.2e-16	***
17_0043	102262	[37723 - 38380]	2.69	<2.2e-16	***
17_0045	128692	[53376 - 54126]	2.39	<2.2e-16	***
17_0049	102969	[37422 - 38075]	2.73	<2.2e-16	***
17_0051	117833	[42683 - 43384]	2.74	<2.2e-16	***
17_0057	115784	[42055 - 42736]	2.73	<2.2e-16	***
17_0058	153575	[59843 - 60662]	2.55	<2.2e-16	***
17_0062	135546	[49160 - 49900]	2.74	<2.2e-16	***
17_0065	103471	[38401 - 39065]	2.67	<2.2e-16	***
17_0067	130263	[49950 - 50710]	2.59	<2.2e-16	***
17_0070	117152	[43154 - 43854]	2.69	<2.2e-16	***
17_0074	114243	[44092 - 44779]	2.57	<2.2e-16	***
17_0075	121728	[46409 - 47135]	2.6	<2.2e-16	***
17_0077	110874	[42282 - 42970]	2.6	<2.2e-16	***
17_0081	155074	[58857 - 59671]	2.62	<2.2e-16	***
17_0082	116802	[44040 - 44741]	2.63	<2.2e-16	***
17_0086	98244	[37995 - 38648]	2.56	<2.2e-16	***

Table A9: Enrichment in upstream region 1kb at 5' from the gene across all individuals (low5_high60)

Ind.	Number sites	95% Quantiles	Enrich.	*p-value*	
17_0003	14863	[10736 - 11101]	1.36	<2.2e-16	***
17_0005	15041	[10672 - 11045]	1.39	<2.2e-16	***
17_0006	14190	[10575 - 10938]	1.32	<2.2e-16	***
17_0009	12208	[9205 - 9540]	1.3	<2.2e-16	***
17_0012	15651	[11394 - 11768]	1.35	<2.2e-16	***
17_0015	16363	[11446 - 11828]	1.41	<2.2e-16	***
17_0018	15626	[11274 - 11639]	1.36	<2.2e-16	***
17_0019	15403	[11317 - 11687]	1.34	<2.2e-16	***
17_0043	10212	[7545 - 7858]	1.33	<2.2e-16	***
17_0045	13426	[10193 - 10544]	1.29	<2.2e-16	***
17_0049	10786	[7816 - 8127]	1.35	<2.2e-16	***
17_0051	12077	[8468 - 8797]	1.4	<2.2e-16	***
17_0057	11442	[8356 - 8686]	1.34	<2.2e-16	***
17_0058	16087	[11528 - 11902]	1.37	<2.2e-16	***
17_0062	13050	[9455 - 9797]	1.36	<2.2e-16	***
17_0065	10853	[8079 - 8403]	1.32	<2.2e-16	***
17_0067	13636	[9715 - 10068]	1.38	<2.2e-16	***
17_0070	11606	[8455 - 8786]	1.35	<2.2e-16	***
17_0074	11707	[8712 - 9046]	1.32	<2.2e-16	***
17_0075	12391	[9125 - 9465]	1.33	<2.2e-16	***
17_0077	11151	[8156 - 8480]	1.34	<2.2e-16	***
17_0081	15748	[11305 - 11674]	1.37	<2.2e-16	***
17_0082	11861	[8696 - 9024]	1.34	<2.2e-16	***
17_0086	10392	[7617 - 7933]	1.34	<2.2e-16	***

Table A10: Enrichment in downstream region 1kb at 3' from the gene across all individuals (low5_high60)

Ind.	Number sites	95% Quantiles	Enrich.	*p-value*	
17_0003	16922	[9225 - 9560]	1.8	<2.2e-16	***
17_0005	17242	[8605 - 8937]	1.97	<2.2e-16	***
17_0006	16629	[9013 - 9348]	1.81	<2.2e-16	***
17_0009	14127	[7878 - 8192]	1.76	<2.2e-16	***
17_0012	18147	[9678 - 10024]	1.84	<2.2e-16	***
17_0015	19331	[9484 - 9837]	2.0	<2.2e-16	***
17_0018	18075	[9658 - 9999]	1.84	<2.2e-16	***
17_0019	17113	[9507 - 9846]	1.77	<2.2e-16	***
17_0043	11967	[6310 - 6596]	1.85	<2.2e-16	***
17_0045	15637	[8639 - 8961]	1.78	<2.2e-16	***
17_0049	12778	[6550 - 6841]	1.91	<2.2e-16	***
17_0051	13604	[6878 - 7179]	1.94	<2.2e-16	***
17_0057	13790	[6995 - 7295]	1.93	<2.2e-16	***
17_0058	18489	[9504 - 9850]	1.91	<2.2e-16	***
17_0062	15442	[7722 - 8034]	1.96	<2.2e-16	***
17_0065	12609	[6678 - 6968]	1.85	<2.2e-16	***
17_0067	15781	[8206 - 8526]	1.89	<2.2e-16	***
17_0070	13878	[7075 - 7373]	1.92	<2.2e-16	***
17_0074	13660	[7340 - 7644]	1.82	<2.2e-16	***
17_0075	14458	[7671 - 7983]	1.85	<2.2e-16	***
17_0077	13460	[7030 - 7324]	1.88	<2.2e-16	***
17_0081	18441	[9278 - 9623]	1.95	<2.2e-16	***
17_0082	13819	[7425 - 7730]	1.82	<2.2e-16	***
17_0086	11814	[6350 - 6630]	1.82	<2.2e-16	***

Table A11: Unpaired t-test comparing methylation levels (%) of short genes (up to four exons, n=7,529) and long genes (with five or more exons**,** n=7,127), and between their first exons and introns. The choice of number of exons to classify short or long genes was arbitrary.

	mean		***t***	***df***	***p-value***
	short	**long**			
all gene	32.68	50.69	32.59	2442.30	< 2.2E-16
all exons	33.98	54.67	32.23	1416.20	< 2.2E-16
exon 1	22.07	40.37	20.37	292.38	< 2.2E-16
exon2	31.08	51.24	21.04	370.99	< 2.2E-16
Exon 3	38.05	57.76	18.44	361.46	< 2.2E-16
Exon 4	44.88	57.28	7.69	283.66	2.4E-13
all introns	30.87	45.84	15.95	937.29	< 2.2E-16
Intron 1	22.96	35.56	10.20	377.88	< 2.2E-16
Intron 2	30.96	47.06	10.39	374.78	< 2.2E-16
Intron 3	39.47	48.83	4.41	265.97	1.5E-05
	exons	**introns**			
Short genes	33.98	30.87	3.16	1046.30	1.6E-03
Long genes	54.67	45.84	15.45	1920.00	< 2.2E-16

Table A12: List of Gene Ontology (GO) terms significantly enriched in genes that were methylated in all 24 samples (n =3,866) compared to the remaining genes (n=4,606). Information about 8,472 genes were present in all 24 samples and presented GO annotation. Fisher's exact test was used with the weight algorithm, which accounts for GO topology using R package *TopGO.* The 30 most significant terms were represented here.

GO term	Category	Description	Fold enrich.	*p-value*
GO:0043227	CC	membrane-bounded organelle	1.7	< 1e-30
GO:0005622	CC	intracellular	1.6	< 1e-30
GO:0010467	BP	gene expression	1.6	2.0e-29
GO:0044267	BP	cellular protein metabolic process	1.6	5.8e-26
GO:0035639	MF	purine ribonucleoside triphosphate binding	1.4	1.3e-23
GO:0032555	MF	purine ribonucleotide binding	1.4	1.8e-23
GO:0032991	CC	protein-containing complex	1.7	6.2e-23
GO:0005488	MF	binding	1.1	3.0e-18
GO:0016070	BP	RNA metabolic process	1.6	3.4e-16
GO:0043043	BP	peptide biosynthetic process	1.9	1.1e-12
GO:0003723	MF	RNA binding	1.8	5.9e-11
GO:0006886	BP	intracellular protein transport	2.0	1.0e-09
GO:0016192	BP	vesicle-mediated transport	1.9	4.0e-08
GO:0006457	BP	protein folding	2.1	9.2e-07
GO:0008026	MF	ATP-dependent helicase activity	2.2	4.0e-06
GO:0008168	MF	methyltransferase activity	1.9	8.0e-06
GO:0097659	BP	nucleic acid-templated transcription	1.4	1.6e-05
GO:0022613	BP	ribonucleoprotein complex biogenesis	2.1	1.4e-04
GO:0016310	BP	phosphorylation	1.4	1.5e-04
GO:0005694	CC	chromosome	1.9	1.6e-04
GO:0004842	MF	ubiquitin-protein transferase activity	1.8	2.8e-04
GO:0140101	MF	catalytic activity, acting on a tRNA	1.7	3.3e-04
GO:0016301	MF	kinase activity	1.3	4.1e-04
GO:0003735	MF	structural constituent of ribosome	1.7	6.6e-04
GO:0000287	MF	magnesium ion binding	1.9	7.6e-04
GO:0032259	BP	methylation	2.2	8.3e-04
GO:0005543	MF	phospholipid binding	1.8	4.3e-03
GO:0005815	CC	microtubule organizing center	1.9	8.3e-03

Table A13: List of Gene Ontology (GO) terms significantly enriched in genes that consistently non-methylated in all 24 samples (n=3,176) compared to the remaining genes (n=5,296). Information about 8,472 genes were present in all 24 samples and presented GO annotation. Fisher's exact test was used with the weight algorithm, which accounts for GO topology using R package *TopGO.* The 30 most significant terms were represented here. 'BP' represents biological process, 'CC' category represents cellular component, and 'MF' represents molecular function.

GO term	Category	Description	Fold enrich.	*p-value*
GO:0016020	CC	membrane	1.4	< 1e-30
GO:0004930	MF	G protein-coupled receptor activity	2.4	2.1e-25
GO:0031224	CC	intrinsic component of membrane	1.4	1.3e-23
GO:0007186	BP	G protein-coupled receptor signaling pathway	2.1	2.2e-19
GO:0004252	MF	serine-type endopeptidase activity	2.0	1.8e-13
GO:0015074	BP	DNA integration	1.9	1.8e-13
GO:0055085	BP	transmembrane transport	1.5	2.5e-12
GO:0008408	MF	3'-5' exonuclease activity	2.3	1.6e-11
GO:0005576	CC	extracellular region	1.8	1.2e-09
GO:0016705	MF	oxidoreductase activity	2.1	2.8e-09
GO:0003887	MF	DNA-directed DNA polymerase activity	2.1	1.1e-08
GO:0042302	MF	structural constituent of cuticle	2.6	1.8e-08
GO:0046906	MF	tetrapyrrole binding	1.9	3.2e-08
GO:0005506	MF	iron ion binding	2.0	1.5e-07
GO:0005549	MF	odorant binding	2.5	1.7e-06
GO:0004970	MF	ionotropic glutamate receptor activity	2.0	2.6e-06
GO:0022843	MF	voltage-gated cation channel activity	2.6	3.1e-06
GO:0006508	BP	proteolysis	1.3	6.8e-06
GO:0050877	BP	nervous system process	2.3	1.6e-05
GO:0008061	MF	chitin binding	2.2	2.0e-05
GO:0005102	MF	signaling receptor binding	2.1	4.1e-05
GO:0006030	BP	chitin metabolic process	2.0	6.5e-05
GO:0005230	MF	extracellular ligand-gated ion channel a...	2.0	7.6e-05
GO:0030001	BP	metal ion transport	1.7	2.7e-04
GO:0005272	MF	sodium channel activity	2.2	4.4e-04
GO:0030594	MF	neurotransmitter receptor activity	2.1	5.1e-04
GO:0005215	MF	transporter activity	1.6	6.5e-04
GO:0015672	BP	monovalent inorganic cation transport	1.6	9.3e-04
GO:0016917	MF	GABA receptor activity	2.5	1.2e-03
GO:0004984	MF	olfactory receptor activity	2.5	1.2e-03

Table A14: Numbers and methylation enrichments in each transposable element family in intergenic regions using at least 12 samples Enrichment was estimated comparing the number of methylated CpGs compared to the expected null distribution, which was estimated based on randomizations (section 2.3.5). *P-values* were calculated using Fisher's exact test.

TE class	TE family	CpG sites	Mean	mCpG sites	95%Quantiles (null dist.)	Enrich.	*p-value*	
DNA Transposon	*Academ*	1475	17.8% [0.7%]	385	[165–215]	2.0	<2.2e-16	***
	Chapaev	257	25.6% [1.9%]	79	[23–43]	2.4	4.2e-14	***
	EnSpm	2275	13.1% [0.4%]	488	[267–329]	1.6	<2.2e-16	***
	Harbinger	1549	13.2% [0.6%]	279	[174–226]	1.4	6.2e-09	***
	HAT	13455	15.2% [0.2%]	3185	[1676–1827]	1.8	<2.2e-16	***
	Helitron	3269	7.4% [0.3%]	265	[385–460]	0.6	1.7e-18	***
	Mariner	4505	29.3% [0.5%]	1768	[546–634]	3	<2.2e-16	***
	MuDr	582	4.1% [0.5%]	25	[59–90]	0.3	2.4e-12	***
	PiggyBac	155	35.1% [2.6%]	62	[12–28]	3.1	2.1e-17	***
	Polinton	25487	1.3% [0.0%]	281	[3399–3612]	0.1	<2.2e-16	***
	Sola	258	39.0% [2.1%]	122	[23–44]	3.7	<2.2e-16	***
LTR Retrotransposon	*BEL*	3487	11.3% [0.3%]	661	[434–512]	1.4	3.0e-19	***
	Copia	5585	25.5% [0.4%]	1945	[675–774]	2.7	<2.2e-16	***
	Gypsy	12976	16.7% [0.2%]	3519	[1639–1790]	2.1	<2.2e-16	***
	HERV	420	4.0% [0.7%]	24	[42–68]	0.4	7.0e-07	***
Non-LTR Retrotransposon	*CR1*	313	21.5% [1.4%]	106	[29–52]	2.6	1.6e-21	***
	Crack	537	15.2% [1.1%]	130	[55–86]	1.9	1.3e-12	***
	Jockey	7659	3.2% [0.1%]	207	[952–1066]	0.2	<2.2e-16	***
	R1	5548	10.4% [0.3%]	847	[666–765]	1.2	1.1e-07	***
	RTE	5843	14.8% [0.4%]	1036	[705–805]	1.4	<2.2e-16	***
PLE	*Penelope*	885	45.5% [1.1%]	528	[94–134]	4.7	<2.2e-16	***

Table A15: Numbers and methylation enrichments in each transposable element family within genes using at least 12 samples Enrichment was estimated comparing the number of methylated CpGs compared to the expected null distribution which was estimated based on randomizations (section 2.3.5). *P-values* were calculated using Fisher's exact test.

TE class	TE family	CpG sites	Mean	mCpG sites	95%Quantiles (null dist.)	Enrich.	*p-value*	
DNA Transposon	*Academ*	160	35.2% [2.7%]	75	[68–93]	0.9	0.21	
	Chapaev	36	51.5% [5.2%]	19	[11–23]	1.1	0.31	
	EnSpm	480	29.5% [1.3%]	189	[208–250]	0.8	1.5e-04	***
	Harbinger	573	26.1% [1.3%]	195	[249–296]	0.7	3.2e-11	***
	HAT	2,959	38.7% [0.6%]	1465	[1,360–1,468]	1.0	0.03	*
	Helitron	1,186	13.2% [0.7%]	179	[544–612]	0.3	<2.2e-16	***
	Mariner	2,235	49.9% [0.6%]	1458	[1,022–1,112]	1.4	<2.2e-16	***
	MuDr	151	20.9% [2.6%]	34	[59–83]	0.5	3.0e-10	***
	PiggyBac	92	42.3% [3.1%]	60	[34–53]	1.4	3.8e-04	***
	Polinton	4,873	4.9% [0.2%]	256	[2,311–2,445]	0.1	<2.2e-16	***
	Sola	110	48.6% [2.9%]	80	[42–63]	1.5	4.9e-08	***
LTR Retrotransposon	*BEL*	1,657	12.4% [0.5%]	379	[777–859]	0.5	<2.2e-16	***
	Copia	3,686	29.7% [0.5%]	1585	[1,721–1,839]	0.9	7.5e-11	***
	Gypsy	4,119	23.6% [0.4%]	1426	[1,924–2,051]	0.7	<2.2e-16	***
	HERV	66	31.6% [4.5%]	19	[23–39]	0.6	1.7e-03	**
Non-LTR Retrotransposon	*CR1*	31	33.4% [3.8%]	31	[24–41]	1.0	0.40	*
	Crack	25	38.5% [4.2%]	25	[23–39]	0.8	0.08	***
	Jockey	148	5.1% [0.3%]	148	[1,039–1,132]	0.1	<2.2e-16	***
	R1	373	14.3% [0.6%]	373	[741–819]	0.5	<2.2e-16	*
	RTE	756	38.1% [0.9%]	756	[762–843]	0.9	0.01	***
PLE	*Penelope*	414	65.3% [1.1%]	414	[217–261]	1.7	<2.2e-16	

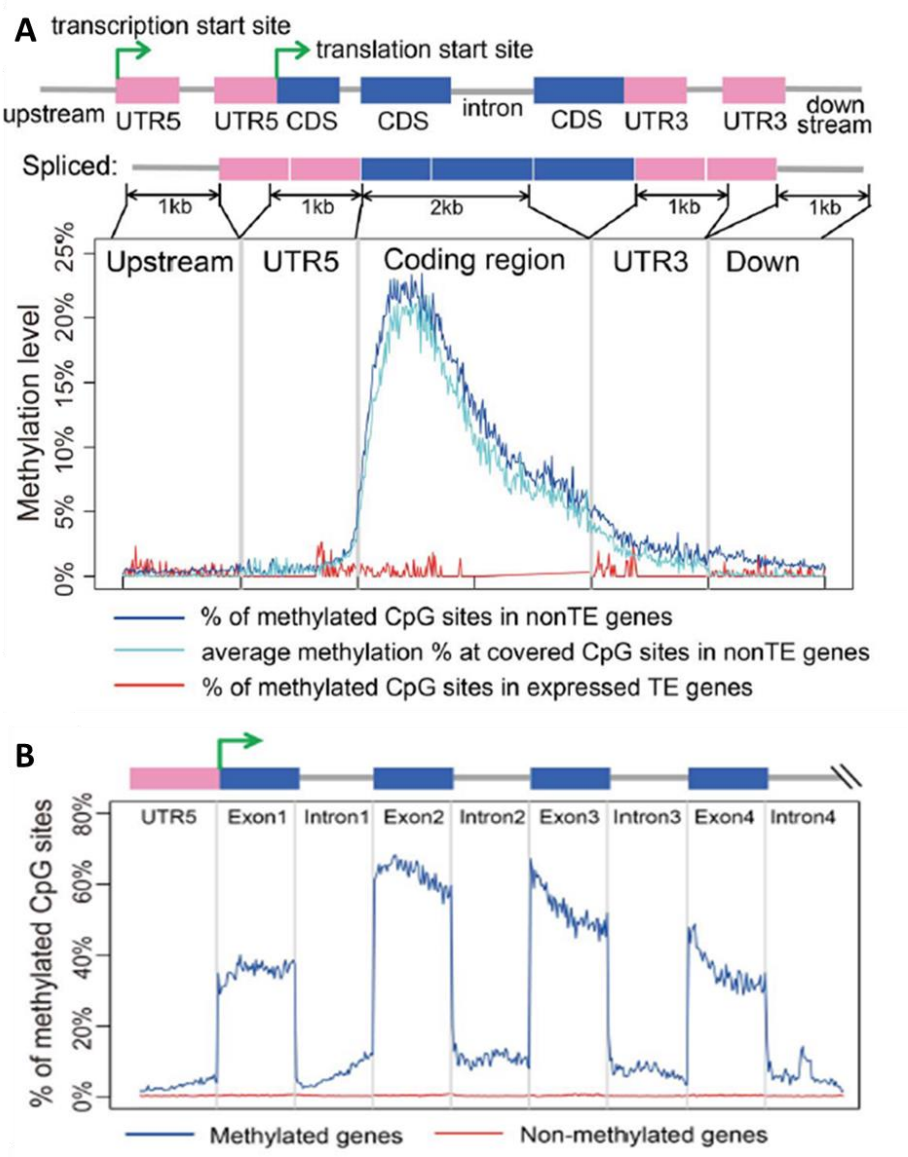


Figure A1: Methylation patterns in CpG context in *Nasonia vitripensis* (Hymenoptera), used to illustrate typical patterns in Holometabola. (**A**) Methylation levels at different genomic features, at 1 kbp upstream, 1 kbp UTR, first 2 kbp coding regions, 1 kbp 3' UTR and 1 kbp downstream regions for transposable element genes (TE genes) and non-TE genes. Methylation levels are enhanced at 5' side of the coding regions. (**B**) Averaged methylation levels represented at first four exons and introns. Methylation is mainly targeted at exons, while introns are not considerably methylated. Figure taken from Wang *et al.* (2013).

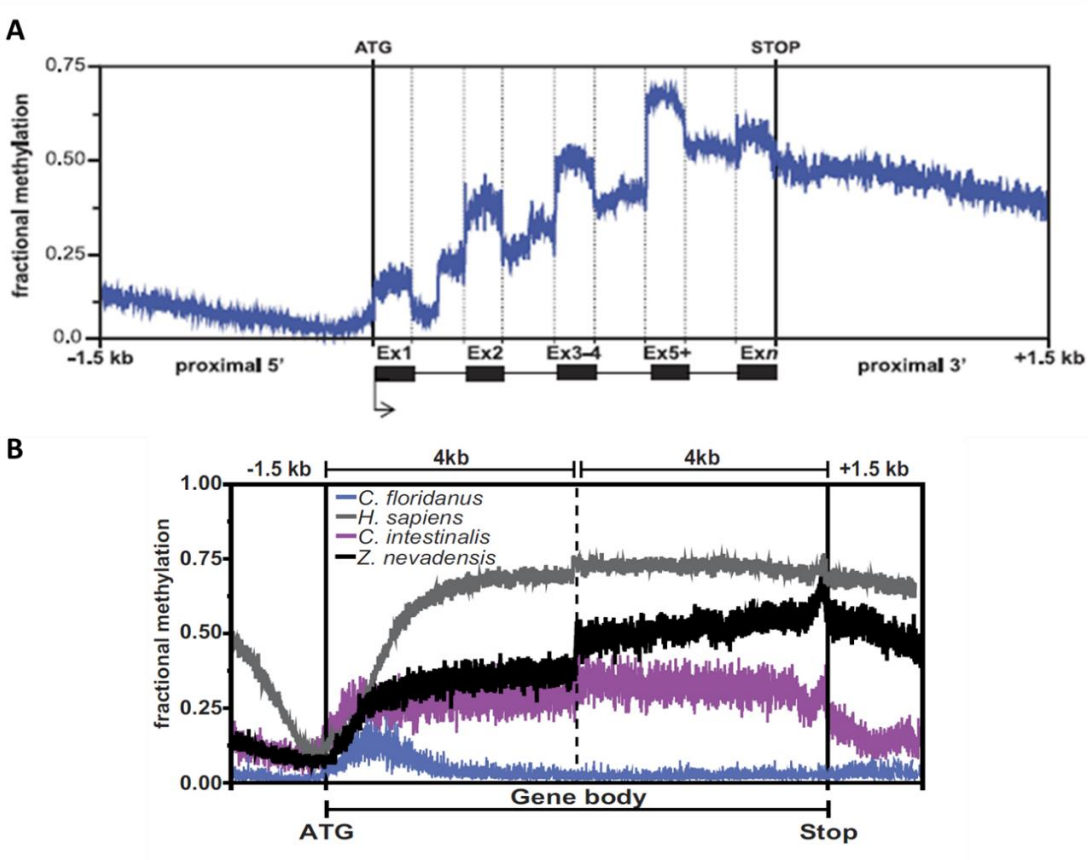


Figure A2: Methylation patterns in CpG context in *Zootermopsis nevadensis* (Isoptera), used to illustrate typical patterns in "Hemimetabola". (**A**) Average fractional methylation (*i.e.* proportion of methylated cytosines) at multi-exon genes, at 1.5 kbp upstream, exons and introns, and 1.5 kbp downstream. Methylation levels tend to increase from 5'→ 3', with less accentuated difference between exons and introns. A similar pattern is found in *T. cristinae.* (**B**) Comparative fractional methylation within the first and last 4 kb of gene bodies (exons + introns): *Z. nevadensis* (black), in *C. floridanus* (Hymenoptera; blue), a non-insect invertebrate (*Ciona intestinalis*; purple), and mammal (*Homo sapiens*; grey). Methylation levels in both *Z. nevadensis* and in *T. cristinae* are more similar to the patterns in these Chordata organisms than to Holometabola insects.

Chapter 3

Patterns and drivers of DNA methylation variation in natural populations of *Timema cristinae* stick insects

3.1. Summary

Epigenetic factors can contribute to phenotypic diversity. For instance, DNA methylation can influence gene regulation, and thus phenotypic plasticity. However, little is yet known about how and why methylation varies in wild populations. Here, I investigated whole-genome methylation profiles in natural populations of the *Timema cristinae* stick insects, depicting the factors shaping genome-wide methylation patterns. I tested the hypotheses that natural methylation variation is structured in geographical space and correlated with environmental factors such as host-plant and climate. We further tested for association between genetic and methylation variation. Using data obtained from whole-genome bisulfite sequencing, I found that methylation variation in CpG context tends to cluster following the geographical distribution of populations. Multivariate analysis revealed this pattern is better explained by genetic variation than by geographical distance only. Environmental factors were not significantly correlated with genome-wide methylation patterns. Binomial mixed models revealed moderate heritability in methylation status (0.67 [0.15 – 1.0 95%CI] across all sites), suggesting variation can accumulate given limited dispersal in space.

3.2. Introduction

Organisms often vary phenotypically within and between populations. These differences might result from genetic variation, shaped by the balance between natural selection and random neutral processes. In addition, the phenotype might arise as a direct interaction with the surrounding conditions, varying according to either an internal or external environmental signal (West–Eberhard, 2003). Together, the individuals' ability to tune in to their environment and genetic variation allow populations to persist and evolve, a process that can happen very rapidly (Reznick & Ghalambor, 2001; Prentis *et al.*, 2008; Scoville & Pfrender, 2010). Yet there are many gaps in the understanding of how the environment directly influences the phenotype, so that it is still debatable how organisms can adjust to environmental changes (Forsman, 2015; Foust *et al.*, 2016). Currently, there is mounting evidence that phenotypic diversity can also be caused by variation in epigenetic modifications, which could play a role in the response to complex environments (Schlichting and Smith, 2002; Hu and Barrett, 2017; Richards *et al.*, 2017).

Epigenetic mechanisms describe molecular processes that can affect gene expression and its function without a change in the underlying DNA sequence. These processes can involve: methylation of cytosine residues in the DNA, remodelling of chromatin structure through histone modifications, and gene regulation mediated by small RNAs (Bird, 2007; Law and Jacobsen, 2010). Among these epigenetic mechanisms, DNA methylation is by far the most studied one. DNA methylation describes the reversible addition of a methyl group when a cytosine is followed by a guanine residue in the genome (*i.e.* CpG sites). The symmetric conformation of CpG dinucleotides allows the methyltransferase to transmit the epigenetic information to newly generated DNA strands during mitosis (Goll and Bestor, 2005; Richards, 2006). DNA methylation is present in most major eukaryotic groups (Feng *et al.*, 2010; Zemach *et al.*, 2010), and it is known to play roles in: modulating gene expression; genomic imprinting; alternative splicing; and maintaining genome integrity by supressing transposable elements' activity (Law and Jacobsen, 2010; Schübeler, 2015). Not

surprisingly, these epigenetic mechanisms are intimately linked with cell differentiation during embryogenesis, and they may determine which genes will be transcriptionally active in different tissues (Reik, 2007). For this to occur, extensive demethylation happens in the genome between generations to assure the pluripotency of the embryo and its correct development in plants and mammals (Reik, 2007; Crevillén *et al.*, 2014). This is why DNA methylation variation does not tend to be meiotically transmitted, although exceptions to this rule are being discovered each day (Verhoeven *et al.*, 2010; Jiang *et al.*, 2013; Wang *et al.*, 2016; Richards *et al.*, 2017).

The role that DNA methylation plays at the genomic and cellular levels can have an effect on the phenotype, and ultimately influence evolutionary processes. One of the most celebrated examples is the toadflax (*Linaria vulgaris*). Its natural floral polymorphisms are associated with methylation changes of the *cis*-regulatory region of the gene responsible for the dorsal-ventral asymmetry (*Lcyc*; Cubas *et al.* 1999). This epigenetic allele (*i.e.* epiallele) is heritably stable and co-segregates with the phenotype. Some other examples have been described, although only a few have been shown to be stably transmitted over generations independently from the genetic background (see Manning *et al.*, 2006; Paszkowski and Grossniklaus, 2011).

Changes in DNA methylation status may occur in response to environmental triggers. Internal cues, such as hormones, can act to affect short and long-term methylation modifications (Stevenson, 2017). For example, oestrogen is known to regulate the *de novo* DNA methyltransferase (DNMT3) expression and to affect several tissues during cell differentiation (*e.g.* regulating sex-specific gene isoform expression in mice; Nugent *et al.*, 2015). In addition, DNA methylation may respond to external environmental triggers (Johnson and Tricker, 2010; Feil and Fraga, 2012). Change in diet affects coat colour in mice, a process related to DNA methylation modifications on the *Agouti* gene (Morgan *et al.*, 1999; Waterland and Jirtle, 2003). In honeybees, DNA methylation changes in response to differential feeding with royal jelly, and ultimately influences the development of larva into

queens or workers (Kucharski *et al.*, 2008; Lyko *et al.*, 2010; Foret *et al.*, 2012). DNA methylation's property to be environmentally-sensitive along with its role in many biological processes suggest this epigenetic mechanism could be involved in phenotypic plasticity, acting as a mediator between the external environment and genome regulation (Bossdorf *et al.*, 2008; Verhoeven *et al.*, 2016). Although knowledge about DNA methylation's role in molecular pathways and in cell signalling is rapidly improving, the ecological and evolutionary consequences of epigenetic mechanisms remain largely unknown.

To obtain a comprehensive understanding of the role DNA methylation variation might play in facilitating phenotypic plasticity and evolution, it is essential to place these processes in an ecological perspective and study their patterns, drivers and consequences in natural populations (Bossdorf *et al.*, 2008; Richards, 2008; Hu and Barrett, 2017; Richards *et al.*, 2017). Usually, the significance of DNA methylation is evaluated using genetically identical organisms (*e.g.* inbred lines) and their response to stress in laboratory settings (*e.g.* Johannes *et al.*, 2009; Verhoeven *et al.*, 2010). Although these studies are valuable to unveil the mechanisms underlying DNA methylation changes and the molecular pathways leading to them, an imperative next step is to explore these processes in natural conditions, in the complex environments where organisms live and evolve (Richards 2008, 2011; Richards *et al.* 2010; Herrera *et al.* 2014). By studying realistic scenarios with genetically and environmentally heterogeneous populations, one can investigate the intertwined factors acting simultaneously on natural methylation variation, which are possibly missed in laboratory experiments (Herrera and Bazaga, 2011; Ledón-Rettig, 2013; Herrera *et al.*, 2014).

To begin with, it is important to investigate the magnitude and structure of methylation variation in different populations in order to depict its patterns in nature. Then, one can estimate the origin and the forces driving this variation (Bossdorf *et al.*, 2008). Namely, DNA methylation variation can result from: (1) stochastic changes, (2)

environmental effect, and (3) genetic control (Fig. 2 in Chapter 1); and possibly be further shaped by forces of natural selection and drift (Bossdor *et al.*, 2008; Richards *et al.*, 2017). Stochastic changes in the methylation status often occur due to a failure of enzymes called methyltransferases to faithfully maintain genome-wide methylation patterns (Law and Jacobsen, 2010). As a consequence, variation can arise from spontaneous epimutations, which tend to happen at a much higher rate compared to genetic mutations (Becker *et al.*, 2011; Schmitz *et al.*, 2011; van der Graaf *et al.*, 2015). In addition, as outlined above, DNA methylation can respond to environmental triggers. As such, methylation variation could emerge from the interaction with the environment, which ultimately affect the phenotype (Johnson and Tricker, 2010; Herrera *et al.*, 2012; Zhang *et al.*, 2013; Duncan *et al.*, 2014). However, it is still controversial whether the environment can promote heritable methylation modifications, and it if can, to what extent it is transmitted (*i.e.* inheritance can be restricted to only a few subsequent generations; Richards *et al.*, 2017). Lastly, methylation variation can arise from genetic control, which can act via *cis* or *trans* regulation (Taudt *et al.*, 2016). This genetic control over methylation variation has been demonstrated in *Arabidopsis thaliana.* Not only the genetic background could partially explain DNA methylation variation, but also where changes specific in genetic sequence are related to changes in methylation (Becker *et al.*, 2011; Dubin *et al.*, 2015). These findings fuelled the debate about the dependency of DNA methylation variation on the underlying genetic variability, on whether the epigenetic variation can exist and perpetuate in the absence of genetic control (Richards, 2006; Dubin *et al.*, 2015).

To date, it has been shown that high levels of methylation variation exist in the wild, often exceeding estimates of genetic variation (Platt *et al.*, 2015; Groot *et al.*, 2018), and that epigenetic variation can be structured in space (Herrera and Bazaga, 2010; Herrera *et al.*, 2016; Smith *et al.*, 2016). Moreover, some studies have revealed significant correlations between DNA methylation variation and phenotypic diversity in different habitats (Lira-Medeiros *et al.*, 2010; Herrera and Bazaga, 2011; Foust *et al.*, 2016), even in the absence of

genetic variability (Richards *et al.*, 2012; Liebl *et al.*, 2013; Medrano *et al.*, 2014). Finally, there is evidence DNA methylation differentiation might persist after gametogenesis (Herrera *et al.*, 2013; Hagmann *et al.*, 2015; van der Graaf *et al.*, 2015), and that changes arisen from the interaction with the environment might be inherited at least to the subsequent generation (Johannes *et al.*, 2009; Verhoeven *et al.*, 2010; Preite *et al.*, 2018). Yet the great majority of studies have been performed in plant populations, with very few examples in vertebrates (*e.g.* Liebl *et al.*, 2013; Skinner *et al.*, 2014; Lea *et al.*, 2016; Carja *et al.*, 2017) and even fewer in invertebrates (*e.g.* Kille *et al.*, 2013; Ardura *et al.*, 2017). Surveys in non-model organisms have been reliant on techniques like amplified fragment length polymorphism (AFLPs) and its methylation sensitive version (methylation sensitive amplified polymorphisms; MSAPs) to capture genetic and epigenetic variation in the wild. These methods screen anonymous loci, and thus cannot specify the genomic region tagged by DNA methylation. Hence, they offer limited information content other than general variation defined by methylated or non-methylated state (Schrey *et al.*, 2013; Trucchi *et al.*, 2016).

A new cost-effective method is sodium bisulfite sequencing (BS-seq), which allows the estimation of genome-wide methylation at base pair resolution (Cokus *et al.*, 2008). This is a promising technique for ecological and evolutionary studies of DNA methylation, as it provides a much higher resolution in the investigation of methylation variation and a genomic context behind it (*e.g.* Becker *et al.*, 2011; Dubin *et al.*, 2015; Platt *et al.*, 2015; Lea *et al.*, 2017). In addition to this technique, research in DNA methylation can make use of some methodological approaches developed in population genetics to investigate natural variation. In fact, the questions raised in the process of understanding the magnitude, patterns and implications of methylation variation resemble the ones addressed in population genetics (Richards, 2006; Bossdorf *et al.*, 2008). Furthermore, genetic models can be used as a proxy while studying methylation variation, in order to capture the aspects

that are inherent only to epigenetics (*e.g.* being capable of changing with environmental change and the reset of methylation marks between generations; Herrera *et al.*, 2016).

In this context, insects are a promising group to study natural methylation variation. This clade contains well-known examples of plasticity, including caste differentiation and seasonal polyphenisms (Simpson *et al.*, 2011). Studying insects and their ability to encode multiple and diverse phenotypes might help to understand their great ecological success (Moczek, 2010; Lo *et al.*, 2018). Although DNA methylation is conserved among eukaryotes, its genomic patterns vary across taxa (Feng *et al.*, 2010; Zemach *et al.*, 2010). This means some of the conclusions raised in one taxonomic group cannot necessarily be extrapolated to others (Lea *et al.*, 2017). In insects, methylation occurs at much lower levels than in vertebrates; it is sparsely distributed across the genome and mainly targeting the gene bodies (Suzuki and Bird, 2008). Although DNA methylation functions are not well understood in insects, they are possibly different from the roles played in vertebrates, and perhaps even more from the ones played in plants (Chapter 2). As such, studying natural DNA methylation variation in this group of organisms could bring novel insights into the ecological and evolutionary importance of this epigenetic mechanism.

This study used *Timema cristinae* (Phasmatodea: Timematodea; Vickery, 1993), a species of stick insect native to South California, to address some of these topics. Extensive population genetics work has been performed in this species (*e.g.* Sandoval, 1994b; Nosil and Crespi, 2006; Nosil *et al;*, 2008; Gompert *et al.*, 2014), which provides a good starting point to investigate similar questions using DNA methylation instead of genetic variation. In particular, this work aimed to investigate how the *T. cristinae* DNA methylation profile varies across different populations, and which mechanisms are underlying its diversity. A population survey was conducted to test the hypothesis that (1) methylation variation is structured in geographical space. This led to the non-mutually exclusive predictions that (a) methylation variation is correlated with genetic variation, and that (b) it is associated with the environment. Finally, this study also tested the hypothesis that (2) there is some

heritability of methylation levels. In other words, the more closely related the individuals are, the more similar the methylation patterns will tend to be. To this end, a multifaceted dataset on *T. cristinae* natural populations was generated, including information about: methylation variation (using BS-seq); genetic variation (using newly acquired genetic data and reanalysis of some previously published data); environmental variables (*e.g.* abundance of host plants, elevation, and climatic variables); and geographical distance. This study revealed that natural DNA methylation variation tended to group in geographical space, and that the differentiation between populations increased with spatial distance. Underlying this pattern was the considerable correlation between DNA methylation and genetic variation, which was stronger than the association with geographical distance or with environmental variables. Binomial mixed models revealed a considerable relatedness of methylation patterns, mirroring the matrix of pairwise kinship estimated using genetic variation (*i.e.* kinship matrix). This result suggests there is some heritability of methylation variation, although the drivers could not be identified. Taken together, the findings from this study show the general patterns in natural methylation variation are strongly associated with its genetic background.

3.3. Materials and Methods

3.3.1. Study system

T. cristinae are plant-feeding and wingless stick insects native to the chaparral in the Santa Ynez Mountains, in Southern California. These insects rely on crypsis to escape detection by visual predators, and have evolved body colouration that matches the leaves and stems of the host plants they rest on (Sandoval, 1994a). Although *T. cristinae* can feed on a variety of plants, it is primarily found on two host species: *Ceanothus spinosus* (Rhamnaceae) and *Adenostoma fasciculatum* (Rosaceae). These two species of plant differ considerably in their leaf morphology, with *Ceanothus* plants presenting broad flat leaves, and *Adenostoma* plants exhibiting thin needle-like leaves (Fig. 1; Fig. 3 in Chapter 1). The

ecotypes in *T. cristinae* are characterized by the host plants they are found on, which determines the '*Ceanothus* ecotype' and the '*Adenostoma* ecotype' (Nosil *et al.*, 2006; Nosil, 2007). The most obvious difference between the ecotypes is the frequency of the typical *T. cristinae* morphs, characterized by presence or absence of a dorsal white stripe in their green body (respectively the 'striped' and 'green' morphs; Fig. 3 in Chapter 1). The striped morph is more frequently found in *Adenostoma*, and the green morph in *Ceanothus* plants (Sandoval, 1994a). Previous experiments showed the striped morph is more cryptic and suffers less predation on the needle-like leaves of *Adenostoma*, whereas the green unstriped morph is more cryptic and suffers less predation on the broad leaves of *Ceanothus* plants (Sandoval, 1994a; Nosil and Crespi, 2006). That is, divergent selection promoted by differential predation between the two host species has contributed to ecological isolation between the two ecotypes (Sandoval, 1994a; Nosil and Crespi, 2006). The third morph has a dark body colour (*i.e.* melanistic) and is often found on both host plant species, but in much rarer frequencies (Sandoval 1994a,b). The morphs segregate as a polymorphism controlled by a major locus, which means the frequencies of green, striped and melanistic alleles vary between the ecotypes (Comeault *et al.*, 2015; Lindtke *et al.*, 2017). Besides colour and pattern, these two ecotypes differ in a suite of other traits, including size, host plant preference, mate choice, and cuticular hydrocarbons (CHCs), molecules with roles in anti-desiccation and in insect communication (Nosil *et al.*, 2006; Nosil, 2007; Chung *et al.*, 2014; Riesch *et al.*, 2017).

The landscape where *T. cristinae* is found is characterized by a mosaic distribution of patches, which vary in size and abundance of the two host plant species. In this context, gene flow between patches with different selection regimes can occur despite the effects of maladaptation, creating a balance between natural selection and gene flow that affects the allele frequencies in *T. crisitinae* (Sandoval, 1994b). In addition, the limited dispersal between non-adjacent patches (*i.e.* allopatric populations) contributes to low gene flow and to the accumulation of genetic differentiation by neutral processes, resulting in patterns of

isolation by distance (Sandoval, 1994b). This clear understanding about population genetics in *T. cristinae* in terms of the interplay between genotype, phenotype and the surrounding environment (Nosil and Crespi, 2006; Gompert, Comeault, *et al.*, 2014; Comeault *et al.*, 2015) provides a good opportunity to test the key questions raised in this study.

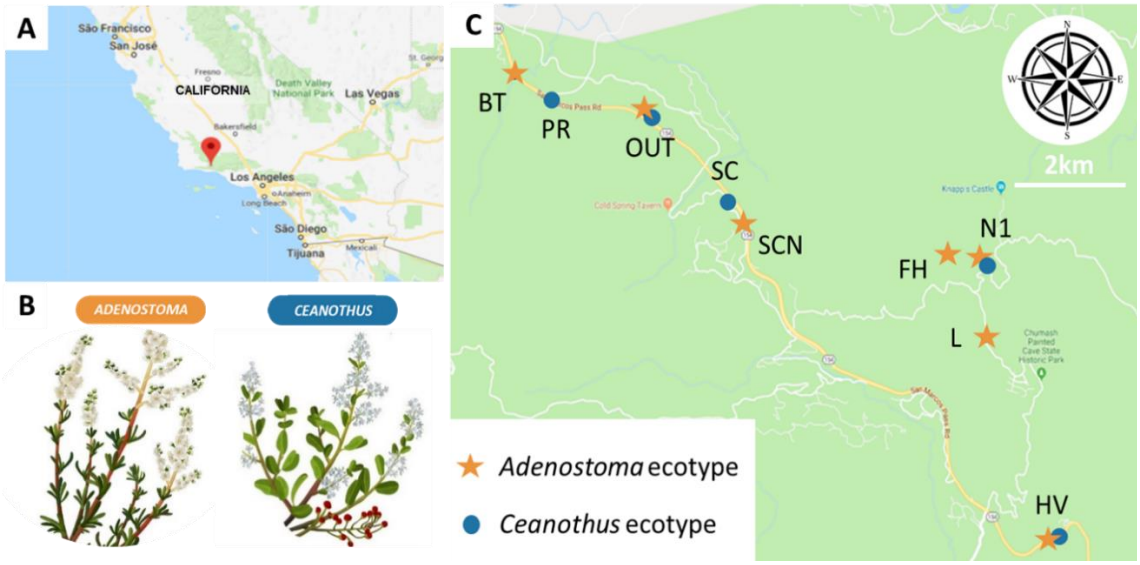


Figure 1: Map detailing geographic position of *T. cristinae* populations included in the sampling plan. (**A)** Location in Southern California where the species is found (Santa Ynez Mountains, Los Padres National Forest). (**B**) Representation of the two main host plants where *T. cristinae* is found, which characterizes the two ecotypes: '*Adenostoma* ecotype' and '*Ceanothus* ecotype' (Nosil *et al.*, 2006). (**C)** Selected populations for the survey. This figure is the same as Fig. 4 in Chapter 1.

3.3.2. Sampling design

A sampling strategy was designed to select *T. cristinae* populations with different combinations of factors that could be shaping methylation variation (Bossdorf *et al.*, 2008; Richards *et al.*, 2017). The aim of this sampling design was not only to capture substantial representation of natural DNA methylation variation, but also to disentangle some of the co-varying factors. Here, a 'population' is defined as all insects collected within a homogeneous patch of a single host species (*i.e.* a locality), as has been done in previous *Timema* studies (*e.g.* Sandoval 1994a,b; Nosil *et al.*, 2002; Sandoval and Nosil, 2005). This work focused on

the following key factors as criteria in the selection of populations: abundance of host plants, elevation, climatic variables and geographical distance between populations.

3.3.2.1. Abundance of host plants

To be able to test the association between *T. cristinae* ecotypes and methylation variation, I selected localities with different abundances of the two host plant species (*i.e. Adenostoma* and *Ceanothus*). *T. cristinae* distribution follows an altitudinal cline, where high elevations are dominated by *Adenostoma* plants while lower elevations are dominated by *Ceanothus* plants. As such, when selecting populations, I considered not only the host plant species at the locality, but also its surroundings: whether the landscape was dominated by *Adenostoma*, *Ceanothus* or if it comprised mixed patches of both plant species (Fig. 2). To obtain this information about abundance of host plants at landscape level, I compared the relative numbers of the two most frequent *T. cristinae* morphs (*i.e.* striped and green). This was based on the fact that in areas where *Adenostoma* is dominant there is a higher frequency of striped individuals as result of local adaptation, whereas more individuals with the green morph are expected in areas dominated by *Ceanothus* (Nosil, 2007; Nosil *et al.*, 2008). Thus, even though a locality is characterized by *Adenostoma* plants, the green allele will probably be more prevalent at a larger scale if the surroundings are dominated by *Ceanothus* patches. Constant migration from the surroundings and gene flow most likely result in a higher frequency of green morphs compared to striped ones in such *Adenostoma* populations (Fig. 2). The database containing information about *T. cristinae* sampling records from previous years (Nosil *et al.*, 2018) was accessed to obtain the relative morph frequencies for each population. The percentage of striped individuals (the '% striped' variable hereafter) was calculated in each population by dividing the total number of striped specimens over the sum of striped plus green individuals, then the mean was generated over the years of sampling. Populations with different levels of '% striped' were chosen from

both *Adenostoma* and *Ceanothus* patches. This design aimed to disentangle effects that are associated with host-plant adaptation from effects originating from migration and gene flow.

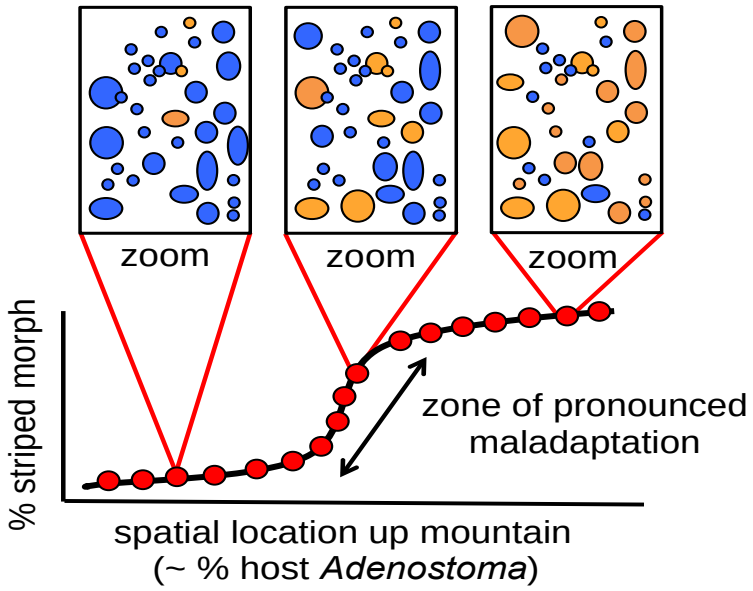


Figure 2: Distribution of host plants and morph frequency along the altitudinal cline in Santa Ynez Mountains. Patches of *Adenostoma* are represented in orange and of *Ceanothus* are represented in blue. There is a growing number of *Adenostoma* patches with increasing elevation, followed by a higher percentage of striped morphs. Thus, the percentage of striped morph in a population tends to represent the surrounding environment in abundance of host plants. Populations of *Adenostoma* and *Ceanothus* ecotypes were selected from different points along this cline for this study. Figure made by Patrik Nosil.

3.3.2.2. Elevation and climatic variables

Based on the evidence in the literature of an association between climate and/or elevation and differential DNA methylation status (*e.g.* Richards *et al.*, 2012; Nicotra *et al.*, 2015), I selected localities that differed in these two environmental factors. Climate information was obtained using the WorldClim database at resolution of 1km^2 for each locality (http://www.bioclim.org). Because the bioclimatic variables were highly correlated, a principal component analysis (PCA) was performed to summarise the total variance between the different localities and reveal the strongest patterns (following

Comeault *et al.*, 2015) using R (3.3.1; R Core Team). All recorded *T. cristinae* localities (Nosil *et al.*, 2018) were used to obtain principal components that represented the general information across the species distribution. Only the first two principal components were retained as they together explained around 92% of the variance (PC1=66.4% and PC2=25.5%). PC1 represents annual temperatures and the temperatures in the coldest and wettest periods of the year. This axis also represents how constant the temperatures are across different seasons (*e.g.* isothermality, minimum temperature in annual range, seasonality), and amount plus seasonality of precipitation. Meanwhile, PC2 represents temperatures in the warmest month, and in the warmest and driest quarters (Fig. 3; Table 1). Elevation was estimated based on the GPS coordinates from the localities using QGIS 2.16.2 (QGIS Development Team 2016).

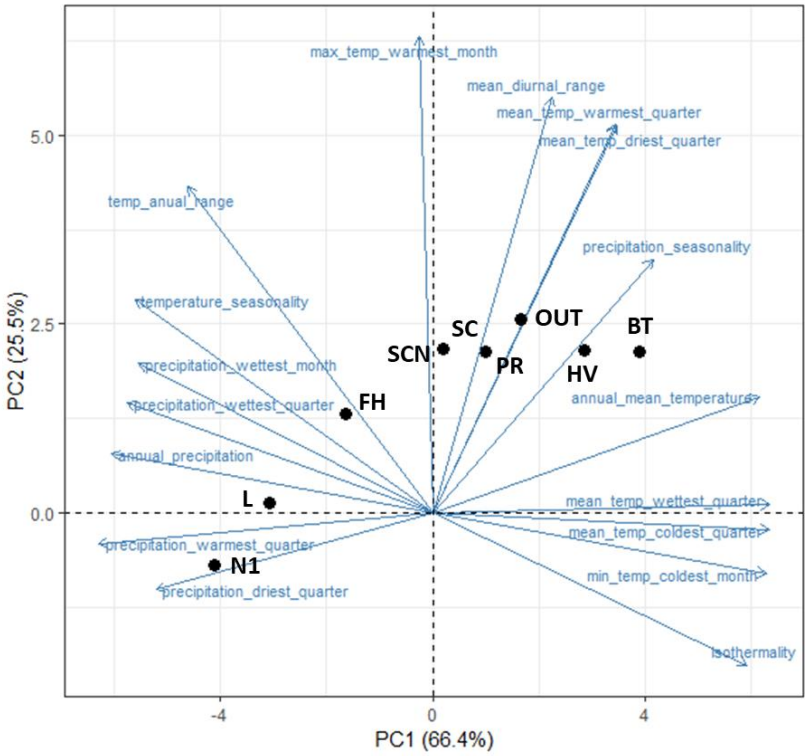


Figure 3: First two principal components on bioclimatic variables (bioclim) represented only on the *T. cristinae* localities used in this study. PCA was performed using all recorded sites (Nosil *et al.*, 2018). Both PC1 and PC2 are strongly correlated with elevation. Correlation between each variable and the main two PCs can be found on Table B1.

A strong correlation between the bioclimatic variables and elevation was found at the candidate localities (PC1 and elevation: adjusted r^2=0.97, P < 2.2e-16; PC2 and elevation:

adjusted r^2=-0.84, P < 2.2e-16, linear models). Thus, elevation was not used in the subsequent analyses, and the first two PCs from bioclimatic variables were used instead.

3.3.2.3. Geographical distance

The localities were also chosen based on the geographic distance between them. Populations tend to be more genetically differentiated the greater the geographic distance between them and the lower the species' dispersal capacity (Jenkins *et al.*, 2010; Shafer & Wolf, 2013). That is, limited gene exchange between populations allows them to accumulate differences arising by drift (Sexton *et al.*, 2014). This rationale applies to genetic differences between populations, because genetic variation is inherited, and so differences accumulate over time particularly between populations with limited gene flow/migration. However, the same logic is challenged when studying DNA methylation variation, given methylation marks are expected to be reset between generations or are imperfectly transmitted (Richards, 2006; Herrera *et al.*, 2014; van der Graaf *et al.*, 2015). In this study, I inquired whether a pattern of isolation by distance, similar to that seen for genetic variation, was present in *T. cristinae* methylation variation, and further investigated the mechanisms that could explain such a pattern. For instance, if methylation variation does not follow an isolation by distance pattern, one can assume there is little heritability of this epigenetic mechanism. On the other hand, if it does, that will point to some heritability of methylation marks (*i.e.* either via epigenetic inheritance, or as a result of genetic control; Herrera *et al.*, 2016). To investigate this, I selected localities that ensured there was variation in geographical distance in the dataset. Distance between localities was estimated based on the GPS coordinates from candidate localities using QGIS 2.16.2 (QGIS Development Team 2016). I selected sites varying from adjacent patches of the two host species in geographic contact with one another (*i.e.* 'parapatric' populations) to patches that were geographically separated by up to 11km (*i.e.* 'allopatric' populations; Fig. 1C).

3.3.2.4. Final selection of localities

Considering all the criteria, 12 localities were selected (Fig. 1C; Table 1). Each locality contains a different combination of the criteria cited above. Here, the '% striped' variable was not significantly correlated with elevation or with either of the climate PCs (elevation: adjusted r^2=0.15, P=0.12; climate PC1: r^2=0.05, P=0.05; PC2: r^2=0.02, P=0.05, linear models). Although there was a noticeable association between '% stripe' and host plant species among the chosen localities (r^2=0.27, P=0.05, linear models), it was much lower compared to when all locations in the Nosil *et al.* (2018) dataset were considered (r^2=0.59; P=5.07e-11, linear models).

Table 1: Populations selected in the sampling plan, including the location code, host plant ('A' for *Adenostoma* and 'C' for *Ceanothus*), geographic coordinates and elevation (metres). Percentage of striped individuals was estimated based on the average relative number of striped individuals within a population in the *T. cristinae* database, which contains sampling records from previous years (Nosil *et al.*, 2018). Climate information was obtained from the first two principal components from the bioclim variables.

Locality	Host	Latitude	Longitude	Elevation	% striped	Climate PC1	Climate PC2
BT	A	34.536	-119.862	306	5.9%	-4.11	1.78
FH	A	34.518	-119.801	813	87.3%	2.30	1.12
HV	A	34.488	-119.787	376	77.9%	-2.79	1.93
HV	C	34.488	-119.787	374	63.4%	-2.79	1.93
L	A	34.509	-119.796	820	90.0%	3.88	-0.28
N1	A	34.517	-119.797	893	70.3%	5.02	-1.32
N1	C	34.517	-119.797	893	39.0%	5.02	-1.32
OUT	A	34.532	-119.843	464	50.8%	-1.44	2.43
OUT	C	34.532	-119.843	462	30.6%	-1.44	2.43
PR	C	34.533	-119.857	364	2.9%	-0.71	1.98
SC	C	34.523	-119.832	570	3.9%	0.24	2.07
SCN	A	34.521	-119.830	585	71.0%	0.24	2.07

3.3.3. Sampling

Individuals from the selected *T. cristinae* populations were all sampled on the same date (25th April 2017) in the Californian spring. The sampling methods and manipulation of the samples were described in Chapter 2. Briefly, specimens were collected using sweep nets and kept in plastic containers at room temperature. The following day, individuals were digitally photographed under standard conditions (Riesch *et al.*, 2017), flash frozen using liquid nitrogen and preserved at -80ºC temperature. All procedures were performed to assure the methylation status was not considerably affected by variation in sampling conditions. This way, one can assume the methylation levels match the patterns present in the wild.

3.3.4. DNA methylation variation

DNA methylation variation was estimated for two female individuals from each population using whole-genome BS-seq (24 individuals in total; Table 1 in Chapter 1). As mentioned before, this high-throughput protocol generates genome-wide information about methylation at a base resolution. Methylation information was estimated for each individual, following the methods detailed in Chapter 2. This workflow involved the removal of potential single nucleotide polymorphisms (SNPs) that could confound the estimate of DNA methylation variation at a specific site (*i.e.* single methylation polymorphisms; SMPs). The final tables for each individual contained information about: number of reads with methylated cytosines (*i.e.* unconverted cytosines), number of reads with non-methylated cytosines (*i.e.* number of thymines), and the proportion of reads with methylated cytosines for each genomic position. This work focused at methylation in CpG dinucleotides because methylation most often targeted this context not only in *T. cristinae*, but also in other animals (Suzuki and Bird, 2008; Feng *et al.*, 2010; Zemach *et al.*, 2010). The final individual tables for CpG sites without the potentially confounding SNPs, had a mean coverage of 2.7

reads per site. 60% of the sites had coverage greater or equal to 2x, dropping to 36% for greater or equal to 3x and then 13% for coverage greater or equal to 5x per site (statistics averaged among all individuals; see Chapter 2).

The function *unite* in the R package methylKit (v1.0.0; Akalin et al., 2012) was used to generate a single table containing methylation information at each site in all 24 individuals (*i.e.* SMPs). That is, information was only retained at sites that were covered in all individuals simultaneously. I removed sites with coverage outliers above the 99.9th percentile to avoid PCR bias (*i.e.* above 60 reads). Very few sites provided methylation information simultaneously among all samples. For example, only 2% of the sites were retained when the minimum coverage of two reads per site was used, and only 0.2% with the minimum of five reads (296,732 and 36,896 SMPs respectively, compared to the number of sites in the final individual tables). Thus, I applied the minimum coverage of two reads (*i.e.* 'less stringent coverage') for analyses using the general genome-wide patterns, to be able to represent a higher methylation variability. For more refined analyses comparing each site individually, I applied the minimum coverage of five reads (*i.e.* 'more stringent coverage') aiming to preserve more information at each SMP. All the reported statistics were generated using R (3.3.1; R Core Team 2016).

3.3.5. Genetic variation

Restriction site associated DNA sequencing (RAD-seq) was used to generate genome-wide single nucleotide polymorphism (SNP) data, following previous studies in the system (Comeault *et al.*, 2015, 2016). For this, I used DNA from the exact same individuals used to generate the methylomes (Table 1 in Chapter 1). This way, information about DNA methylation and genetic variation was available for each individual. To obtain a better estimation of genetic diversity at population level, I expanded the sample size by reanalysing RAD-seq datasets from previous studies in *T. cristinae* (Comeault *et al.*, 2015; Lindtke *et al.*, 2017; Riesch *et al.*, 2017). Around 15 accessions were randomly selected from

each population with previously published data (Table B2), and new RAD-seq data were acquired from the populations that had not been previously sequenced (*i.e.* BT, OUT, SC, and SCN populations; Table B2). The latter were collected along with the individuals used to obtain the methylomes (sampled in 25th April 2017), and preserved in 100% ethanol at -20^{0}C. Finally, genetic data for the six individuals used in the rearing experiment (Chapter 4) were also processed with these other datasets. Details about how these data were processed were described below. In the end, SNPs from all datasets were called together, which ensured there were enough samples from different populations to reliably calculate the genotypic probabilities.

3.3.5.1. Library preparation and sequencing

DNA from the 24 individuals used for BS-seq was extracted following the procedures described in Chapter 2. A subsample of this extraction was used in RAD-seq protocol to produce genetic variation data. As mentioned above, around 15 individuals from each population that had not been previously sequenced (Table B2) were used to obtain population level genetic variation. Genomic DNA was isolated using DNeasy Blood and Tissue Kits (Qiagen). Library preparation was done using a combination of Parchman *et al.*'s (2012) protocol designed for Illumina sequencing, which has been successfully implemented in *Timema* stick-insects (Comeault *et al.*, 2015; Riesch *et al.*, 2017) and the protocol of Peterson *et al.* (2012), modified to produce paired-ended libraries. Following these protocols, genomic DNA was first digested with the restriction endonucleases *EcoRI* and *MseI* (New England Biolabs). The samples were then incubated with T4 DNA ligase (New England Biolabs) and with the following oligonucleotides: (1) Illumina adapter sequences followed by custom barcodes with 8-10 base pairs plus some extra base pairs to adjust to *EcoRI* cut sites; and (2) adapters to the *MseI* cut site with standard Illumina multiplexing read index (adapted from Peterson *et al.*, 2012). The resulting fragments were amplified by polymerase chain reaction (PCR) using 20 cycles, and then pooled, resulting in

an individually barcoded restriction-site associated DNA library. This library was sequenced using an Illumina HiSeq2000 platform with V3 reagents at the National Center for Genome Research (Santa Fe, New Mexico, USA). After retrieving the Illumina sequencing reads, I removed the barcodes and the *EcoRI* cut site base pairs from the reads (following a Perl script developed in Nosil *et al.,* 2012) and split the reads by individual.

3.3.5.2. Variant calling and genotypic probabilities

RAD-seq datasets from previously published studies (Comeault *et al.*, 2015; Lindtke *et al.*, 2017; Riesch *et al.*, 2017; Table B2) were re-analysed to estimate genetic diversity at population level; and they were processed along with the newly acquired data from this step onwards. In summary, the good quality reads were aligned to the most recent *T. cristinae* reference genome (Nosil *et al.* 2018) using bowtie 2.3.4.1 (Langmead and Salzberg, 2012) with the single-end or paired-end argument depending on the library type. The mapped reads were sorted and indexed using SAMTOOLS 1.8 (Li *et al.*, 2009). Variants were called following a custom Perl script (Comeault *et al.*, 2015), which uses SAMTOOLS *mpileup* and BCFTOOLS using the full prior, and requiring the probability of an allele to be lower than 0.5 to call a variant, under the null hypothesis that all samples were homozygous for the reference allele. The insertion and deletion polymorphisms were not included in the final table. For each variant, the posterior mean genotype was estimated for each individual at each locus as two times the probability of the homozygous minor allele genotype plus the probability of the heterozygous genotype. These steps led to 3,870,412 SNPs in total. From those, only 533,420 were retained after discarding SNPs for which there were sequence data for less than 50% of the individuals, low confidence calls with a phred-scale quality score lower than 20, SNPs with more than two alleles, and filtering for a minor allele frequency of 0.01. Custom Perl scripts were used along with a custom C++ program (alleleEst 0.1b) to estimate the genotypic probabilities using a Bayesian model (Gompert *et*

al., 2013). The values were stored in BIMBAM format with values ranging from 0 to 2 representing the minor allele dosage.

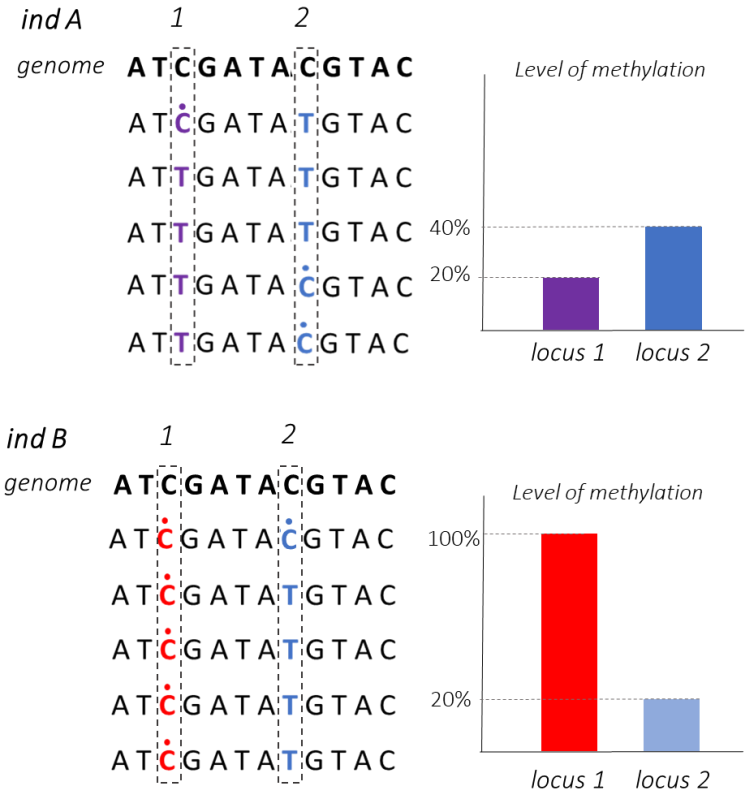


Figure 4: Hypothetical example of differences in DNA methylation levels between two samples, here represented by *indA* and *indB* and in the loci *1* and *2*. Levels of methylation in each locus are estimated based on the proportion of read counts with methylated cytosines (*i.e.* number of non-converted cytosines in BSseq data) over the total coverage (*i,e,* 5 reads in this example).

3.3.6. General patterns of geographical structure

3.3.6.1. Clustering analyses

Hierarchical clustering analyses were used to estimate general patterns of DNA methylation variation. For this analysis, the methylation levels were estimated using the proportion of read counts with methylated cytosines over the total coverage in every genomic position (Fig. 4). Euclidean distances were then calculated to estimate the dissimilarity between pairs of samples (Eq. 1, Fig. 5) using the function *dist* in R (3.3.1; R Core Team 2016):

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \quad (1)$$

where *A* and *B* represent two hypothetical individuals to have the Euclidean distance *d* estimated based on methylation levels of *n* sites (see Fig. 4 for the example using two loci). Here, this operation was done pairwise among the 24 individuals in the methylation levels of 296,732 sites. The resultant distance matrix was used as input for the agglomerative hierarchical clustering analysis. Analyses were performed using *hclust* function in R using the 'Ward D' agglomerative criterion. In addition, a principal components analysis (*prcomp* function in R) and k-means were used to evaluate the grouping trend (*kmeans* function).

$$\boldsymbol{d(indA, indB)^2} = (A_1 - B_1)^2 + (A_2 - B_2)^2$$

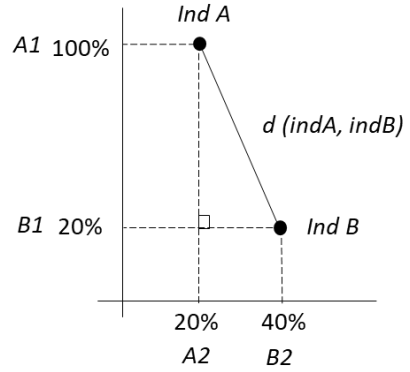


Figure 5: Euclidean distance between individuals A and B from Fig. 4 The distances are estimated at each locus, then summed to obtain the squared distance. The Euclidean distance in this example is 0.82, considering percentage using decimals. The same method can be used for *n* loci.

3.3.6.2. Regression analysis

Distances between methylation levels were estimated for every pair of individuals using the Euclidean distance function *dist* in R (Figs. 4-5). These distances were compared with other distances for every pair of individuals, namely climatic, genetic, host plant species, and geographical distances. The climatic distances were also estimated using Euclidean distances, based on the first two principal component axes. The genetic distances were obtained using the RAD-seq data. The aligned reads from the 24 individuals were used

to obtain the genetic distance matrix from RapidNJ (2.3.0.2; Simonsen *et al.*, 2008). This software calculates pairwise evolutionary distances between individuals to ultimately generate a neighbour-joining tree that represents the given distance matrix as well as possible. Its algorithm takes multiple alignment nucleotide sequences as input and counts the observed number of purine-transitions (A and G), of pyrimidine-transitions (C and T) and of transversions (purine to pyrimidine or vice versa). Once these observed mutational events have been counted, the most likely (ML) estimate of the distance between two sequences s_1 and s_2 is computed based on Kimura's two-parameter model of sequence evolution (Kimura, 1980; Elias and Lagergren, 2007; Simonsen and Pedersen, 2011):

$$d_{(s1,s2)} = \frac{1}{2} ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4} ln\left(\frac{1}{1-2Q}\right) \quad (2)$$

where P and Q are the rate of transitions and transversions respectively. The resultant pairwise distances are all compiled in the genetic distance matrix (Simonsen *et al.*, 2008). Host plant differences were coded for each pair of individuals as '0' if they were collected in the same host plant species or as '1' if they were different ecotypes. Finally, geographical distances were estimated in metres using QGIS "*Point Distance*" tool from the "*Vector... Analysis Tools*" menu. The logarithm of the pairwise geographical distances (in metres) was calculated to perform the regression between genetic and methylation distances and geographical distances (following Rousset, 1997). To avoid calculating logarithm of 0, individuals from the same population (*i.e.* geographic distance of 0 metres) were considered to be 128 metres apart, as this is the maximum dispersal ability within their lifetime (Sandoval, 2000). The statistical tests between different matrices were estimated using linear models in R.

To complement the linear models' regressions, the correlation between elements in the distance matrices was assessed by a Mantel randomization test. This test randomizes the *n* individuals rather than using the pairwise observations of different variables (Mantel, 1967). That is, it computes the significance of the correlation through permutations of the

rows and columns of one of the input distance matrices. The statistic test is the Pearson correlation coefficient *r*. This analysis provides a test of the null hypothesis of no linear relationship between two distance matrices. The tests were performed using *mantel* function from *vegan* R package (Oksanen *et al.* 2018), using 10,000 permutations.

3.3.6.3. Bayesian regression

In addition, a multivariate analysis using Bayesian regressions was performed to determine whether methylation differences between individuals were better explained by genetic variation, geographical distance, climatic distances, ecotype, or combinations between the variables. It was used to correct for any bias in the data introduced by dependency among the pairwise data points. The model was based on Clarke *et al.* (2002) and Gompert *et al.* (2014), and follows the equation:

$$\text{Logit}\left(Y_{ij}\right) = \beta_0 + \beta_{gen}X_{ij}^{gen} + \beta_{geo}X_{ij}^{geo} + \beta_{clim}X_{ij}^{clim} + \beta_{host}X_{ij}^{host} + \lambda_i + \lambda_j + \epsilon_{ij} \quad (2)$$

where the dependent variable (Y_{ij}) was the methylation distances between individuals *i* and *j*. The variables X_{ij}^{gen}, X_{ij}^{geo}, X_{ij}^{clim}, X_{ij}^{host} are the pairwise distances in genetic variation, geographical distance, climatic distances and host plant species, respectively, and the β denote the fixed effect regression coefficients. The model includes λ_i, which is a random effect representing the average deviation of the pairwise Y values (*i.e.* methylation distances) involving individual *i* from what is expected from its X distances to the other individuals. In other words, the λ terms represent the dependency in the data, accounting for the pairwise comparisons by allowing each individual to have its own deviation from the baseline expectation (Clarke *et al.*, 2002; Gompert *et al.*, 2014). ϵ_{ij} represents the residual errors, and λ_i and ϵ_{ij} are assumed to be independent. All variables were centred and standardized. The model uses Bayesian framework and Markov Chain Monte Carlo to estimate the regression, and a deviance information criterion (DIC) to evaluate model fit via *rjags* R package (Plummer, 2003). Three parallel chains with 10,000 iterations and a burn-

in of 2,000 iterations were used. The equation above represents the full model, but the analyses were also performed using different combinations of the variables.

3.3.7. Binomial Mixed Models

To estimate the heritability of methylation levels at each site, I used the Mixed model Association for Count data via data AUgmentation (MACAU) method, developed by Lea *et al.* (2015). Briefly, this model tests whether a variable (predictor) has effect on methylation levels at a specific site, and allows us to control for relatedness (*i.e.* kinship) in the samples. I tested for a relationship between ecotype and DNA methylation levels at each site using the 24 individuals from the population survey. In this Chapter, I interpret the model's outputs to estimate mean heritability of methylation levels. In Chapter 4, I evaluate the associations between the SMPs and ecotype arising from the same analysis. This separation was done to avoid overlap between the two Chapters.

3.3.7.1. MACAU

The model tests whether a variable (predictor) has an effect on methylation levels at a specific site. For instance, the predictor of interest can be a phenotype, an environmental factor or genotypic values. Its binomial model can handle methylation count data, modelling the number of reads with methylated cytosine (y_i) and the total coverage (r_i) to estimate the level of methylation (π_i) for each site:

$$y_i = \text{Bin}\ (r_i, \pi_i) \qquad (3)$$

Thus, here the variability in coverage is used to estimate the methylation levels. Given that the coverage may differ considerably across sites and individuals, this approach offers many advantages compared to other models, which tend to ignore this issue by using proportion of methylated cytosines to estimate the methylation status. Using the same example provided by the developers: if only the methylation proportion is considered, a site where

5 out of 10 reads are designated as methylated is treated identically to a site where 50 out of 100 reads are designated as methylated, even though the accuracy of the estimated proportion will be worse when there are fewer reads. This assumption reduces the power to detect true associations between a predictor and the variation in methylation levels (Lea *et al.*, 2015).

In addition, MACAU can control for population structure when testing for the relationship between methylation variation and the predictor. DNA methylation levels are often heritable. In humans, for example, the average estimated heritability levels are 18-20% in the whole blood (Taudt *et al.*, 2016). As such, closely related individuals will tend to exhibit more similar methylation patterns than non-related individuals will. The similarity in methylation levels in SMPs or genomic regions (*i.e.* methylation 'relatedness') can arise from the genetic control determining methylation patterns, or via pure epigenetic inheritance, where methylation patterns are not reset and are transmitted to the next generation (Jablonka and Raz, 2009; Taudt *et al.*, 2016). As such, analyses that do not consider relatedness can lead to spurious conclusions if the predictor in question co-varies with kinship. To control for this, MACAU incorporates a matrix of pairwise genetic kinship, which is treated as the variance-covariance matrix for the heritable component of the random effects' variable. The kinship matrix contributes to the value of the response variable, but does not affect the non-heritable part of the response variable (Lea *et al.*, 2015; Lea *et al.*, 2017). The kinship matrix can be estimated using the genetic variation in the dataset, and modelled as the 'genetic random effects'.

The methylation levels at each site are modelled in MACAU using a logit link linear function based on the following variables:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = w_i^T\alpha + x_i\beta + g_i + e_i \qquad (4)$$

$$g = (g_1, \dots, g_n)^T \sim \mathrm{MVN}(0, \sigma^2 h^2 K) \qquad (4)$$

$$e = (e_1, \dots, e_n)^T \sim \mathrm{MVN}(0, \sigma^2(1-h^2)I) \qquad (6)$$

In equation 5, w_i corresponds to c-vector of covariates, and α corresponds to its coefficients; x_i is the predictor of interest for individual i and β is its coefficient. Variable g is an n-vector of genetic random effects due to population structure or kinship, and e is an n-vector of environmental independent noise. As mentioned above, the genetic random effects g are estimated using a relatedness matrix K, which can be calculated based on genotypic data following tools described in Zhou *et al*. (2013). Its $\sigma^2 h^2$ element corresponds to the genetic variance component, where h^2 is the heritability of logit(π) at each site. K has been standardized in the model to ensure tr(K)/n = 1, so that h^2 lies between 0 and 1 and can be interpreted as heritability of the methylation levels (Zhou et al. 2013). *MVN* represents the multivariate normal distribution applied when g and e are estimated. Ultimately, MACAU tests the null hypothesis $H_0 : \beta = 0$ for every site, using a MCMC algorithm based approach to determine an approximate maximum likelihood estimate $\hat{\beta}$, its standard error se($\hat{\beta}$) and its corresponding *p-value*. The MCMC sampling steps are also used to produce uninformative priors to estimate the heritability h^2 and its standard error se(h^2).

3.3.7.2. Estimating heritability

To run MACAU, I used the table with the most stringent coverage (*i.e.* minimum of five reads per site) to assure a higher power analysis at each SMP locus. To assure the SMPs were variable enough to run the model, I selected sites where at least two individuals had methylation levels above 25% (> 0.25) or below 75% (< 0.75). This step excluded the sites that were consistently hypomethylated or consistently hypermethylated (following Lea *et al.* 2016, see rationale at Appendix B 'Testing MACAU' section). This step retained 35% of the CpG sites with the selected coverage (yielding 13,050 SMPs in total).

The first two principal component axes of climatic variation were used as covariates along with bisulfite conversion efficiency estimated for each sample (using the non-methylated Lambda phage; see Chapter 2, Table A1). The genotypic probabilities were

extracted from the same 24 individuals from the BIMBAM file to calculate the kinship matrix following Zhou *et al.* (2013). To obtain an estimate of genome-wide heritability of methylation variation, I calculated the mean of the outputted parameter h and its corresponding standard error (used to estimate the 95% confidence intervals). The analyses were performed with 100,000 sampling steps and burn-in of 50,000 iterations, with the filtering ratio threshold equal 1.

3.4. Results

3.4.1. General patterns of methylation in natural populations

3.4.1.1. Clustering analyses

The hierarchical clustering analyses showed considerable differences between the individuals, represented in the dendrograms by the height of dissimilarities between the tips (Fig. 6A), suggesting a considerable intraspecific variability in the methylation levels. These findings are supported by the principal components analysis, as each PC axis almost equally explains the variance in the data around 4.5% [4.0% – 5.0%] (mean [95% confidence interval]; Fig. B2). At first sight, the dendrograms appear to group individuals according to their position in geographical space (Fig. 6). The k-means analyses (using k=2) separate two outlier individuals (*i.e.* 17_0077 and 17_0086) from the others. If these two samples are excluded, the k-means analysis groups specimens coming from eastern and western parts of the distribution (Fig. B3). This pattern supports the hypothesis the methylation variation is structured in geographical space.

3.4.1.2. Regressions

The regression analyses indicated genetic diversity is associated with geographic distance (r^2=0.29, P < 2.2e-16, linear models; Fig. 7A), as has been previously shown (Sandoval, 1994; Nosil *et al.*, 2008). Following the results from clustering analyses, the

regressions revealed methylation distances are also positively correlated with geographical distances (r^2=0.09, P = 1.2e-12, linear models; Fig. 7B). That is to say methylation and

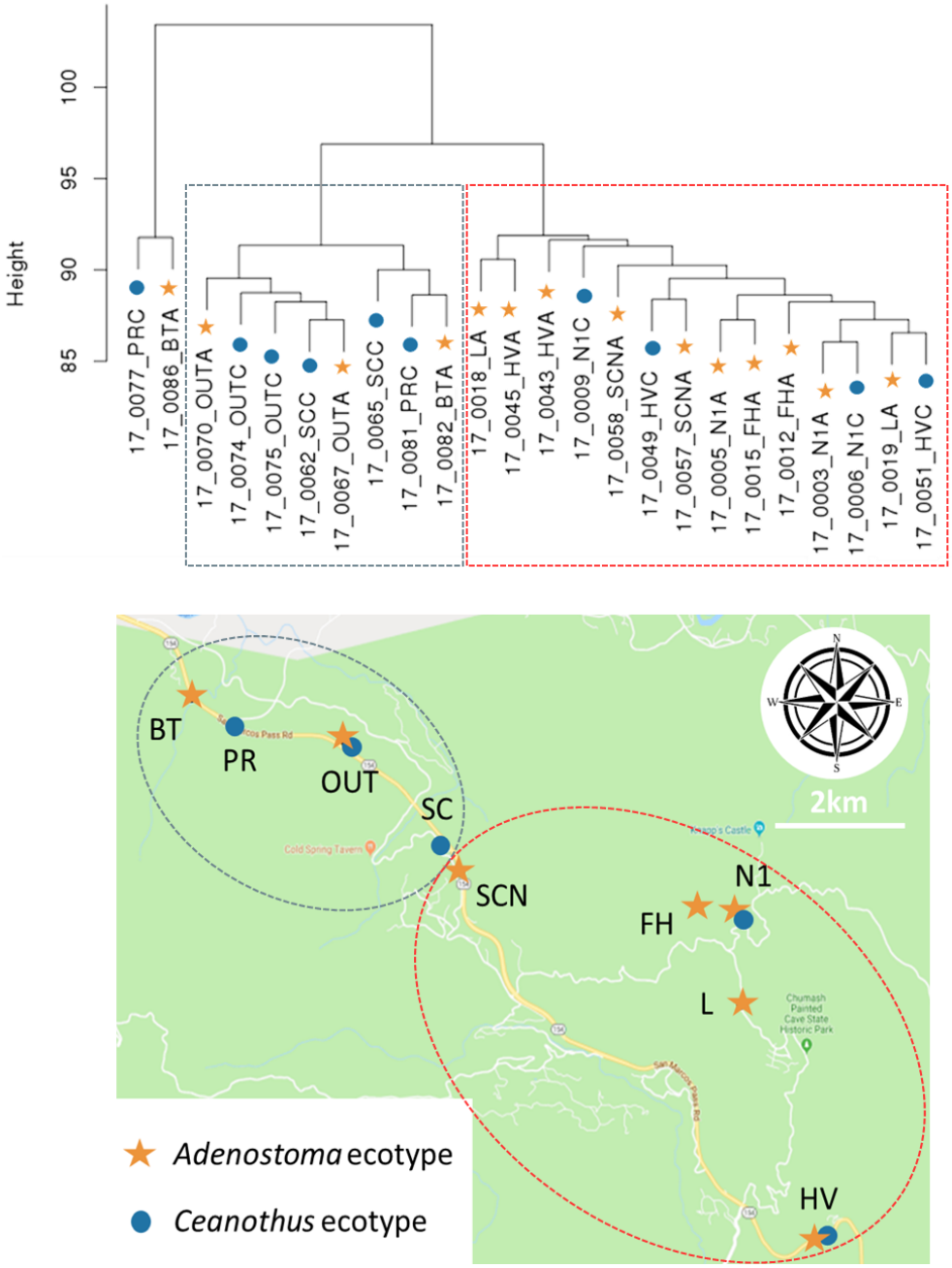


Figure 6: Hierarchical clustering in the population survey using Euclidean distances to estimate dissimilarity and 'Ward.D' algorithm of clustering. The table with the less stringent coverage was used to include more sites in the analysis (*i.e.* minimum coverage of 2 reads, 296,732 sites in total). Dendrogram height represents the Euclidean distance between individuals (Eq. 1, Fig. 5). The mean pairwise Euclidean distance was 90.1 [89.9 – 90.4]. Individuals do not tend to cluster by ecotype, but by geographical position. Blue line colour corresponds to the western side of the distribution and red to the eastern side. In the dendrograms, the last letter of the population code corresponds to ecotype (*e.g.* BTA represents the *Adenostoma* ecotype and PRC the *Ceanothus* ecotype).

genetic distances tend to co-vary, with a pattern following isolation-by-distance premises (r^2=0.42, P < 2.2e-16, linear models; Fig. 8). Accordingly, the Mantel tests revealed a strong significant correlation between methylation and geographic distances (r=0.30, P = 1.0e-06), between genetic and geographic distances (r=0.45, P = 1.0e-06), and between methylation and genetic distances (r=0.65, P = 1.0e-06). This pattern could also be associated with the climatic characteristics in geographical space. The results show the relationship between geographic distance and climatic distances is significant (r^2=0.23, P<2.2e-16, linear models; r=0.4, P=1.0e-06 Mantel test). However, regarding the correlations with climatic variation, neither genetic variation (r^2=0.02, P=4.3e-02, linear models; r=0.12, P=0.11, Mantel test) or DNA methylation differences seem to be associated with climatic distances (r^2=0.0, P=0.65, linear models; r=0.02, P=0.41, Mantel test).

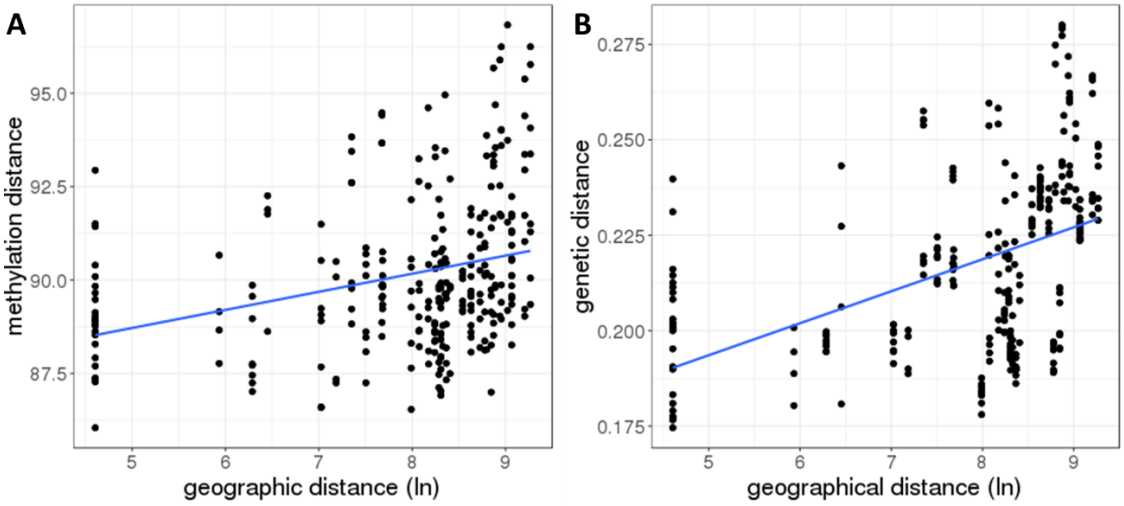


Figure 7: Relationship between **(A)** genetic differences and the logarithm of geographic distances (r^2=0.29, P < 2.2e-16, linear models), and between **(B)** methylation levels and the logarithm of geographic distances (r^2=0.09, P < 1.2e-12, linear models). Methylation distances were estimated from Euclidean distances based on methylation levels, and genetic ones from neighbour joining tree built based on RAD-seq alignments.

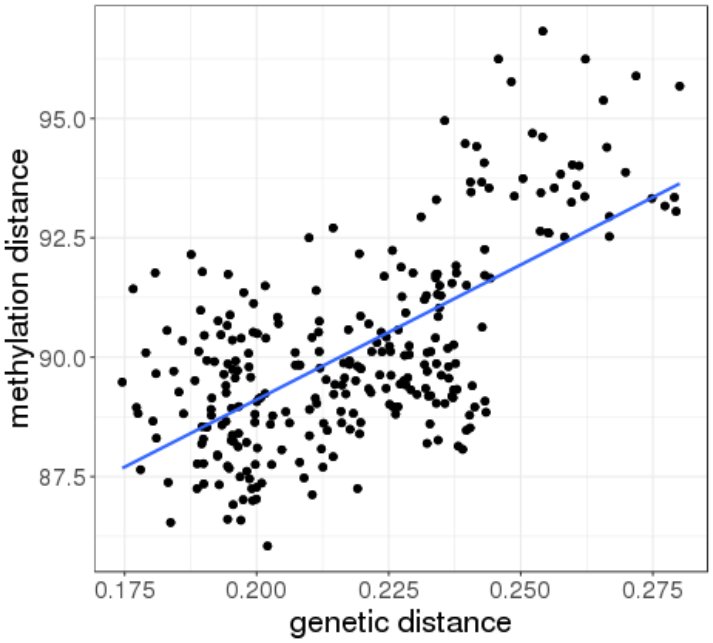


Figure 8: Relationship between methylation differences and genetic distance. Methylation distances were estimated from Euclidean distances based on methylation levels, and genetic ones from neighbour joining tree built based on RAD-seq alignments. r^2=0.42 (P < 2.2e-16, linear models).

3.4.1.3. Bayesian regression

Results from the Bayesian regression analyses showed the model with genetic variation and host differences together received the best support explaining methylation variation, according to the lowest DIC (325.4; Table 2). However, the DIC values were always reduced whenever genetic variation was included in the model, suggesting it had a significant explanatory power in defining the underlying general differences in DNA methylation patterns. Thus, although there was a marked correlation between genetic variation and geographical distance, the reduced model with only genetic distances (DIC=325.7) was better than the one with logarithm of geographical distances only (DIC=523.0).

3.4.2. Heritability of methylation variation

The MACAU analysis returned a high estimate of genome-wide heritability of methylation levels, with a mean heritability of 0.67 across all sites and mean standard error

of 0.26 (*i.e.* 95% confidence interval of 0.15 – 1.0). The large confidence intervals are expected given a relatively small number of individuals was used in the analyses (Lea *et al.*, 2015). The model estimates the heritability of methylation levels based on the kinship matrix obtained from the genetic differences between individuals. Thus, the similarity in methylation variation seems to mirror the genetic kinship, supporting the hypothesis of genetic background explaining DNA methylation variation. However, it is possible that part of this relatedness is due to the non-erasure of epigenetic marks between generations (*i.e.* pure epigenetic inheritance). In other words, MACAU might not be able to distinguish the heritability derived from the genetic control from pure epigenetic inheritance. Therefore, further experiments are required to disentangle the drivers of the heritability in *T. cristinae*.

Table 2: Bayesian regression using differences in genetic variation, geographical distance, climate and host plants to explain methylation variation. Full model uses all variables simultaneously. Deviance information criterion (DIC) was used to compare the models. Models including genetic variation had a better support (lowest DIC) than when it is not considered.

	Mean deviance	*Penalty*	*DIC*
Full model	299.1	29.1	328.2
Genetic + geography + climate	300.7	28.2	328.8
Genetic + geography + host	298.2	28.0	326.2
Genetic + climate + host	299.5	29.6	329.1
Geography + climate + host	490.4	28.1	518.5
Genetic + geography	300.0	27.3	327.3
Genetic + climate	298.8	26.7	325.5
Genetic + host	297.8	27.6	325.4
Geography + climate	491.6	27.0	518.6
Geography + host	495.5	27.6	522.8
Climate + host	595.1	26.8	621.9
Genetics	299.1	26.6	325.7
Geography	496.8	26.2	523.0
Climate	596.6	26.4	623.0
Host	621.2	26.4	647.6

3.5. Discussion

Although there is mounting evidence of the molecular mechanisms and biological processes involved in DNA methylation (Bird, 2007; Law and Jacobsen, 2010; Schübeler, 2015), how these processes happen in nature remains largely unknown. Studies exploring DNA methylation from an ecological and evolutionary perspective have started to elucidate how it varies in natural populations and to identify some of the drivers underlying this diversity (Richards, 2008; Nicotra *et al.*, 2015; Platt *et al.*, 2015; Foust *et al.*, 2016; Verhoeven *et al.*, 2016). However, the great majority of these studies were performed in plants. This work assessed the general aspects of genetic and epigenetic diversity in natural populations of *T. cristinae* (*i.e.* a genome-wide comparison). To my knowledge, this was the first investigation on patterns of methylation in natural populations of insects.

3.5.1. DNA methylation is structured in geographical space

The population survey showed substantial intra-specific DNA methylation variation in *T. cristinae*. The marked differences between individuals indicate each one of them has a very specific methylation background. These findings suggest DNA methylation varies within population. In this work, only two individuals were used to represent a population, hence future studies with a larger sample size should be carried out to determine the extent of this diversity. This study focused at investigating how DNA methylation variation is distributed in geographical space to capture the effects of forces that are acting on it, possibly cumulatively over many generations (Herrera *et al.*, 2016). The results showed DNA methylation variation was spatially structured, and revealed a marginal trend to group according to the geographical distribution. In addition, DNA methylation differences increased with geographical distance. This pattern is similar to previous works in *T. cristinae*, which showed that genetic divergence is consistently and significantly greater between allopatric populations than between adjacent populations (Sandoval, 1994b; Nosil

and Crespi, 2004). In other words, the results suggested that spatial structure in DNA methylation variation could be driven, to some extent, by the same factors driving genetic variation. That is to say differences tend to accumulate the more distantly the populations are separated, due to the effect of limited dispersal and reduced gene exchange (Jenkins *et al.*, 2010; Herrera *et al.*, 2016).

3.5.2. DNA methylation is strongly associated with genetic background

This resemblance between DNA methylation and genetic patterns in geographical space was reflected by the strong correlation between them. Although genetic differentiation and geographical distance are intimately interconnected, the multivariate analyses revealed the models with genetic variation were always preferred over the models with geographical distances. In other words, this result suggests DNA methylation variation in *T. cristinae* varies in space mostly because of its genetic basis, and not simply because stochastic epigenetic differences accumulated between further apart populations. It is challenging to disentangle genetic variation from physical geographic distance with a sampling design such as the one used in this study. To this end, one of the factors should be controlled for to test the strength of this association. For example, by further sampling from one single population to ensure there is no geographical variation, but that genetic variation is present.

The strong correlation between methylation and genetic variation suggests a substantial amount of DNA methylation variation in *T. cristinae* is determined by its genetic basis. In other words, that some of the methylation patterns could be under genetic control, either by factors that *cis* or *trans*-regulate methylation state. That is, even though methylation differences might result in important changes in the phenotype, if it is strictly under genetic control, it would represent a proximate cause of these changes, and not the ultimate cause (Richards, 2006). In plants, for example, the association between genetic and methylation variation is due to SNP alleles in structural variants, such as transposable

elements (TEs) insertions, repeats or inversions, which tend to be highly methylated (Pecinka *et al.*, 2013; Taudt *et al.*, 2016). In addition, methylation variation associated with ecologically relevant traits and local adaptation has been shown to be strongly associated with a few genetic variants, including one plant-specific methyltransferase. In mammals, some SMPs have been reported to be sequence dependent, and this is normally related to differential transcription factor binding in enhancers or promoters – regions that are regulated according to the methylation state (Taudt *et al.*, 2016; Onuchic *et al.*, 2018).

However, there might be cases where the associated genetic background facilitates the epigenetic change. That is, when a genetic mutation or a transposable element insertion on a regulator gene occurs and creates a facilitating change to be modulated by the methylation state. A good example is the already cited case in mice about diet and coat colour changes that involves a TE that is inserted upstream of the *Agouti* gene (Waterland and Jirtle, 2003). The methylation state of this transposon leads to different expression levels of *Agouti*, which means that there can be no epigenetic variation if this TE is absent. Thus, in this case methylation variation is associated with genetic variation, but the genotype alone cannot explain the phenotype (Richards, 2006). Therefore, future analyses should be conducted in *T. cristinae* to deepen the understanding of this association between DNA methylation and genetic variation.

Investigating the relationships between methylation and genetic variation in *T. cristinae* could elucidate how they are likely to occur in insects. This study revealed a strong association between genetic and methylation variation at genome-wide levels. Further investigations should estimate this association at a finer scale, such as to look for a co-variation between specific SNPs and SMPs. With these analyses, it will be possible to identify the genetic bases of DNA methylation patterns and the regulatory mechanisms determining them (*i.e. cis* or *trans* regulatory pathways affecting methylation levels), and to investigate its contribution to phenotype – how much it depends on genetic variants to generate phenotypic variation.

3.5.3. Genome-wide methylation variation is not strongly associated with climate or host plant

This study showed there was not a significant correlation between the pairwise differences in methylation and in climatic variables or host plant. Hence, these environmental factors do not explain the genome-wide variation in DNA methylation as much as genetic variation. However, it can be difficult to distinguish whether the association between epigenetic and genetic variation is a direct effect of genetic control, or if there are common factors shaping DNA methylation and genetic variation simultaneously, such as an environmental factor (Richards *et al.*, 2017). For example, some genetic and epigenetic variation could be jointly affected by natural selection, and thus be significantly correlated because of local adaptation. Studies have dealt with this issue by using reproductive inbred lines or asexual populations. In the absence of genetic variation, researchers could estimate the significance of the correlation between environment and methylation variation (Massicotte *et al.*, 2011; Herrera *et al.*, 2012; Yu *et al.*, 2013; Preite *et al.*, 2015). Others have evaluated populations with low levels of genetic variation, such as following a recent bottleneck (*e.g.* invasive species; Richards *et al.*, 2012; Liebl *et al.*, 2013), or without any association between genetics and the environmental gradient (Foust *et al.*, 2016).

This present work focused on exploring the genome-wide associations between DNA methylation and environment. Thus, although the association between DNA methylation and environmental variables was low, the results here did not discard the possibility that some regions in the methylome were associated with environment. Chapter 4 explores the association between DNA methylation variation and ecotype in natural populations at a finer scale (*i.e.* at each SMP), where the results from MACAU were interpreted. In addition, it details an experiment where specimens sampled from one population were reared either on *Ceanothus* or on *Adenostoma* under controlled conditions. This provided the opportunity to directly test the effects of an environmental factor on DNA methylation variation.

3.5.4. There is some heritability of DNA methylation patterns

We estimated heritability of DNA methylation patterns using MACAU (Lea *et al.*, 2015), which models methylation levels at each site and tests for an association with a predictor of interest. It uses genetic information to estimate the kinship matrix, which is included in the model as a random effect used to estimate the heritability parameter (Lea *et al.*, 2015). The analyses returned a significant mean heritability value. This result aligns with the finding DNA methylation is structured in space, and supports the idea that DNA methylation variation can accumulate over generations and tends to be more differentiated the more isolated the populations are (Herrera *et al.*, 2016). At the same time, this result follows the strong correlation between methylation variation and genetic variation, suggesting heritability of methylation patterns could be elevated because it is associated and/or determined by genetic variants. On the other hand, MACAU statistically estimates the heritability of methylation status by comparing the relatedness in methylation variation between samples to the genetic kinship matrix. This means the model does not identify the drivers underlying this heritability, whether it was the result of genetic control or via an incomplete erasure of methylation marks, and consequent transmission across generations. That is, although this pure epigenetic inheritance is independent of genetic control, it could lead to the accumulation of differences at the population level (Verhoeven *et al.*, 2010; Herrera *et al.*, 2013; Jiang *et al.*, 2013).

Very little is understood about epigenetic inheritance in insects. In fact, it is not known whether DNA methylation reprogramming occurs in insects' gametogenesis. Wang *et al.* (2016) showed stable inheritance of methylation status between generations in *Nasonia* wasps. They revealed that F1 hybrids retained patterns of DNA methylation on alleles that are specific of each parental *Nasonia* species with near "100% fidelity". With this, the authors suggested that either DNA sequence elements in *cis* determined these differential methylation patterns, or they were transmitted across generations via pure epigenetic inheritance. Moreover, Chapter 2 discusses the fact *T. cristinae* and other insect species do

not present the *de novo* DNA methyltransferase. As such, one of the hypotheses that can be raised is that DNA methylation patterns are not erased during gametogenesis, implying maintenance of methylation status across generations and evolutionary time. However, this hypothesis is still speculative, and thus future investigations should be carried to elucidate the molecular mechanisms of epigenetic inheritance in insects. In any case, the heritability of DNA methylation variation in *Timema* suggests it might provide background for evolutionary processes to act on.

3.6. Conclusion

Besides its taxonomically different methylation profile from plants and vertebrates, *T. cristinae* methylation patterns are spatially structured in nature, varying with geographical distance in a very similar trend to that exhibited by genetic variation. Added to the finding that methylation variation is significantly heritable, these results indicate DNA methylation follows a pattern analogous to isolation by distance, accumulating differences with reduced dispersal and gene flow. The results presented here suggest this structure is mainly explained by the genetic variation underlying DNA methylation patterns, which could also account for the heritability of methylation status. Future investigations should be carried to explore the associations between DNA methylation and genetic variations at a more refined scale. In this study, I did not find a strong association between DNA methylation variation and environmental factors, such as host plant or climatic variables. It is possible to conclude that the environmental factors studied here are not a very significant driver of general patterns of methylation variation (*i.e.* genome-wide patterns), without discarding their relevance to specific regions in the methylome.

Appendix B: Supplementary Tables and Figures – Chapter 3

Table B1 describes the relationship among each climatic variable from bioclim database and the first two principal components used in this study. The principal components analysis was performed based on climatic data from all recorded *T. cristinae* populations (Nosil *et al.*, 2018).

Table B1: Principal component analysis on climatic variables from bioclim database. Here are represented the correlation between the variable and the first two principal component and the contribution of each variable to them (in percentage).

Climatic variable	Correlation		Contribution	
	PC1	PC2	PC1	PC2
Annual mean temperature	0.97	0.24	8.3%	1.3%
Annual precipitation	-0.95	0.12	8.0%	0.3%
Isothermality	0.93	-0.32	7.6%	2.3%
Max temp warmest month	-0.04	0.99	0.0%	22.6%
Mean diurnal range	0.35	0.86	1.1%	17.2%
Mean temp. coldest quarter	1.00	-0.04	8.8%	0.0%
Mean temp. driest quarter	0.54	0.81	2.6%	15.0%
Mean temp. warmest quarter	0.54	0.81	2.6%	15.1%
Mean temp. wettest quarter	1.00	0.02	8.8%	0.0%
Min temp coldest month	0.99	-0.13	8.6%	0.4%
Precipitation coldest quarter	-0.82	-0.16	5.9%	0.6%
Precipitation driest quarter	0.65	0.53	3.8%	6.4%
Precipitation seasonality	-0.99	-0.06	8.6%	0.1%
Precipitation warmest quarter	-0.87	0.31	6.7%	2.2%
Precipitation wettest month	-0.90	0.23	7.2%	1.2%
Precipitation wettest quarter	-0.72	0.68	4.6%	10.6%
Temperature annual range	-0.88	0.44	6.8%	4.5%
Temperature seasonality	0.97	0.24	8.3%	1.3%

Contribution corresponds to the squared cosine (cos^2) of the variable divided by total cosine of the component. Squared cosine represents the quality of representation of the variables on factor map. This analysis was performed only for the selected sites using the R packages *FactoMineR* and *factoextra* (Lê, *et al.*, 2008).

Table B2 describes the different genetic datasets used in this study. Accessions from previously published data and newly acquired genomic sequences were obtained to estimate genetic diversity at population level. With this, more genetic variants were obtained to estimate genotypic probabilities. Sites identified as C/T and G/A polymorphisms were identified and added to the list of SNPs to be removed from methylation datasets (see Chapter 2).

Table B2: Details about data used to estimate population genetic diversity. Genetic data was generated following a RAD-seq protocol, but type of sequencing varied between single and paired-ended.

Location	Host	N	Sequencing	Publication
BT	A	15	paired-end	newly acquired
FH	A	15	single-end	Comeault *et al*. 2015
HV	A	15	paired-end	newly acquired
HV	C	15	paired-end	newly acquired
L	A	15	single-end	Riesch *et al*. 2017
N1	A	15	single-end	Lindtke *et al.* 2017
N1	C	15	single-end	Lindtke *et al.* 2017
OUT	A	7	paired-end	newly acquired
OUT	C	6	paired-end	newly acquired
PR	C	13	single-end	Riesch *et al*. 2017
SC	C	15	paired-end	newly acquired
SCN	A	15	paired-end	newly acquired

Testing MACAU

To test for the consistency and repeatability of MACAU's outputs, two independent analyses were run in parallel and the values of the beta were compared (Eq. 4), as this coefficient describes the effect size of the predictor in the methylation levels (Lea *et al.*, 2015). To this end, the table with more stringent coverage was used (*i.e.* minimum coverage of five reads covering all 24 individuals), with 36,896 SMPs. The parameters applied were the same as described in section *3.3.7.2*, regarding the predictor, the covariates, the kinship matrix, MCMC samplings and burn-ins.

As a result, there was not a correlation between the beta outputs (r^2= 0.0, P=0.11, linear models). That is, there was not a convergence between the effect sizes of the predictor on the methylation levels, implying the output is not repeatable (Fig. B1A). The SMPs with highest beta were mainly those which presented very low methylation levels, bordering zero. This suggests MACAU requires a minimum variance at each SMP to generate consistent outputs.

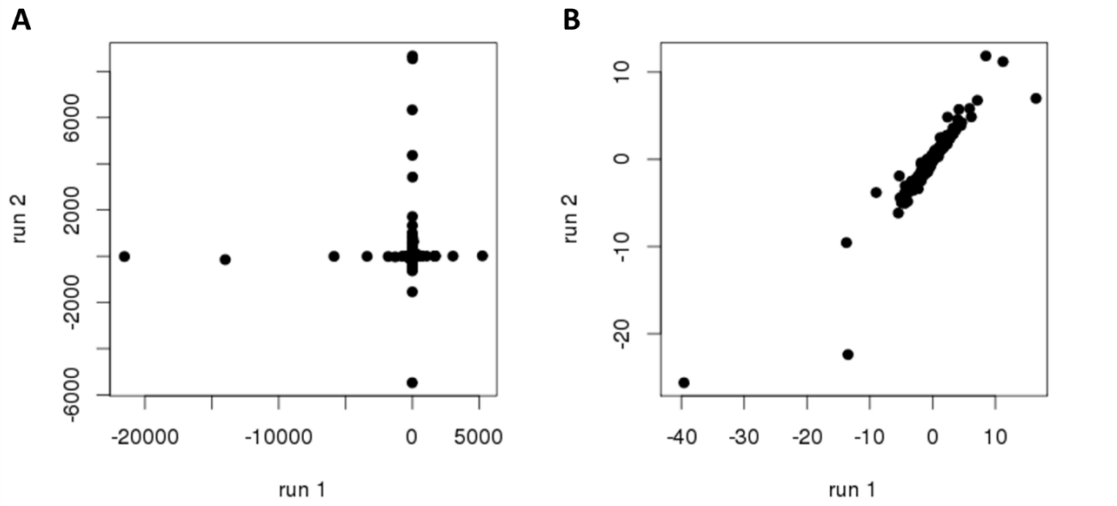


Figure B1: Correlation between output beta values (*i.e.* denotes the predictor effect size) from two independent MACAU runs using the same dataset: (A) the dataset with 38,896 SMPs not filtered for its variability; (B) the dataset after selecting at least two individuals with methylation levels > 0.25 or < 0.75. This step removed consistently hypomethylated and consistently hypermethylated positions, as it has at least two individuals that fall outside these levels, yielding 13,050 SMPs. Output from (A) show very different values for beta (r^2= 0.0, P=0.11), while the outputs converge in (B), suggesting a better consistency and reliability of the results.

In the main empirical work applying MACAU (Lea *et al.*; 2016), sites with low variance were removed from the SMPs dataset. All sites that were consistently hypomethylated (average DNA methylation level < 0.10) and consistently hypermethylated (average DNA methylation level > 0.90) were excluded, as well as sites in which the standard deviation was below 0.05. After a few attempts, I obtained SMPs with standard deviation above 0.05 (*i.e.* higher variance at each SMP) by selecting sites that had methylation levels ≥ 0.25 or ≤ 0.75 in at least two samples. With this subset dataset, the correlation between betas from two independent runs was high (r^2=0.86, *P*=2.2e-16; Fig. B1B), suggesting a good convergence between the outputs. This step retained only 35% of the CpG sites with the selected coverage (yielding 13,050 SMPs in total), but it retained more sites than when applying the threshold used in Lea *et al.* (2016; yielding 12,858 SMPs). In summary, this filtering step was applied as it improved the repeatability of the outputs. Future works could test the application of other filtering strategies to obtain consistent outputs in MACAU.

Figure B2 refers to the PCA performed on methylation variation in each individual from the population survey. This graph suggests a great variation surrounding each individual. Figure B3 represents the first two PCs associated with k-means analyses.

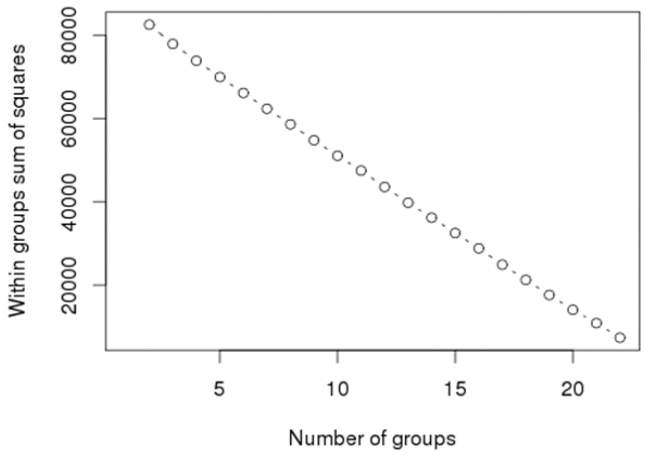


Figure B2: Sum of squares within every possible group in the first batch of data. In general, clustering methods aim to define clusters such that the total intra-cluster variation (known as total within-cluster variation or total within-cluster sum of square) is minimized. For this data, the elbow method cannot define which total within-cluster sum is the smallest, suggesting the variation in each sample is larger compared to every possible grouping.

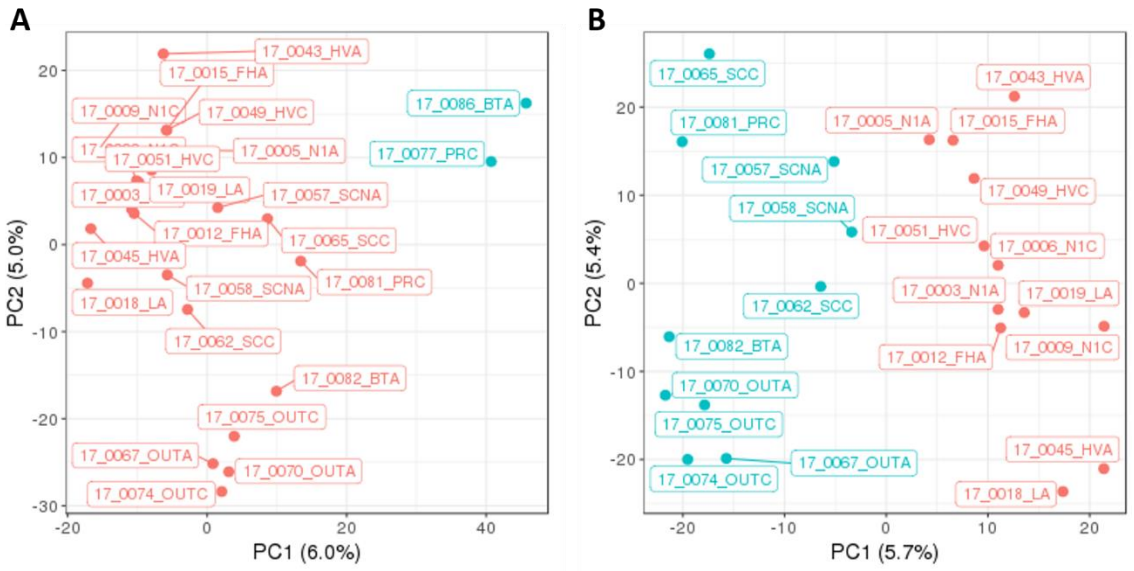
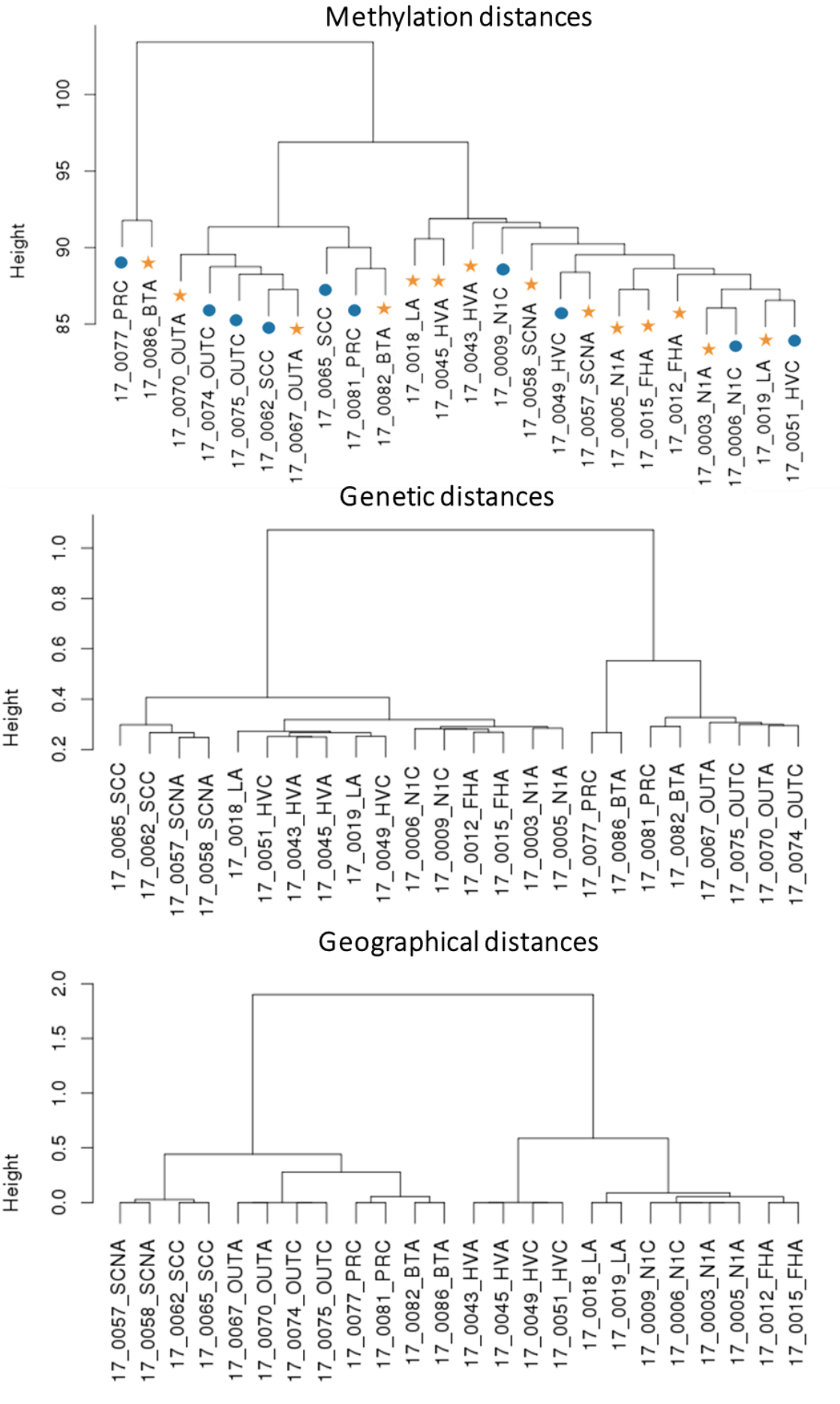


Figure B3: First two principal components from a PCA on the first batch of sequencing using the least stringent coverage (minimum 2 and maximum of 60 reads per site, yielding 296,732 sites in total). **A**. K-means analysis (k=2) groups the two outlier samples (17_0077_PRC and 17_0086_BTA in blue font; see Fig. 6). **B**. When these two samples are excluded from the analysis, k-means divides the samples according to their geographical position. Blue font and line colour correspond to the western side of the distribution and red to the eastern side.

Figure B4 represents the distances within all variables considered in this study, on: methylation, genetic, geographical space, climate, and host plant. Methylation differences were obtained based on the Euclidean distances between individuals, estimated from methylation levels in each position from the joint table (with minimum coverage of 2 and maximum of 60 reads per site). Genetic distances were estimated using the neighbour joining tree between each individual (Simonsen *et al.*, 2008); the geographical distance using QGIS tools; the climatic distances using Euclidean distances on the first PC2 loadings; and the host plants using "0" as same host and "1" as different hosts.


Methylation distances
Height
85
90
95
100
17_0077_PRC
17_0086_BTA
17_0070_OUTA
17_0074_OUTC
17_0075_OUTC
17_0062_SCC
17_0067_OUTA
17_0065_SCC
17_0081_PRC
17_0082_BTA
17_0018_LA
17_0045_HVA
17_0043_HVA
17_0009_N1C
17_0058_SCNA
17_0049_HVC
17_0057_SCNA
17_0005_N1A
17_0015_FHA
17_0012_FHA
17_0003_N1A
17_0006_N1C
17_0019_LA
17_0051_HVC
Genetic distances
Height
0.2
0.4
0.6
0.8
1.0
17_0065_SCC
17_0062_SCC
17_0057_SCNA
17_0058_SCNA
17_0018_LA
17_0051_HVC
17_0043_HVA
17_0045_HVA
17_0019_LA
17_0049_HVC
17_0006_N1C
17_0009_N1C
17_0012_FHA
17_0015_FHA
17_0003_N1A
17_0005_N1A
17_0077_PRC
17_0086_BTA
17_0081_PRC
17_0082_BTA
17_0067_OUTA
17_0075_OUTC
17_0070_OUTA
17_0074_OUTC
Geographical distances
Height
0.0
0.5
1.0
1.5
2.0
17_0057_SCNA
17_0058_SCNA
17_0062_SCC
17_0065_SCC
17_0067_OUTA
17_0070_OUTA
17_0074_OUTC
17_0075_OUTC
17_0077_PRC
17_0081_PRC
17_0082_BTA
17_0086_BTA
17_0043_HVA
17_0045_HVA
17_0049_HVC
17_0051_HVC
17_0018_LA
17_0019_LA
17_0009_N1C
17_0006_N1C
17_0003_N1A
17_0005_N1A
17_0012_FHA
17_0015_FHA

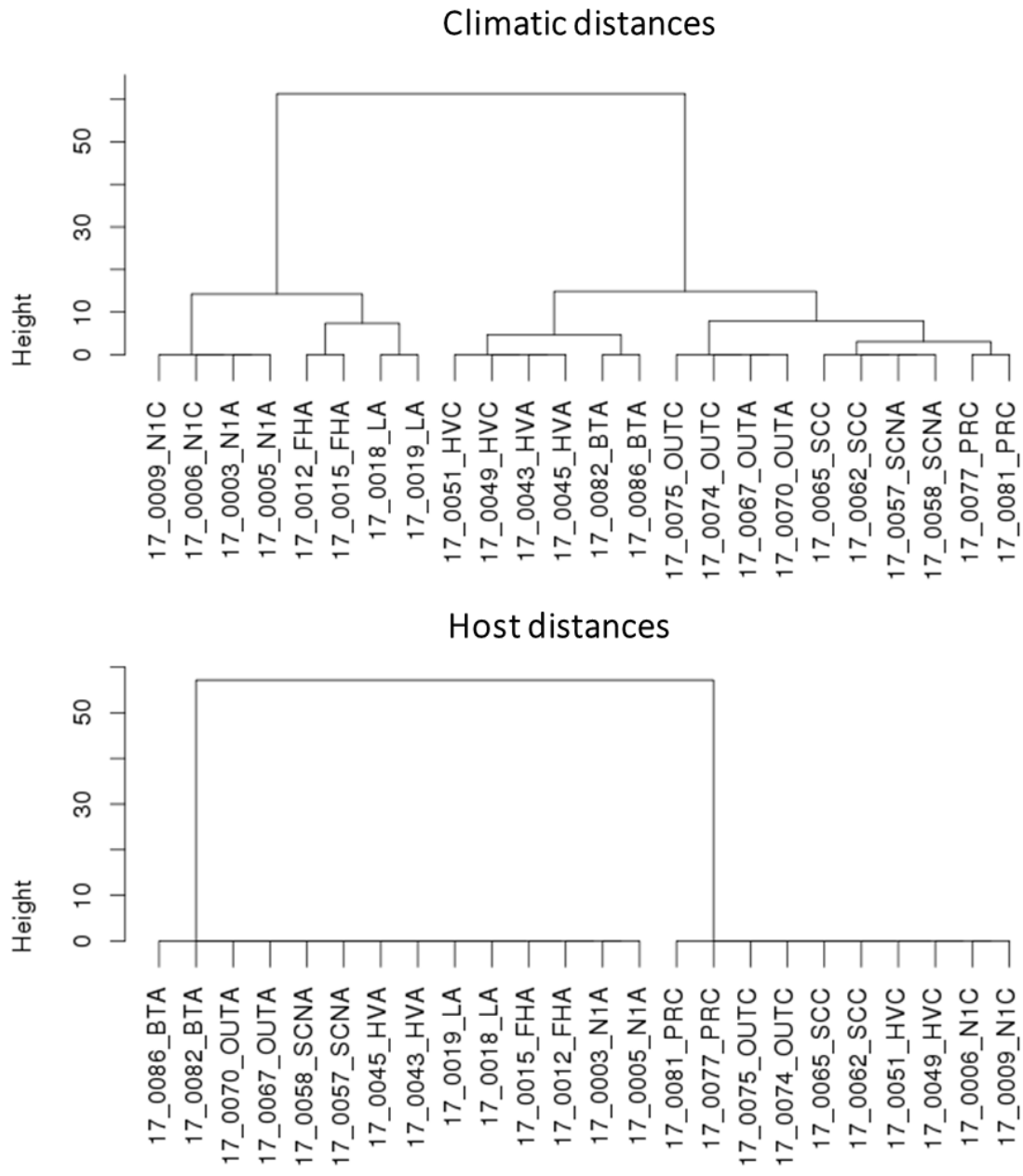


Figure B4: Dendrograms representing the distances used in this study. The algorithm used was 'Ward.D' on R function *hclust*.

Figure B5, B5 and B6 were estimated using methylation distances (Fig. B4) and genetic distances estimated using Euclidean distances on genotypic probabilities. This was performed to allow comparisons between the slopes from methylation and genetic distances, given that they were estimated using the same method (*i.e.* Euclidean distances). The dendrogram generated using this method (Fig. B5) differs slightly from the one using a neighbour joining tree (Fig. B4). The genetic differences seem to be less associated with methylation (Fig. B6) and less associated to geographical distance (Fig. B7). Comparison between the slopes suggests a higher accumulation of differences in methylation with geographical distance (Herrera *et al.*, 2016).

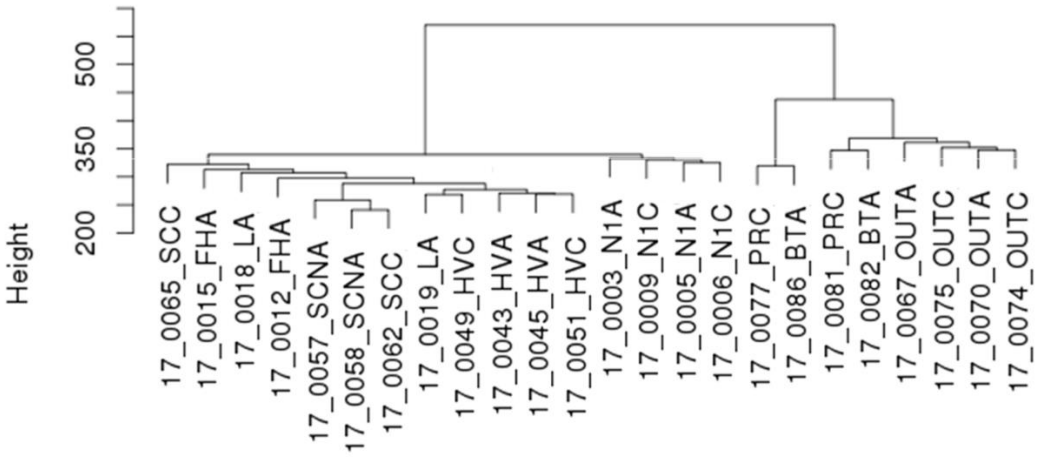


Figure B5: Genetic distances estimated using Euclidean distances on genotypic probabilities. Dendrogram estimated using algorithm 'Ward.D'.

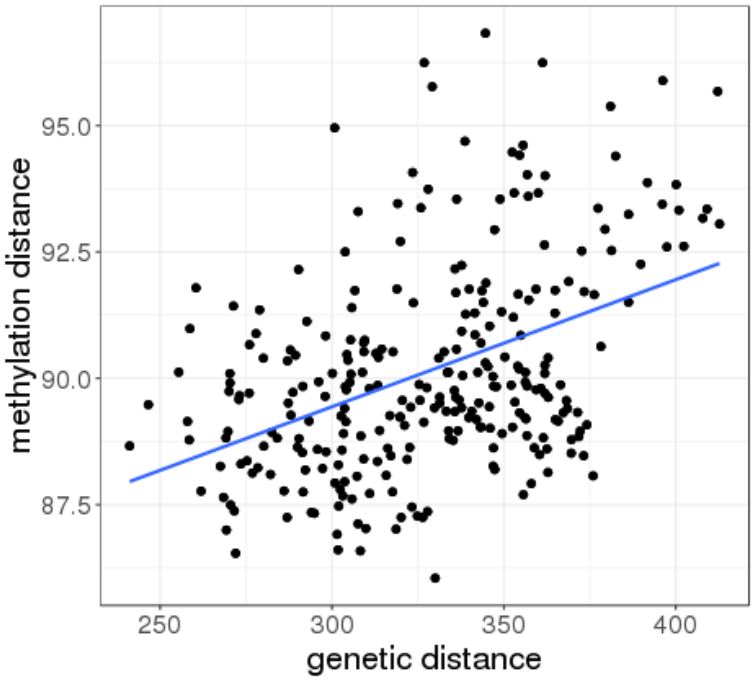


Figure B6: Relationship between methylation and genetic distances between the 24 populations used in this study. The methylation distances were estimated using Euclidean distance on the proportion of methylated cytosines over total coverage. The genetic distances were also estimated using Euclidean distances, here on the genotypic probabilities obtained from RAD-seq data. The association between these two variables is smaller using this method to estimate genetic distances (r^2=0.19, P< 2.2e-16, linear models).

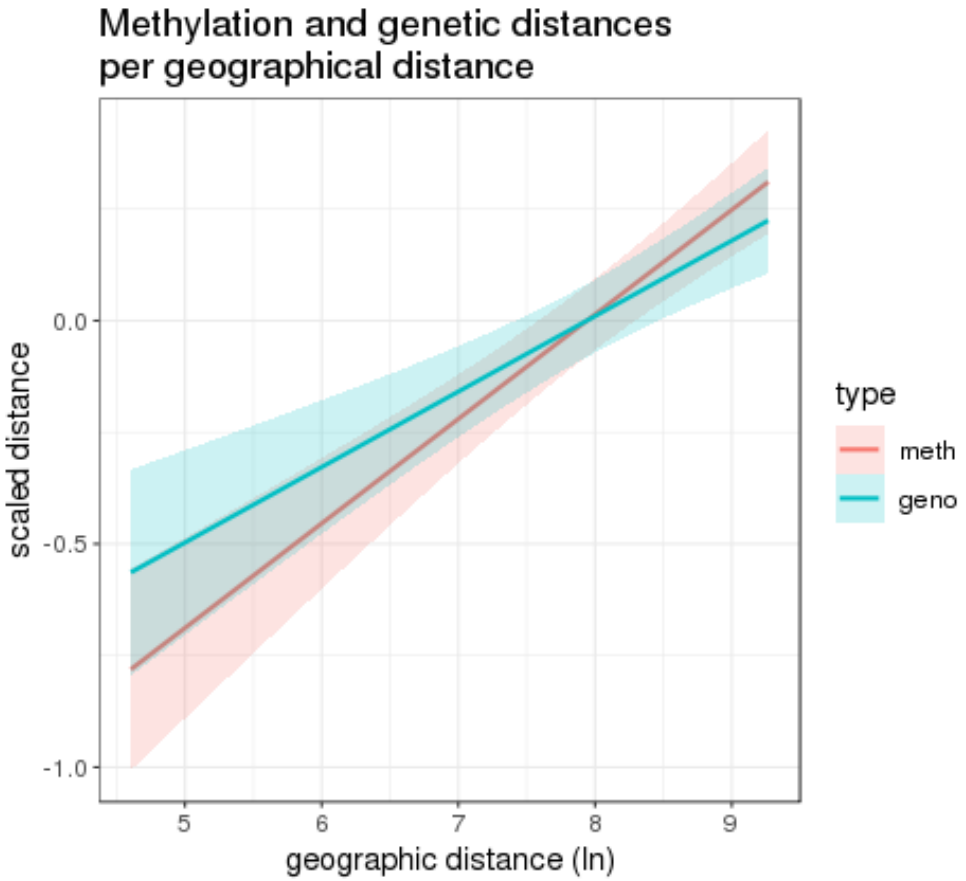


Figure B7: Comparison between the slopes on the methylation and genetic distances over geographical distance. The distances in methylation and genetic were estimated as described in (Fig B4). Here the distances were scaled for comparison. Association between methylation variation and geographical distance is r^2=0.09, P= 1.2e-12; and between genetic variation and geographical distance is r^2=0.04, P< 3.8e-07.

Chapter 4

Differential DNA methylation patterns and host plant use in *Timema cristinae* stick insects

4.1. Summary

Herbivore insects cope with a set of nutrients and chemical defences from the host plant they use. This interaction becomes challenging in a host shift, which likely involves new selective pressures. Insects' dietary requirements and performance at metabolization of species-specific chemical compounds may determine the success of colonization. As phenotypic plasticity might help organisms to tolerate the new conditions, DNA methylation could be a mechanism involved in host shift. However, evidence of this process remains elusive. Here, I combined a population survey and a rearing experiment in *Timema cristinae* stick insects to test the hypotheses that DNA methylation variation (1) is associated with different host plants and (2) changes following a host shift. Methylome scans using binomial mixed models were performed independently in each dataset to estimate candidate regions associated with plant use. One gene with functions in digestion and nutrient uptake was output in both analyses. The rearing experiment suggested methylation levels responded to host shift in a 'non-adaptive' way. Despite a few limitations to claim statistical significance of some of the results, this work highlights the possible effects host shift exert in methylation patterns in *T. cristinae*.

4.2. Introduction

Interactions between different species have always been a major interest of ecologists and evolutionary biologists (Darwin, 1859; Ehrlich and Raven, 1964; Brown and Kodric-Brown, 1979; Elton, 2001). By studying the inter-related dynamics between species, it is possible to shed light on the different factors shaping these associations and on how they can affect each other's evolution. In particular, insect-plant interactions are ubiquitous, and it is extensively acknowledged that these interactions have contributed to the great diversity of extant species in both groups (Bernays and Graham, 1988; Janz *et al.*, 2001; Agosta, 2006). Plant-feeding insects are extremely species-rich, corresponding to one quarter of all described species in the world[3] (May, 1990; Schoonhoven *et al.* 2005). A process that could explain this great diversity is the formation of new species driven by adaptation to a new plant species (Nylin and Janz, 2009). However, speciation can only occur if the insects are able to persist on the novel host, raising questions regarding the mechanisms that generate, maintain and constrain new associations between insects and plants (Ehrlich and Raven, 1964; Janz *et al.*, 2001; Agosta, 2006; Janz *et al.*, 2006).

The insect-plant interactions are described as a constant arms race between the plants evolving new defence mechanisms against herbivory, including chemical and physical barriers, and the insects counteracting these barriers with detoxification schemes, sequestration of poison and alteration of their gene expression patterns (Silva *et al.*, 2001; Mello and Silva-Filho, 2002; Schoonhoven *et al.* 2005). As such, any interaction with a specific plant involves not only facing a set of nutrients and water content, but also coping with specific chemical defences and toxins (Nylin & Janz 2009). These multiple constraints are one possible explanation as to why herbivorous insects are generally host specific

[3] If microbes, fungi and algae are excluded from the calculation.

(Bernays and Graham, 1988). The interaction becomes more challenging in the context of colonization of a new plant species, which likely involves a very different set of conditions and new selective pressures with which adults and their larvae must cope to preserve life cycle regulation (Agosta, 2006; Savković *et al.*, 2016). That is, insects' dietary requirements and performance at metabolization of species-specific chemical compounds in the new host may determine the success of its colonization (Nylin and Janz, 2009). In this context, plastic responses could help organisms to tolerate the new conditions and allow enough time for a population to become established (*i.e.* where standing genetic variation and/or new mutations can provide heritable phenotypes to respond to the novel selection pressures; Pigliucci, 2005; Ghalambor *et al.*, 2007). Insects with a broader niche width in host plant use (*i.e.* capable of exploring a variety of plant species) can be considered more plastic compared to those with a more specialized diet, as they can respond to a wider set of environments and probably handle physiological adjustments (Agosta, 2006). However, the mechanisms that allow the colonization of new host plants and niche expansion are still largely unknown.

Epigenetics, particularly changes in DNA methylation status, could be a molecular mechanism involved in the colonization of new environments (Richards *et al.*, 2012; Ardura *et al.*, 2017). DNA methylation affects processes that can ultimately influence phenotypic variation, such as gene expression, alternative splicing and transposable element silencing (Law and Jacobsen, 2010). Changes in DNA methylation patterns can change in response to biotic or abiotic triggers, and thus can sometimes lead to changes in the phenotype (Bossdorf *et al.*, 2008). It has been proposed that DNA methylation could be involved in phenotypic plasticity and contribute to an organism's ability to acclimatize to new conditions, acting as a mediator between external environment and internal molecular machinery and gene expression (Duncan *et al.*, 2014; Verhoeven *et al.*, 2016; Richards *et al.*, 2017). In other words, changes in DNA methylation that are sensitive to environmental triggers could provide a rapid source of phenotypic variation within an individual's lifetime.

Indeed, manipulation of methylation has been shown to affect patterns of plasticity, such as in caste formation in bees (Kucharski *et al.*, 2008; Foret *et al.*, 2012), in resource use of nectar-living yeasts (Herrera *et al.*, 2012), and in plant reaction norms (*i.e.* changes in the average phenotype as a function of the environment; Lande, 2009; Bossdorf *et al.*, 2010; Zhang *et al.*, 2013). This ability to finely adjust to environmental cues could be especially important when faced with novel environments (Herrera *et al.*, 2012; Richards *et al.*, 2012). In addition, studies of invasive species report an excess of DNA methylation variation relative to genetic variation, suggesting this epigenetic mechanism could facilitate establishment on a new habitat by compensating for bottlenecks and founder effects (Richards *et al.*, 2012; Liebl *et al.*, 2013; Ardura *et al.*, 2017).

Despite the potential implications of the effects of DNA methylation on phenotypic plasticity and on colonization of new habitats, empirical examples remain scarce. In addition, knowledge about DNA methylation patterns and function in insects is quite limited, with studies being considerably biased towards investigations of the evolution of eusociality (Bewick *et al.*, 2017; Glastad *et al.*, 2018). Hence, whether DNA methylation is involved in the process of host plant use by insects and in adjusting to a new plant species is unknown. To address some of these topics, I used *Timema cristinae* (Phasmatodea: Timematodea; Vickery, 1993), a species of stick insect native to South California. These herbivorous insects live in patchy environments in the Californian chaparral, feeding and mating on their host plants (Sandoval *et al.*, 1994a). The species is found in patches of two main host plant species: *Adenostoma fasciculatum* and *Ceanothus spinosus* (Sandoval, 1994b). Previous studies have shown that *T. cristinae* metapopulations probably go through many episodes of colonization and local extinction on different patches of these host plant species (Sandoval, 1994b; Farkas *et al.*, 2013).

I used natural populations to first test the hypothesis that (1) there is an association between methylation variation and environment. I predicted *T. cristinae* individuals would have different methylation patterns depending on the host plant they feed on in the wild

(*i.e.* ecotype). To this end, a population survey was conducted to investigate the natural state of methylation variation in the established populations. In addition, an experiment was performed to test the hypothesis that (2) DNA methylation is sensitive to the environment. To this end, adult insects were collected from a population where *Ceanothus* is the dominant plant, and reared on the same 'home' host, and on *Adenostoma,* to simulate a host shift (Fig. 1). I predicted the specimens reared in the home host would present similar methylation patterns found in the *Ceanothus* natural populations. In addition, I predicted the individuals reared in *Adenostoma* would respond to the environmental change by altering the methylation status resembling the patterns found in *Adenostoma* natural populations (Fig. 2). In both studies, binomial mixed models were used to test for associations between methylation variation and host plants, with genetic variation fitted as a random effect (see Chapter 3). This approach is analogous to genome scans used in population genetics, and tests for regions in the genome that are associated with a specific environmental factor (*e.g.* Hohenlohe *et al.*, 2010). There was an association between methylation variation and ecotype in natural populations, and there was an indication that some methylomic regions were sensitive to host shift. A gene with functions related to digestion and nutrient uptake was output in both analyses. The results suggest that if a true methylation change occurred in the experiment, it happened in the opposite direction from the presumed local optimum. Despite the study's limitations on not being able to correct for multiple testing in the binomial mixed models analyses and regarding the low statistical power arising from the small sample size in the transplant experiment, this work highlights the possible effects host shift exert in methylation patterns.

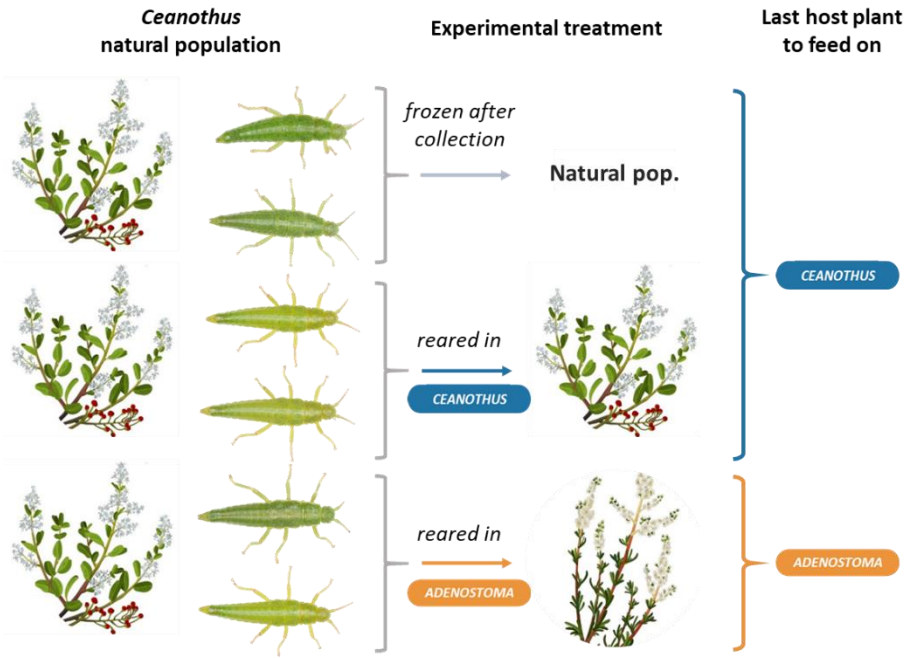


Figure 1: Rearing experiment performed in adult female *T. cristinae* of *Ceanothus* ecotype, from a population where *Ceanothus* is the dominant plant (SC locality). Two specimens were flash frozen right after collection to be used as the experimental control. Two females were reared in *Ceanothus*, and two in *Adenostoma* for 10 days, and then preserved. These six females had their methylome sequenced to test the lability of DNA methylation status after shifting to a different host plant. I tested the association between methylation levels and the host plant species the insects last fed on. Insect pictures are only examples and do not represent the individuals used in this study.

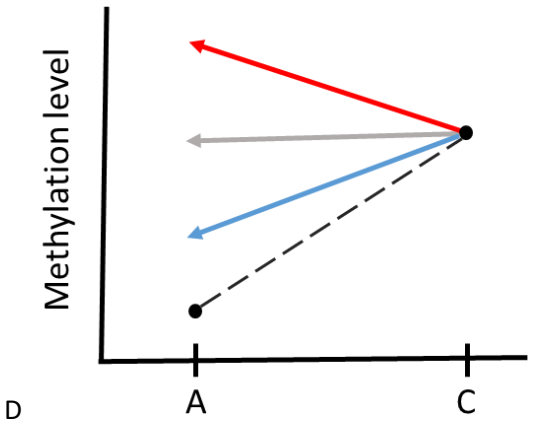


Figure 2: Hypothetical scenario of a methylation change in response to host shift. Individuals from a dominant *Ceanothus* population were reared in *Adenostoma* (A) and *Ceanothus* cuttings (C). The points represent the methylation levels found in *Adenostoma* and *Ceanothus* ecotypes in natural populations. Assuming the insects reared in *Ceanothus* would present the same levels found in the source natural population, there are three possible directions of response to the host shift treatment. If the methylation levels in each native habitat are optimal, then the change could play a beneficial role if the direction of the change goes in the same direction as the optimum state (blue arrow). On the other hand, it could be detrimental if it goes in an opposite direction (red arrow), as it results in a response that is far from the optimum (*i.e.* a 'non-adaptive' response). The grey arrow represents a lack of reaction to the host plant environment. Dashed line is represented for a reference of an 'optimal response'. Figure inspired by Ghalambor *et al.* (2007).

4.3. Material and Methods

4.3.1. Study system

T. cristinae stick insects are normally found on two morphologically distinct species of plant: *Ceanothus spinosus* (Rhamnaceae) and *Adenostoma fasciculatum* (Rosaceae). As described in Chapter 3, *T. cristinae* ecotypes are defined by the host plants they are found on, characterizing the '*Ceanothus* ecotype' and the '*Adenostoma* ecotype' (Nosil, 2007). The ecotypes differ in characteristics related to host plant use, with the most evident one being the presence or absence of a highly heritable dorsal white stripe, which characterizes the 'striped' and 'green' morphs (Fig. 3 in Chapter 1). The striped morph is more frequently found in *Adenostoma*, and the green morph in *Ceanothus* plants (Sandoval, 1994a). Previous experiments showed the striped morph is more cryptic and suffers less predation on the needle-like leaves of *Adenostoma*, whereas the green unstriped morph is more cryptic and suffers less predation on the broad leaves of *Ceanothus* plants (Sandoval, 1994a). That is, divergent selection promoted by differential predation between the two host species contributes to ecological isolation between the two ecotypes (Sandoval, 1994a; Nosil and Crespi, 2006). Previous studies showed these two ecotypes differ in a suite of other traits, including size, host plant preference, mate choice, and cuticular hydrocarbons (CHCs), molecules with roles in anti-desiccation and in insect communication (Nosil *et al.*, 2006; Nosil, 2007; Chung *et al.*, 2014; Riesch *et al.*, 2017).

4.3.2. Sampling

Specimens used in the population survey were sampled as described in Chapter 3. Previous studies have shown that the *Adenostoma* environment likely offers some physiological challenges to *T. cristinae* individuals compared to *Ceanothus*, as lifetime fecundity is significantly reduced when they are reared on this host species (Sandoval and Nosil, 2005; Nosil and Sandoval, 2008). Thus, for this work, the consequences of a host shift

from *Ceanothus* to *Adenostoma* were evaluated on DNA methylation. To this end, samples were collected at *Stagecoach* (SC; latitude 34.523, longitude -119.832), a locality where *Ceanothus* is the dominant plant. Specimens were sampled on 6th May 2016, by shaking bushes of *Ceanothus* plants and collecting insects falling onto sweep nets.

4.3.3. Rearing experiment

A rearing experiment was performed to test the effects of host plant on *T. cristinae* DNA methylation. The experiment consisted of three treatments (Fig. 1). The first one was the experimental control, where the individuals were flash frozen one day after sampling to represent the natural DNA methylation status in the population. The second treatment involved rearing the specimens on the same host plant they were collected on (*i.e. Ceanothus* plants) for ten days. Finally, for the third treatment the samples were reared in *Adenostoma* plants for ten days to simulate a host shift. Similarly-sized adult females were used for each treatment (*i.e.* six individuals in total; Table 1). Individuals were digitally photographed under the same standard conditions used for the population survey (Riesch *et al.*, 2017). The samples were flash frozen using liquid nitrogen immediately after their designated rearing time, then preserved at -80^{o}C temperature.

Table 1: Details about individuals used in the rearing experiment. All individuals were collected from the same population (SC - *Ceanothus*) on the same date (6th May 2016). Natural population treatment involved flash freezing the individuals one day after sampling, while other treatments involved rearing the specimens on the designated host plant.

Ind.	*Morph*	*BL*	*BW*	*HW*	*Treatment*
16_0116	green	2.2	0.4	0.2	natural pop.
16_0122	green	2.2	0.4	0.2	*Adenostoma*
16_0137	green	2.2	0.4	0.2	*Ceanothus*
16_0138	green	2.3	0.4	0.2	*Adenostoma*
16_0142	striped	2.2	0.4	0.2	natural pop.
16_0182	green	2.2	0.4	0.2	*Ceanothus*

Morphometric measurements (in centimetres) body length (BL), body width (BW) and head width (HW) were estimated using in ImageJ 1.4.882 (Abràmoff *et al.*, 2004), following previous works in *T. cristinae* (Comeault *et al.*, 2014).

4.3.4. DNA methylation variation

DNA methylation variation was estimated using whole-genome bisulfite sequencing (BS-seq). The methods used to obtain and process BS-seq data for the population survey were described in detail in Chapter 2. The same procedures were used to obtain DNA methylation variation for the samples used in the experiment. Specific results obtained while processing the rearing experiment data are described in the following sections.

4.3.4.1. Whole genome bisulfite sequencing (BS-seq)

The BS-treatment and sequencing were performed by Biomedicum Functional Genomics Unit (FuGU, Helsinki) in February 2018. The libraries were sequenced using the Illumina NextSeq 500 system, with High Output 2 x 150 bp runs. The samples were processed in three flow cells (*i.e.* two samples per flow cell) with four lanes each.

4.3.4.2. Read mapping and methylation calls

Filtering was done using Trimmomatic (0.36; Bolger, Lohse, & Usadel, 2014) following the same steps described in Chapter 2. The mean [95% confidence interval] number of reads across the six samples, after filtering, was 31,385,445 [23,882,505 – 38,888,385] (Table C1). Following the steps used in the population survey protocol, samples were subsampled to a maximum of 24 million reads randomly sampled from the *.fastq* files before mapping. This step was performed to minimize differences between the three flow cells.

The reads were mapped using Bismark (0.16.1; Krueger and Andrews, 2011). First, the good quality reads were mapped to the BS-transformed Lambda phage genome to isolate the data from this strain, and to obtain the estimates of BS conversion. This mapping step yielded a mean across samples of 726,076 [383,617 – 1,068,535] reads mapped uniquely to the Lambda phage (mapping efficiency of 3.0% [2.6% – 3.4%]; Table C1). The proportion of methylated cytosines in the phage was 0.4% in CpG context, which means the

mean mapping efficiency was 99.6% across the six samples. The reads that were not mapped to the phage (23,138,640 [22,775,084 – 23,502,196]) were then aligned to *T. cristinae* BS-converted reference genome, yielding 9,819,082 [9,535,164 – 10,103,000] uniquely mapped reads (mapping efficiency of 42.4% [41.9% – 42.9%]; Table C2). The '*--cytosine_report*' option from the '*bismark_methylation_extractor*' tool was used to extract the methylation call for every single cytosine in each context from the mapped files. In this work, I focused on methylation in CpG dinucleotides because this is the context most often targeted by methylation in *Timema* and in other animals (Suzuki and Bird, 2008).

I removed potential single nucleotide polymorphisms (SNPs) that could confound the estimate of methylation variation at a specific site (*i.e.* single methylation polymorphisms or SMPs; see Chapter 3). Similar to the population survey (Chapter 3), at the end of these steps tables of each individual, at each genomic position were obtained with information on: the number of reads with methylated cytosines (*i.e.* unconverted cytosines), the number of reads with non-methylated cytosines (*i.e.* number of thymines), and the proportion of reads with methylated cytosines (*i.e.* methylation levels).

The function *unite* in the R package methylKit (v1.0.0; Akalin et al., 2012) was used to generate a single joint table with variable methylation information at sites that were present across all 6 samples. That is, this function only retained information at sites that were covered in all individuals. The sites with coverage outliers above the 99.9th percentile were removed to avoid PCR bias (*i.e.* above 60 reads). The most stringent cut-off of at least five reads covering a site was used at this analysis, aiming to preserve more information at each SMP. This joint table comprised 103,873 sites. All the reported statistics were performed using R (3.3.1; R Core Team 2016).

4.3.5. Genetic variation

A restriction site associated DNA sequencing (RAD-seq) was used to generate genome-wide single nucleotide polymorphism (SNP) data, following previous studies in the

system (Comeault *et al.*, 2015, 2016). Description of the methods to obtain genetic variation are detailed in Chapter 3 (section *3.3.5*), where the genotypic data for both the population survey and experiment were processed together.

4.3.6. Clustering analyses

Hierarchical clustering analyses were performed on the six samples used in the experiment to obtain general patterns of methylation variation. The methylation levels were used to calculate the pairwise Euclidean distances between all individuals at each site (see section *3.3.5.2* in Chapter 3). The outputted distance matrix was used in the hierarchical clustering analysis, applying the 'Ward D' agglomerative criterion. Analyses were performed using *hclust* function in R (3.3.1; R Core Team 2016). In addition, I performed the same analyses using the genetic data, for comparison. For this, I used the genotypic probabilities stored in BIMBAM format, and calculated the pairwise Euclidean distances between the six individuals.

4.3.7. Methylome scan: binomial mixed models

To determine the effects of host plant on methylation levels I used the approach Mixed model Association for Count data via data AUgmentation (MACAU), developed by Lea *et al*. (2015). For each site, the model estimates the host plant effect on the methylation level, while controlling for relatedness in the samples. It does so by incorporating a matrix of pairwise kinship, which is treated as the variance-covariance matrix for the heritable component of the random effects (Lea *et al.*, 2015; Lea *et al.*, 2017). The kinship matrix can be estimated using the genetic variation in the dataset, and modelled as the 'genetic random effect'. In addition, MACAU works directly with count data, which maximises power in analyses of bisulfite sequencing datasets (see Chapter 3).

4.3.7.1. Methylome scan in the population survey

To perform the methylome scan on the population survey, I used the joint table with single methylation polymorphisms (SMPs) from the 24 samples with minimum coverage of five reads per site (see section *3.3.5* in Chapter 3). MACAU does not generate consistent outputs with SMPs that have methylation levels around 0% or 100% in all samples. These sites that are consistently hypomethylated or consistently hypermethylated, respectively, tend to return spurious outputs in MACAU, in a way that the values for the beta coefficient (*i.e.* the effect size of the predictor 'host plant' in the methylation levels, Eq. 4 in Chapter 3) were not repeatable (see Appendix B 'Testing MACAU' in Chapter 3). To assure the SMPs were sufficiently variable to run the model, I selected sites where at least two individuals had methylation levels above 25% (> 0.25) or below 75% (< 0.75). This step excluded the sites that were consistently hypomethylated or consistently hypermethylated (following Lea *et al.* 2016), retaining 13,050 sites. With this table, the effect of ecotype was modelled on methylation levels. That is, whether the host plant that the specimens were collected on could explain methylation variation in specific genomic regions. With this analysis, I could test the association between methylation variation and ecotype in the natural state. The bisulfite conversion rates were used at each sample as a covariate, as well as the climatic variables first principal component (see Chapter 3). The kinship matrix was calculated using the genotypic probabilities obtained from RAD-seq data, and was modelled as the genetic random effects.

4.3.7.2. Methylome scan in the rearing experiment

I performed the methylome scan to test for a host shift effect on methylation variation in the samples used in the rearing experiment. As a predictor, I considered the host plant a specimen last fed on. That is, the two specimens that were reared in *Adenostoma* were compared to the ones that were reared in *Ceanothus* and to the ones that were preserved right after collection in their natural habitat (Fig. 1). The joint table with SMPs in the six

samples was used, selecting sites where at least two individuals had methylation levels above 25% (to deviate from consistently hypomethylated sites) or below 75% (to deviate from the consistently hypermethylated sites; yielding 10,540 sites in total). The bisulfite conversion was used as a covariate, and added the rearing treatment as the other covariate (*i.e.* whether the insects were reared in laboratory conditions, or not). The genotypic probabilities obtained from the RAD-seq data were used to estimate a kinship matrix, which was modelled as a random genetic effect. Because it was not possible to correct the outputs for the multiple testing problem (see section *4.4.2.1*), I carried on with the investigations using the SMPs with an output *p-value* lower than 0.01, which were considered as 'putatively significant SMPs'.

4.3.7.3. Triangulation

In the search for candidate sites with effect in both population survey and in the experiment, I overlapped the putatively significant SMPs coming from the two MACAU outputs (*i.e.* SMPs with *p-value* < 0.01, but not corrected for multiple testing). This was obtained by calculating the minimum physical distance in base pairs between the putatively significant sites using R (3.3.1; R Core Team 2016) and then finding those that had an overlap of 10kbp or less. The overlap of randomized sites was not estimated in this study, and thus the expectations for the triangulation were not known to determine if the results are truly valid (*i.e.* to test whether the pattern was expected by chance or not). Thus, results from the triangulation analyses were interpreted and discussed, but they do not represent a statistically significant pattern.

4.3.8. Annotation

The genomic features at the putatively significant sites were obtained using the *T. cristinae* genomic annotation table (Villoutreix *et al. in prep*). Only genes with InterPro or GO accessions were analysed (InterPro EMBL-EBI; Gene Onthology, UniProt), and

considered upstream and downstream regions as 1kb at 5' and 3' from the genes (see Chapter 2). In addition, I also used the *T. cristinae* repeatmasker database (Villoutreix *et al. in prep*) to extract information about transposable elements (TEs), focusing only on TEs families that contained more than 400 repeats across the *T. cristinae* genome (Chapter 2).

4.3.9. Transcriptome

To evaluate expression at these different regions, I accessed an available transcriptome, previously published (1KITE project; Misof *et al.*, 2014). Briefly, the reads from these datasets were filtered using Trimmomatic (0.36; Bolger, Lohse, & Usadel, 2014) using default parameters and subsequently aligned to the *T. cristinae* reference genome 1.3c2 (Nosil *et al.*, 2018) using STAR (Dobin *et al.*, 2012). The basic two-pass mapping was used, and all reads were mapped in the first step, discarding alignments with a ratio of mismatches greater than the 5% of the mapped length (--twopassMode Basic --twopass1readsN -1 --outFilterMismatchNmax 999 –outFilterMismatch-NoverLmax 0.05). The aligned reads were then summarized into transcripts read counts and normalized read counts (FPKM) using TopHat and Cufflinks (Trapnell *et al.*, 2013).

4.4. Results

4.4.1. Clustering analyses

Results from the clustering analysis on the population survey were reported in Chapter 3. I showed DNA methylation variation does not cluster according to ecotype, but that it has a tendency to group in geographical space following the distance between populations. This is because DNA methylation variation was significantly associated with genetic variation, which is known to be more differentiated the more geographically distant two populations are from each other (Sandoval *et al.*, 1994; Jenkins *et al.*, 2010), implying there is an isolation-by-distance pattern in methylation variation. On the other hand, the dendrograms estimated from methylation differences on the rearing experiment samples

did not seem to reflect the ones generated with genetic variation (Fig. 3). Instead, individuals that were reared in *Adenostoma*, seemed to share similarities in the methylation patterns that are related to the host shift treatment. There was no correlation between DNA methylation variation and genetic variation (adjusted r^2=0.00, P= 0.93, linear models), and a low correlation between DNA methylation variation and host plant (r^2=0.14, P=0.18, linear models).

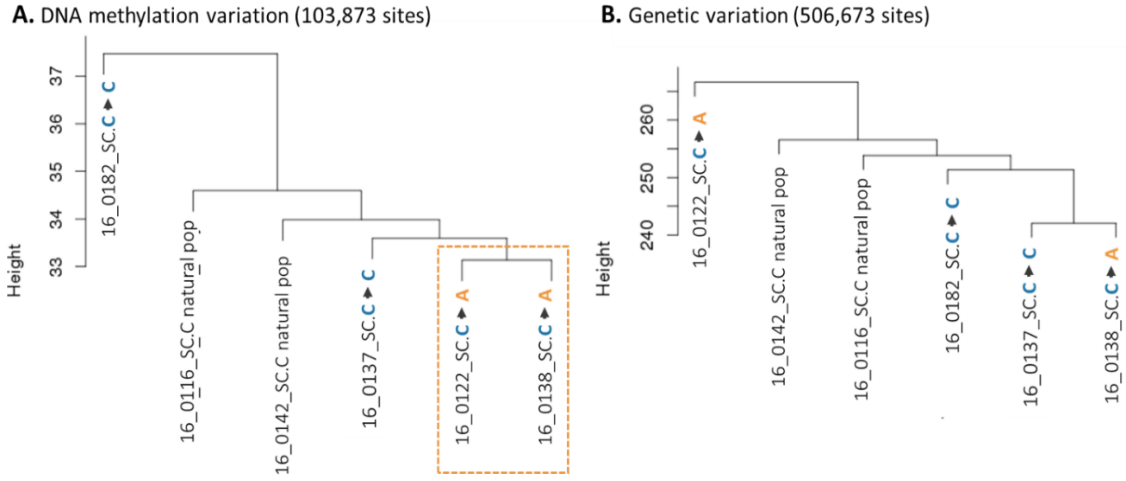


Figure 3: Hierarchical clustering using the samples from the experiment on **(A)** DNA methylation variation, estimated using methylation levels; and **(B)** on genetic variation, estimated using genotypic probabilities between the samples. Euclidean distances were used to estimate dissimilarity and 'Ward.D' algorithm of clustering. All individuals are from *Ceanothus* ecotype, and here they were represented according to the experimental treatment: 'natural pop' as the individuals representing the natural population methylation status (*i.e.* no rearing treatment), C -> C as those reared in the same host plant of origin, and C -> A as reared in the shifted host *Adenostoma*. The dendrograms differ between the two datasets, suggesting there can be some disentanglement between epigenetic and genetic variation. Despite the small sample size, these results suggest the individuals reared on the shifted host (C -> A) seem to share some similarities in methylation patterns, which is not expected from the genetic dendrogram.

4.4.2. Methylome scans

4.4.2.1. Association between methylation variation and host plants

In the population survey, differences in methylation levels between ecotypes were found across the genome. In total, 62 sites were identified as putatively significant (*i.e. p-value* < 0.01), and the results are represented in Fig. 4. The distribution of *p-values* was conservative, with relatively few sites with a *p-value* < 0.01 (Fig C1). This suggests this data

may be prone to type II error, or false negative findings. Thus, the corrections and false discovery rates that are normally used for this model (Lea *et al.*, 2015) might not be applicable to this data. Thus, different corrections should be applied to attest the validity of the results despite multiple testing. Although these corrections were not applied in this study, the further investigations were conducted using SMPs with *p-value* < 0.01 considering them as 'putatively significant SMPs'. In any case, MACAU seems to have effectively identified sites displaying differential methylation levels between the two ecotypes (Fig C2). This suggests the model is retrieving at least some of the sites where differences in methylation levels are associated with ecotype. Similar patterns were found for the rearing experiment, where 98 sites were putatively significant (Fig. 4).

4.4.2.2. Triangulation

To search for candidate sites with effect in both population survey and in the experiment, the putatively significant SMPs coming from the two MACAU outputs were overlapped. Given the very small sample of SMPs for the datasets with minimum coverage of five reads, very few regions overlapped in this trial (the closest proximity between the putatively significant sites was around 3kbp; Table 2). To expand the analyses, I used datasets with minimum coverage of three reads per site, which were modelled in MACAU with the same parameters as reported earlier in this Chapter (section *4.3.7*). The outputs from the population survey and rearing experiment analyses returned 220 and 827 putatively significant sites, respectively, out of 37,934 and 51,888 sites in total (Tables C3-C4 for gene ontology enrichment test on putatively significant sites). I found 13 putatively significant regions output in both analyses, separated by a distance up to 10kbp (Table 2). Because there were not clear expectations for the triangulation tests, this analysis does not validate these putatively significant SMPs, but identify some regions for some discussion. Among those, one gene was of particular interest, with only 20bp separating the differentially methylated cytosines from the two outputs (located in the same exon). This

gene (code IPR010629 from InterPro database, EMBL-EBI) was predicted to produce an insect allergen protein, and is widespread among Insecta clade (Randall *et al.*, 2013). This site was output as putatively significant in the methylome scan using the population survey, even with the most stringent coverage (Fig. 4), but was filtered out from the equivalent table in the rearing experiment because one sample presented maximum coverage of three reads at this site (*i.e.* 16_0116_natural pop).

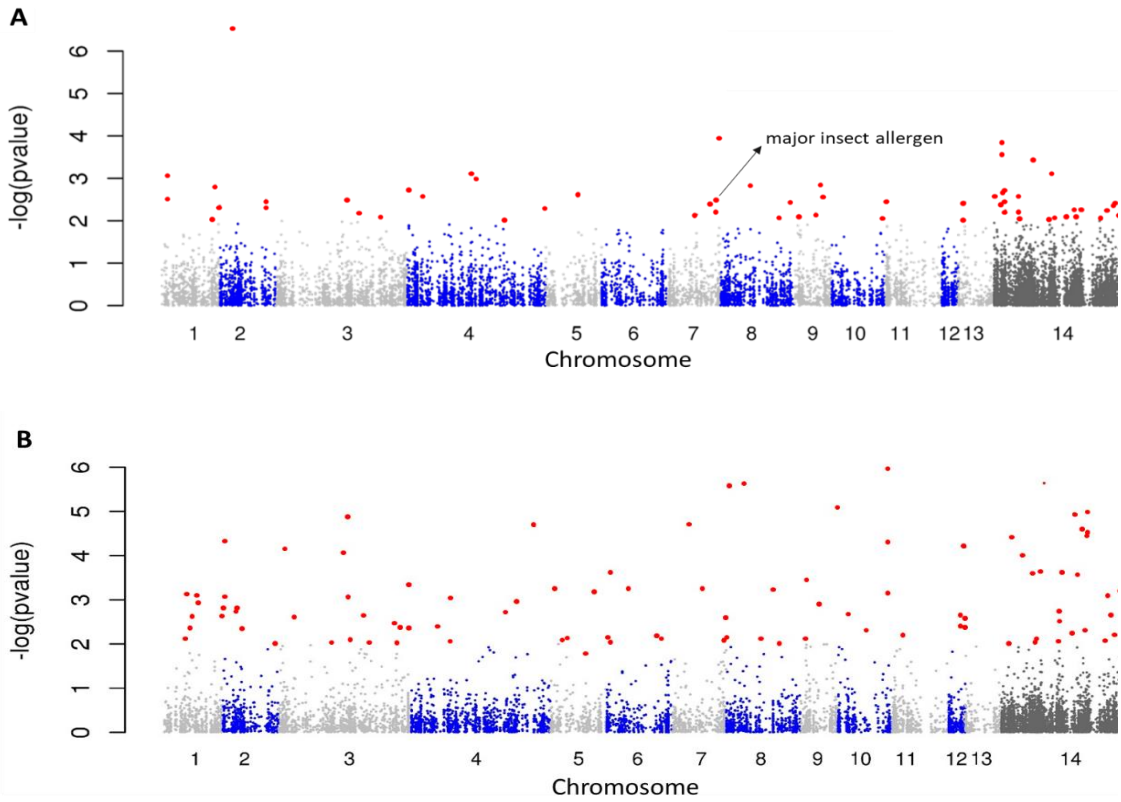


Figure 4: Manhattan plots representing the results from MACAU methylome scan. **(A)** Results from the population survey, testing for association between methylation levels and ecotype, controlling for bisulfite conversion, climatic variables, and using the kinship matrix obtained using RAD-seq as random effects. 62 sites (out of 13,501 sites) were returned as significantly associated with ecotype (*i.e.* putatively significant), scattered across the genome. **(B)** Results using the rearing experiment, testing for association between methylation levels and the plant the insects last fed (*i.e.* natural populations and *Ceanothus* treatment versus *Adenostoma* treatment), controlling for bisulfite conversion, rearing factor (*i.e.* reared individuals versus natural populations) and using the kinship matrix obtained from RAD-seq. 98 sites were identified as associated with the plants (out of 10,540 sites). The mean heritability (h^2) of SMPs was putatively significant in both analyses, with a mean of 0.67 [0.15 – 1.0] in the population survey (Chapter 3) and of 0.62 [0.08 – 1.0] in the rearing experiment.

I focused on the insect allergen gene because it was the one with the lowest physical distance between the two putatively significant sites (*i.e.* it had the best overlap), and

because I could draw a biological explanation linking this result with the interaction between *T. cristinae* and its host plants. In the population survey, the methylation levels are higher in individuals from *Ceanothus* ecotype in this site compared to the ones from *Adenostoma* (Fig. 5A). In the rearing experiment, where the insects were collected on a *Ceanothus* population, the specimens used for the control without rearing and those reared in *Ceanothus* presented similar levels of methylation, mirroring the pattern found in the population survey. However, the insects reared on the shifted host, *Adenostoma*, seemed to present a hyper-methylated status at this site (Fig. 5B).

4.4.3. Characterization of the insect allergen gene

In *T. cristinae*, this insect allergen gene was found at a single location on linkage group 7 (LG7), and is composed of three exons. In addition to this gene, other accessions on the *T. cristinae* annotation were found matching the major insect allergen function; relatives of this gene family were found as three locations *in tandem* at linkage group 11 (LG11; Table 3, Table C5). However, I found some key differences between the accessions on the different linkage groups (Table C6). While the major allergen gene in LG7 had differential methylation between the two ecotypes in natural populations, the complex at LG11 was generally non-methylated in all 24 samples.

Although I did not obtain expression data for any of the samples studied, in this study, I performed preliminary analyses by comparing the transcription levels between the two regions using previously published transcriptomes (Misof *et al.*, 2014). In contrast with the methylation results, the major allergen gene in LG7 did not have any transcripts aligned to it, or only showed very marginal transcript counts (1KITE; Misof *et al.*, 2014). While there is no significant trace of transcription in this region in the datasets, the genes in LG11 have more transcripts in both datasets (Table 3). Although these results do not directly represent transcription levels in the natural population samples or in the rearing experiment, they highlight differences in these genes' patterns of expression.

Table 2: Triangulation of putatively significant sites between population survey and rearing experiment testing for association between methylation levels and host plant MACAU.

Scaffold	*Pos1*	*Pos2*	*Dist. (bp)*	*Genom. feature*	*Gene id*	*Gene function*	*TEs family*
LG7 scaf2404	109093	109073	20	exon	g3176.t1 exon2	Insect allergen (IPR010629)	-
LGNA scaf2744	53520	53441	79	intron	g26797.t1 intron10	Chitin synthase (IPR004835)	-
LGNA scaf1756	246431	246339	92	-	-	-	Gypsy
LG11 scaf883	1280194	1280436	242	intron	g15247.t1 intron2	Dynein heavy chain (IPR026983); Molecular microtubule motor activity (GO:0003777); Microtubule-based movement (GO:0007018)	-
LG9 scaf157	3295336	3295653	317	exon	g7990.t1 exon1	-	Helitron HRC
LGNA scaf2537	864606	863490	1116	-	-	-	-
LGNA scaf3324	334110	336032	1922	intron/ upstream	g28030.t1 intron1	Molecular nucleic acid binding (GO:0003676)	Gypsy
LGNA scaf3384	154160	156607	2447	intron	g28492.t1 intron2	Exonuclease, phage-type (IPR011604); Molecular DNA binding (GO:0003677); Molecular nuclease activity (GO:0004518)	RTE / -
LG2 scaf4016*	11383	8600	2783	intron / -	g31109.t1 intron1	Peptidase aspartic (IPR008737)	BEL / -
LG7 scaf3745	28703044	28707516	4472	intron	g55987.t1 intron6	Ankyrin repeat (IPR002110); Molecular protein binding (GO:0005515)	-
LG2 scaf1827	2517557	2523326	5769	-	-	-	-
LGNA scaf3914	663596	670710	7114	exon	g29522.t1 exon1	Domain of unknown function (IPR025398)	-
LG8 scaf2963	4414172	4404463	9709	intron	g5289.t1 intron1/ g5290.t1 intron1	Ribosomal protein (IPR001141); Constituent of ribosome (GO:0003735); Ribosome (GO:0005840); Translation (GO:0006412) / BCP1 family (IPR025602)	-

This overlap was obtained by estimating the genomic physical distance in base pairs (Dist. [bp]) between the putatively significant sites in population survey (Pos1) and rearing experiment (Pos2). The table is ordered according to the minimum distance (bp) up to 10kbp. Sites in different genomic features are separated by "/". Gene function represent the genes function obtained from InterPro and GeneOntology databases. *This site overlapped between the outputs using tables with minimum reads covering each site.

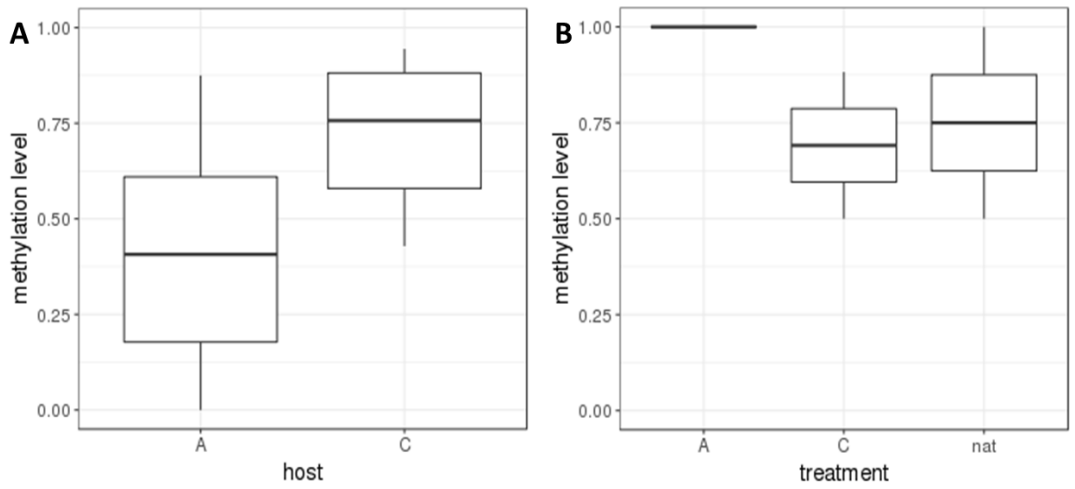


Figure 5: Methylation levels at the sites located in the exon from insect major allergen gene at linkage group 7. **(A)** Putatively significant site outputted from MACAU using the population survey (LG7 scaf2404 position 109093). Differential methylation between ecotypes in natural populations, where *Adenostoma* types [A] generally present lower methylation levels compared to *Ceanothus* [C]. **(B)** Putatively significant site outputted from MACAU using the rearing experiment (LG7 scaf2404 position 109073). Individuals representing natural *Ceanothus* populations [nat] and reared in the same plant *Ceanothus* [C] exhibited similar methylation levels to what was found in the population survey at this site. In contrast, individuals reared in *Adenostoma* [A] were hyper-methylated in this locus, not reflecting the pattern found in natural populations.

Table 3: Details of insect major allergen genes locations in *T. cristinae* genome. Methylation levels are the average methylation in exons with minimum coverage of 5 and maximum of 60 reads per site in at least one sample in the population survey. 95% CI are represented in brackets. Transcription counts reported here correspond to the values found in the 1KITE transcriptome (Misof *et al.*, 2014).

Scaffold	*Range*	*Gene*	*Number exons*	*Methylation levels (%)*	*Transcription counts*
lg7 scaf2404	107,470 - 110,769	g3176.t1	3	64.4 [7.7]	6.2
lg11 scaf2779	75,914 - 76,752	g13444.t1	3	0.0 [0.1]	313.4
	81,099 - 84,449	g13445.t1	4	0.6 [1.0]	2634.1
	99,971 - 100,629	g13446.t1	3	1.0 [1.0]	4247.2

4.5. Discussion

In this work, I combined a population survey with a rearing experiment to investigate the association between methylation variation and environment. The methylome scan output a few single methylation polymorphisms (SMPs) that putatively varied with ecotype. Although these

results do not imply DNA methylation variation is adaptive in different ecotypes, they suggest it might be involved in some aspects of the interaction between *T. cristinae* and their host plants. With the rearing experiment, I tested the property of DNA methylation to change in response to host shift. The individuals that were host shifted seemed to share some similarities in the methylation patterns, a trend that did not directly mirror genetic variation patterns. Methylome scan analyses in both datasets independently output putatively significant SMPs located in an exon from a gene belonging to the insect major allergen family, generally known for its role in nutrient uptake and for its role in detoxifying functions. In this study, the number of samples used in the experiment were very small to obtain a good statistical power in the analyses, and the outputs from the binomial mixed models could not be corrected for multiple testing. Nonetheless, the results collectively suggest DNA methylation differences could be involved in the interaction between *Timema* and their host plants.

4.5.1. Association between methylation variation and host plant

As described in Chapter 3, multivariate analyses in the population survey did not highlight ecotype as a major clustering factor among samples. Instead, genetic variation was the factor that better explained methylation variation between populations. These tests were performed using all individual methylation variation at once (*i.e.* genome-wide variation), not discarding the possibility that some specific regions could be significantly associated with ecotype. In contrast, the analyses in the rearing experiment suggested some similarities in methylation status in the individuals nurtured on the host plant of a different species from the one they were collected on (*i.e.* the shifted host). Considering this pattern seemed not reflect the one when only genetic variation was evaluated (Fig. 3), and that all the samples were likely exposed to the same environmental conditions, this result suggests a potential direct effect of host plant on the methylation levels. Although a better support for the relationships between

individuals is required for a clearer interpretation of the patterns (*i.e.* a larger sample size for more statistical power and more elaborate analysis to estimate comparable trees based on methylation and on genetics), the results highlight a possible disentanglement between methylation and genetic variation in response to host shift.

MACAU was used to perform the methylome scan in the population survey, for a fine-scaled evaluation of the association between ecotype and methylation variation at each site. The output pointed putatively significant sites spread across the genome, denoting potential candidates correlated with ecotype. These results could represent the methylation differences existing in the natural state. In other words, one could assume these differences represent the local optimum in each ecotype, as they were likely on a stable state that could have arisen from a combination of forces that would affect methylation variation (Herrera *et al.*, 2016). On the other hand, methylation differences found in the rearing experiment would most likely reflect the host plant effect only, given other conditions were standardized. In the rearing experiment, I found differentially methylated sites associated with the host plant they last fed on: either on the home *Ceanothus* plants (from the natural population or reared on the same home host plant), or on *Adenostoma*. A triangulation was performed between both analyses' outputs to look for common regions that were putatively significant for differences in methylation associated with ecotype. It is important to point that, in this study, one could not estimate whether any overlap between SMPs was expected by chance, and thus the validity of the results is yet to be confirmed. In any case, this analysis pointed some putatively significant regions located within the gene body with particularly interesting functions to raise some discussion. One of them was located in a chitin synthase gene (IPR004835, InterPro database). Chitin is known to be the main component of insects exoskeletons and inner structures, such as the tracheal cuticles and the peritrophic membrane in the guts surface (Merzendorfer, 2006). This membrane surrounds the food bolus, and it is responsible for enhancing the efficiency of food

digestion and absorption (Cardoso *et al.*, 2019). Another interesting overlap was present in an exon belonging to an insect major allergen family. I focused on this specific region in the next section not only because it was the one with smallest physical distance in the genome with only 20bp apart, but also because it was one of the few candidates that allowed a biological interpretation about the observed patterns.

4.5.2. Insect allergen gene and ecological context

The independent results in both the natural population survey and in the experiment pointing putatively significant SMPs in the insect allergen gene suggests this region would be an interesting candidate to explore the role of methylation in the interaction between *Timema* and their host plants. Insect major allergen genes are widespread in insects (Randall *et al.*, 2013). Although the roles of major allergen proteins have not yet been characterized, they are related to digestion and nutrient uptake (Gore and Schal, 2004; Suazo *et al.*, 2009), and their genes' activity is upregulated after feeding (Dostálová *et al.*, 2011; Nolan *et al.*, 2011). Thus, this gene could somehow be involved in the digestion of the plants ingested by *T. cristinae*. Interestingly, the nitrile-specifier protein (NSP) gene belongs to the major allergen gene family, which in the cabbage white butterfly and its relatives (Pieridae family) produces a detoxifying enzyme to counteract Brassicales' glucosinolate defensive compounds (Fischer *et al.*, 2008). It was described as a key innovation in the evolution of these butterflies, as it had a single evolutionary origin, and it allowed the colonization of Brassicales followed by significantly increased diversification rates (Wheat *et al.*, 2007). This is the best described gene in the family; albeit it is thought to be more derived compared to other insect allergen genes in the literature (Fischer *et al.*, 2008).

I predicted the methylation status on individuals reared in *Ceanothus* and *Adenostoma* would mirror the patterns found in the natural populations. This was based on the notion that

plastic responses that happen in the direction of the optimal phenotype in the new habitat can be advantageous, as they provide broader tolerance and hence higher fitness in the new environmental conditions (Ghalambor *et al.*, 2007; Nicotra *et al.*, 2015). Following this, it was expected the individuals shifted to *Adenostoma* would respond to the new conditions by changing the methylation status in the same direction observed in the *Adenostoma* natural populations. However, I found that the differential methylation patterns in the candidate on the insect allergen gene were distinct between the two results.

Although the levels of methylation at the insect allergen gene in the individuals from the home host reflected the patterns found in the wild, the reaction to the *Adenostoma* rearing environment did not follow the expected direction of methylation change. That is, instead of a reduction in the methylation levels, I observed a hypermethylated status (Fig. 5). This result can be interpreted as a response in the opposite direction from the optimum (*i.e.* as a non-adaptive response). When populations experience new environments, it is likely they respond in a non-adaptive way, because they bring traits and responses evolved elsewhere (*i.e.* selection has not had an opportunity to act on the basis for plasticity; Agosta, 2006; Ghalambor *et al.*, 2015). Although it drives the trait further away from the presumably adaptive peak, this non-adaptive response could also influence evolutionary trajectory in the novel environment. It is predicted to increase the strength of directional selection as it reduces relative fitness in the new environment (Conover *et al.*, 2009). In their study with guppies, Ghalambor *et al.* (2015) showed the main changes in gene expression reacting to a novel environmental cue in the laboratory was contrary to the pattern found in the transplanted populations in the wild. In other words, the controlled conditions in the laboratory allowed individuals to express the non-adaptive plastic response, while those with the same reaction in the natural transplant were possibly removed by strong directional selection – leaving just the ones with a more constrained plastic response. Some authors argue non-adaptive plasticity can increase the

phenotypic variance around the mean due to expression of cryptic genetic variation (Conover *et al.*; Pfennig *et al.*, 2010). In other words, whereas beneficial plastic responses can buffer genetic variants and facilitate their accumulation, a change in the environment followed by a non-adaptive plastic response might release this variation and allow the populations to respond rapidly to the new selective pressures.

Given the above, it is possible the insect allergen methylation levels could respond in a 'non-adaptive' way to a host shift, which would compromise the activity of this gene and consequently performance in the new habitat. In this study, it was not possible to evaluate the consequences of this differential methylation on the insect allergen gene. That is, differential gene expression was not evaluated in this study, and neither did I estimate phenotypic or fitness differences associated with host shift to extrapolate the conclusions about non-adaptive response. In addition, it is important to note the plastic response in methylation reported here could be under genetic control, although more analyses are required to test it. In their work, Dubin *et al.* (2015) showed a significant association between DNA methylation variation and temperature in *Arabidopsis thaliana.* Using several genome-wide analyses, they found a marked association between this variation and the genetic background, suggesting the epigenetic response was under genetic control. A similar approach could be applied to *T. cristinae*, although a bigger sample size would be required, preferably from fewer populations to reduce the genetic structure.

Results from this study suggest methylation patterns could change following an environmental change, a phenomenon that may be linked to rapid and reversible phenotypic plasticity (Huang *et al.*, 2017; Metzger and Schulte, 2018). Previous studies revealed *T. cristinae* performs fairly well on both host plants (*e.g.* Nosil, 2007). In fact, *Timema* species seems to have retained plasticity in host use, being able to process and metabolize a series of host plants (Larose *et al.*, 2019). This suggests *Timema* stick insects possess a diverse molecular machinery

to cope with host shifts, likely involving plastic responses. To expand our knowledge of how methylation is associated with host plant use and how it responds to a host shift, future work should increase the sample size and the methylome coverage to be able to explore a greater number of regions present in all individuals, and to obtain substantial statistical power to validate what in this study was considered 'putatively significant SMPs'. In other words, to validate these results (obtained using 'putatively significant SMPs') future studies should correct for multiple testing to consider the SMPs that are truly significantly associated with ecotype differences in downstream analyses. In addition, not only could careful assessment of phenotypic traits such as gain of body mass and reproductive performance be evaluated, but also the question of whether the changes in methylation following host shift are linked to differences in expression could be investigated.

In this work, I focused on adjustments to new environmental conditions in adult individuals, a process that can be called 'acclimation'. That is, the experiments did not evaluate changes during development, which tend to result in stable phenotypic changes that remain throughout an organism's lifetime (Metzger and Schulte, 2018). Future experiments could perform a host shift in early stages of development to assess how methylation could vary with this process. Such an experiment, coupled with measures of weight gain, survival, and fecundity would provide a good opportunity to find links between methylation variation, phenotype and fitness, and help generate a clearer picture of the relevance of methylation variation to ecological processes in *T. cristinae*.

4.5.3. Evolution of insect major allergen genes

The major allergen gene is normally found in many copies in insect genomes. They can either be found *in tandem* as part of a major gene complex or isolated as a single major domain.

In *Timema cristinae*, this gene was found as a single locus in LG7 and as three loci *in tandem* at LG11 (Table 3). While the major allergen gene in LG7 has differential methylation in different ecotypes in natural populations, the complex at LG11 was generally non-methylated in all samples. In contrast with the methylation results, the major allergen gene in LG7 showed no signal, or very marginal read counts in the available transcriptome datasets (Comeault *et al.*, 2012; Misof *et al.*, 2014). There was not a substantial trace of transcription in this region in either of the datasets, whereas the genes in LG11 seemed to be highly transcribed in both of them (Table 3). Although these results do not directly represent the gene expression patterns in the natural population samples or in the rearing experiment, they shed light on the evolution of these genes and setup a direction for future investigations.

Rodin and Riggs (2003) proposed DNA methylation as a mechanism that facilitates the conversion of duplicate genes into pseudogenes or towards functional diversification. Their models show DNA methylation could alter the roles of duplicated genes and prevent them from becoming pseudogenes by partitioning the functions performed by the ancestral gene between the duplicates, a process called sub-functionalization. That is, the divergence in gene-body methylation could play a functional role in influencing evolution and divergence of paralogs. Indeed, the frequency of functional young gene duplicates is higher in organisms with high levels of DNA methylation (*e.g.* mammals and plants), compared to those with little or no methylation (*e.g.* insects and nematodes; Lynch, 2000). Studies have shown divergence in methylation levels and patterns in paralog genes correlates with their sequence and expression divergences – in the great majority of duplicate pairs, one pair is always hyper-methylated compared to the other one (Keller and Yi, 2014; Wang *et al.*, 2014). Kucharski *et al.* (2016) studied honeybees' odorant binding proteins (*obp*), molecules that facilitate the delivering of external particles to the odorant receptors. The genes for these proteins are found in many copies across the genome, and they showed DNA methylation could have been the mechanism

driving functional diversification of one of these genes (*opb11*) from its non-methylated tandem partner *opb10* by affecting alternative splicing.

Hence, it is possible methylation is related to the evolution of the major allergen genes in *T. cristinae*. For a better understanding of the context of this divergence, future studies could investigate the genetic differences between the copies of the insect allergen genes and their evolution. The divergence between the different copies of the genes could be estimated, and classical diversification tools could be used to determine the rate of evolution of these paralogs. Signatures of selection (*e.g.* dN/dS ratios or Macdonald Kreitman tests) could be used in the future within the *Timema* radiation to look for non-neutral evolution of these gene families, and help elucidate their role in host plant use.

4.6. Conclusion

My findings suggest there can be an association between host plant use in *T. cristinae* and DNA methylation variation in some regions. In addition, they point putatively significant SMPs in a gene that could be relevant to processing food resources (*i.e.* the major insect allergen gene), and hypothesised that some of this variation could be subject to rapid change following an environmental shift. That is, it is possible DNA methylation could be associated to the relationship between these stick insects and their native host plant, even though the modification following the environmental change happened in the opposite direction from the expected pattern in nature, suggesting host shift could trigger 'non-adaptive' responses. Multiple copies of the major allergen gene, with different methylation patterns, were found in *T. cristinae* genome, which could imply a history of sub-functionalization and function diversification. This work highlights the importance of using data acquired from natural populations, combined with a controlled-conditions experiment in the understanding of DNA

methylation's ecological relevance. Although more studies are required to support these conclusions, this work gave a first step towards understanding the importance of DNA methylation and insect-plant interactions and host shift.

Appendix C: Supplementary Tables and Figures – Chapter 4

Table C1: Details about bisulfite sequencing data from the 6 individuals used in the rearing experiment when mapped to Lambda phage.

Ind.	*Treat.*	*Flow cell*	*Reads parsed**	*Reads mapped*	*Mapping efficiency*	*Number mCpG*	*mCpG*	*mCHG*	*mCHH*
16_0116	nat.	4	42963472	1398462	5.8%	77139	0.3%	0.4%	0.4%
16_0122	A	5	43202411	433013	1.8%	25298	0.4%	0.5%	0.4%
16_0137	C	6	24809873	359913	1.5%	19499	0.3%	0.4%	0.4%
16_0138	A	4	26399313	628230	2.6%	36563	0.4%	0.4%	0.4%
16_0142	nat.	5	23188300	456682	2.0%	29624	0.4%	0.5%	0.5%
16_0182	C	6	27749302	1080158	4.5%	58438	0.3%	0.4%	0.4%

Treat= Experimental treatment, where 'nat' corresponds to experimental control (flash frozen right after sampling), 'A' corresponds to rearing treatment in *Adenostoma*, and 'C' to rearing treatment in *Ceanothus*. Flow cell= Information about the flow cell that each individual was sequenced. Details about flow cells 1-3 are described in Chapter 3. Reads parsed= Represents the total number of reads retained after filtering. This step was followed by a random subsampling of 24 million reads in each sample before mapping. Reads mapped= Number of reads uniquely mapped to the unmethylated Lambda phage BS-converted genome, starting from the 24 million reads. Mapping efficiency= Percentage of reads uniquely mapped to Lambda phage. Number mCpG= number of methylated cytosines in CpG context. mCpG, mCHG, and mCHH correspond to the proportion of methylated cytosines in each one of those contexts.

Table C2: Details about BS-seq data from the 6 individuals used in the rearing experiment when mapped to *T. cristinae* BS-converted reference genome 1.3c2. Mapping was performed using the reads that were not mapped to the phage.

Ind.	*Treat.*	*Flow cell*	*Non-map. reads*	*Reads mapped*	*Mapping efficiency*	*Number mCpG*	*mCpG**	*mCHG*	*mCHH*
16_0116	nat.	4	22601538	9588269	42.4%	4659828	9.2%	0.4%	0.4%
16_0122	A	5	23566987	9601128	40.7%	5093721	10.2%	0.5%	0.5%
16_0137	C	6	23640087	10352246	43.8%	6581194	11.3%	0.5%	0.5%
16_0138	A	4	23371770	9935347	42.5%	5115671	9.3%	0.5%	0.5%
16_0142	nat.	5	22731618	9416473	41.4%	5208409	10.4%	0.5%	0.5%
16_0182	C	6	22919842	10021163	43.7%	6146275	11.1%	0.5%	0.5%

Non-map. reads= Number of reads that were not uniquely mapped to the Lambda phage. Reads mapped= Number of reads uniquely mapped to *T. cristinae* BS-converted reference genome starting from the reads that were not mapped to the Lambda phage. Mapping efficiency= Percentage of reads uniquely mapped to *T. cristinae.* Number mCpG= number of methylated cytosines in CpG context. mCpG, mCHG, and mCHH correspond to the proportion of methylated cytosines in each one of those contexts.

** Proportion of methylated CpG is lower than the mean from population survey. The causes of that difference are not understood, but are likely a result of differences in the manipulation of the samples.*

Gene Ontology (GO) of putatively *significant sites outputted at MACAU*

I generated a list of GO terms that were over-represented in genes with differently methylated sites varying with host plant, output from analyses using MACAU (Lea *et al.*, 2015). The number of putatively significant sites in each gene were counted, and their enrichment in certain GO terms were estimated using the R package *TopGO* (v 2.34.0). This analysis was performed independently for the population survey (Table C3) and for the rearing experiment outputs (Table C4), using the minimum coverage of three reads per site. I performed the analysis using genes that presented at least one differently methylated site versus the genes without any hits. Fisher's Exact Tests were used to calculate the significance of the enrichment, coupled with a weight algorithm. This algorithm uses a hierarchical approach to compute the *p-value* of a GO term, conditioning the process based on the neighbouring terms (*i.e.* it accounts for GO topology). Hence, the tests are not independent from each other, which means the multiple testing theory does not apply. Given this, the authors of the R package attest the *p-values* are internally corrected and do not need further correction for multiple testing.

Table C3: List of Gene Ontology (GO) terms significantly enriched in sites associated with host plant in the population survey (minimum of three reads per site). I tested for genes containing at least one putatively significant site (*i.e.* with *p-value* < 0.01; n=101 genes) compared to genes without putatively significant sites (n=4,540). Fisher's exact test was used with the weight algorithm, which accounts for GO topology using R package *TopGO.* Few terms were significant, and here I represented those with *p-value* < 0.05.

GO term	*Category*	*Description*	*Annot.*	*Signif.*	*Fold enrich*	*p-value*
GO:0005198	MF	Structural molecule activity	45	5	4.9	0.0032
GO:0004601	MF	Peroxidase activity	6	2	14.3	0.0072
GO:0006979	BP	Response to oxidative stress	7	2	14.3	0.0075
GO:0020037	MF	Heme binding	19	3	7	0.0084
GO:0000214	CC	tRNA-intron endonuclease complex	1	1	50	0.015
GO:0030130	CC	Clathrin coat of trans-Golgi network vesicule	1	1	50	0.015
GO:0030132	CC	Clathrin coat of coated pit	1	1	50	0.015
GO:0000379	BP	tRNA-type intron splice site recognition and cleavage	1	1	50	0.0198
GO:0006857	BP	Oligopeptide transport	1	1	50	0.0198
GO:0043461	BP	Proton-transporting ATP synthase complex assembly	1	1	50	0.0198
GO:0000213	MF	tRNA-intron endonuclease activity	1	1	50	0.0228
GO:0004385	MF	Guanylate kinase activity	1	1	50	0.0228
GO:0004719	MF	Protein-L-isoaspartate (D-aspartate) O-methyltransferase activity	1	1	50	0.0228
GO:0008889	MF	Glycerophosphodiester phosphodiesterase activity	1	1	50	0.0228
GO:0031072	MF	Heat shock protein binding	1	1	50	0.0228
GO:0009408	BP	Response to heat	2	1	25	0.0392
GO:0051090	BP	Regulation of DNA-binding transcription factor activity	2	1	25	0.0392
GO:0005858	CC	Axonemal dynein complex	3	1	20	0.045

'BP' represents biological process, 'CC' category represents cellular component, and 'MF' represents molecular function. Annot=number of genes with the annotated GO term; Signif=how many genes with the GO term contained putatively significant sites.

Table C4: List of Gene Ontology (GO) terms significantly enriched in sites associated with host plant in the rearing experiment (minimum of three reads per site). I tested only for putatively significant sites (*i.e. p-value* < 0.01; n=427) compared to the remaining sites (n=6,953), within the gene body. Fisher's exact test was used with the weight algorithm, which accounts for GO topology using R package *TopGO.* Few terms were significant, and here I represented those with *p-value* < 0.05.

GO term	*Category*	*Description*	*Annot.*	*Signif.*	*Fold enrich.*	*p-value*
GO:0007018	BP	Microtubule-based movement	75	13	2.6	0.002
GO:0003777	MF	Microtubule motor activity	71	11	2.5	0.004
GO:0005509	MF	Calcium ion binding	107	14	2.1	0.006
GO:0005615	CC	Extracellular space	14	4	4.5	0.010
GO:0046578	BP	Regulation of Ras protein signal transduction	29	6	3.1	0.011
GO:0017048	MF	Rho GTPase binding	26	5	3.1	0.020
GO:0004725	MF	Protein tyrosine phosphatase activity	27	5	3.0	0.023
GO:0016311	BP	Dephosphorylation	37	6	2.4	0.034
GO:0008138	MF	Protein tyrosine / serine / threonine phosphatase activity	5	2	6.5	0.034
GO:0008173	MF	RNA methyl-transferase activity	12	3	4.0	0.034
GO:0001539	BP	Cilium or flagellum-dependent cell motility	5	2	6.1	0.039
GO:0005977	BP	Glycogen metabolic process	6	2	5.0	0.041
GO:0004930	MF	G protein-coupled receptor activity	22	4	2.9	0.044

'BP' represents biological process, 'CC' category represents cellular component, and 'MF' represents molecular function. Annot=number of genes with the annotated GO term; Signif=how many genes with the GO term contained putatively significant sites.

BLASTp on insect major allergen genes

To characterize the major allergen genes, I searched for genes annotated in *T. cristinae* (version 1.3c2; Villoutreix *et al. in prep*) that presented the same protein function (InterPro: IPR010629). I then retrieved the putative protein sequences and performed a BLASTp alignment (Altschul et al., 1997) at National Center for Biotechnology Information website (NCBI). I aligned all the protein sequences to its non-redundant protein sequence database (Table C5), and against each other (Table C6). The best hits were selected, and all the accompanying information reported here.

Table C5: BLASTp output between the putative proteins with major insect allergen function IPR010629 InterPro database) in *T. cristinae* and NCBI's non-redundant protein sequences database.

Gene	*Description*	*Max score*	*Total score*	*Query cover*	*E value*	*Ident*
g3176.t1 *LG7 scaf2404*	allergen [*Periplaneta americana*]	114	114	73%	4.0e-27	36.1%
	putative *Per a* 1 allergen variant [*Periplaneta americana*]	112	112	73%	1.0e-26	35.5%
	major allergen *Per a* 1.0101 [*Periplaneta americana*]	111	111	73%	4.0e-26	35.5%
	major allergen *Bla g* 1.0101 [*Blattella germanica*]	107	107	73%	4.0e-25	36.0%
	major allergen Cr-PII [*Periplaneta americana*]	111	218	74%	7.0e-25	35.6%
g13444.t1 *LG11 scaf2779*	major allergen *Bla g* 1.0101 [*Blattella germanica*]	167	167	94%	6.0e-49	45.7%
	protein G12 isoform X2 [*Aedes aegypti*]	166	166	95%	3.0e-48	40.5%
	major allergen *Bla g* 1.0101 [*Blattella germanica*]	169	338	95%	2.0e-47	45.7%
	protein G12 [*Aedes aegypti*]	162	162	95%	7.0e-47	40.9%
	AAEL010435-PA [*Aedes aegypti*]	163	163	98%	8.0e-47	40.1%
g13445.t1 *LG11 scaf2779*	major allergen *Bla g* 1.0101 [*Blattella germanica*]	159	159	76%	3.0e-45	42.5%
	protein G12 isoform X2 [*Aedes aegypti*]	158	158	76%	2.0e-44	39.5%
	major allergen *Bla g* 1.0101 [*Blattella germanica*]	162	323	81%	6.0e-44	41.5%
	AAEL010435-PA [*Aedes aegypti*]	157	157	85%	7.0e-44	36.2%
	protein G12 isoform X2 [*Aedes aegypti*]	154	154	76%	3.0e-43	39.1%
g13446.t1 *LG11 scaf2779*	major allergen *Bla g* 1.0101 [*Blattella germanica*]	164	164	94%	6.0e-48	45.7%
	Bla g 1.02 variant allergen [*Blattella germanica*]	172	425	97%	1.0e-47	47.6%
	major allergen *Bla g* 1.02 [*Blattella germanica*]	172	424	97%	1.0e-47	47.6%
	major allergen *Bla g* 1.0101 [*Blattella germanica*]	167	333	95%	2.0e-46	45.7%
	G12 [*Culex quinquefasciatus*]	160	160	97%	4.0e-46	41.1%

The majority of genes were related to major allergen genes in cockroaches (*Periplaneta americana,Blattella germanica*) and in mosquitoes, also called G12 (*Aedes aegypti*, *Culex quinquefasciatus*).

Table C6: BLASTp output comparing the putative proteins with major insect allergen function (IPR010629 InterPro database). Gene g3176.t1 is located on linkage group 7, while g13444.t1, g13445.t1 and g13446.t1 are located on linkage group 11 in tandem. Based on the lowest E-values and highest identity score, the genes on LG11 are more similar to each other than when compared to the gene on LG7.

Gene1	Gene2	Max score	Total score	Query cover	E value	Ident
g3176.t1	g13444.t1	79	79	73%	8.0e-23	30.27%
g3176.t1	g13445.t1	81.3	81.3	73%	4.0e-23	28.80%
g3176.t1	g13446.t1	81.6	81.6	73%	7.0e-24	32.80%
g13444.t1	g13445.t1	321	342	100%	2.0e-117	77.32%
g13444.t1	g13446.t1	318	318	100%	3.0e-117	79.90%
g13445.t1	g13446.t1	295	310	80%	2.0e-107	69.07%

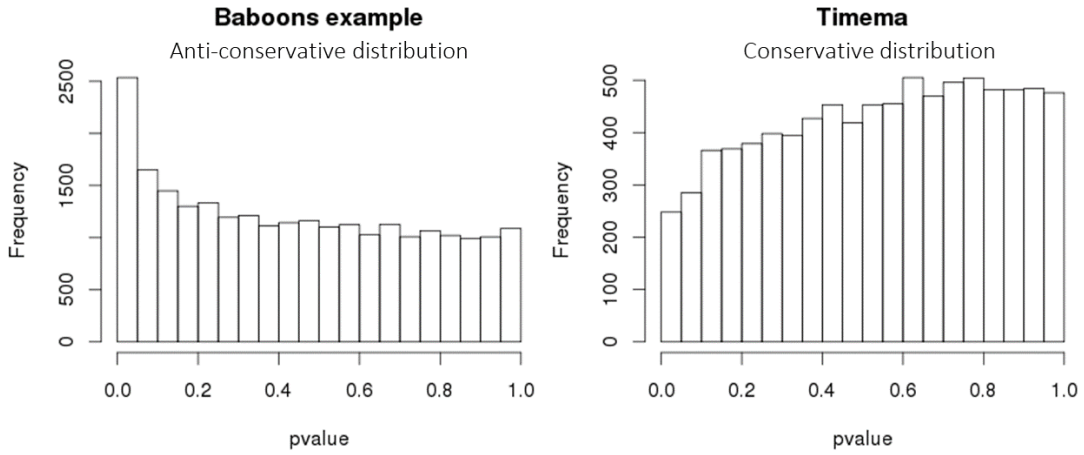


Figure C1: Distribution of *p-values* outputted from MACAU (Lea *et al.*, 2015) using the example dataset on baboons provided by the developers (left), and on *T. cristinae*'s population survey (right). The distribution of *p-values* in baboons is anti-conservative, showing a very high frequency of values below 0.01. This inflated frequency of significant sites can emerge from multiple testing, which can result in false positives. The authors suggest applying false discovery rates corrections to control for this effect. However, *T. cristinae* is conservative, with a depletion of *p-value* < 0.01, below what is expected at a null hypothesis (*i.e.* same frequency of *p-values*). Thus, it is possible the model is returning less significant sites that are expected by chance, characterizing a type II error, or false negatives. Thus, the corrections suggested by the developers do not apply. Because could not find an adequate test to correct for the multiple testing, the investigations in this study were carried with the sites showing p-value < 0.01, considering them 'putatively significant SMPs'.

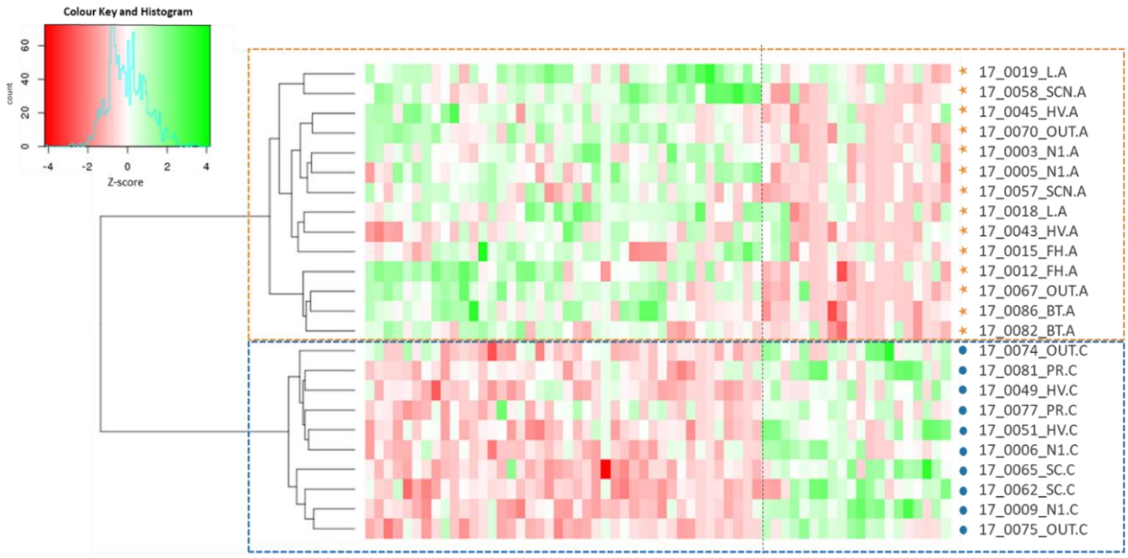


Figure C2. Heatmap using the methylation levels at the 62 sites that were putatively significantly associated with ecotype on the population survey (*p-value* < 0.01). This graph suggests the model managed to detect the regions that were differently methylated between the individuals collected on each host plant. Indeed, the samples cluster by ecotype when only these sites are considered.

Chapter 5

Conclusions and future directions

5.1. General discussion

This thesis encompasses three studies focused on understanding the patterns and the functionality of DNA methylation in *Timema cristinae* stick insects. The aim of this dissertation was to investigate natural DNA methylation variation in realistic scenarios with genetically and environmentally heterogeneous populations. With this, it was possible to explore the intertwined factors acting in natural methylation variation, which are generally missed in laboratory experiments (Bossdorf *et al.*, 2008; Herrera and Bazaga, 2011; Ledón-Rettig, 2013; Herrera *et al.*, 2014). This was done by investigating different scales of DNA methylation variability. First on a species-level context, then on genome-wide differences within-species, and lastly focusing on the relationship between ecotype and methylation differences at single base resolution. Overall, this dissertation has characterized natural DNA methylation variation in *T. cristinae* and its covariance with genetic and environmental factors. In the next section, I discuss the main findings of this work, with a focus on how the results provide insights into our understanding of the importance of DNA methylation in ecological processes. The sections are divided according to the outstanding questions highlighted at Chapter 1 (see Fig. 2 in Chapter 1). To my knowledge, this is the first study to investigate DNA methylation through an ecological perspective in insects.

5.1.1. How does DNA methylation vary between insect species?

DNA methylation is sparsely studied in insects and is mainly focused on clades with labour division and social systems, mostly Hymenoptera (Holometabola). In Chapter 2 of this

dissertation, I depicted the methylation profile in *T. cristinae* stick insects. This study highlighted the similarities and especially the differences between this system and the normally studied ones. In *T. cristinae*, such as in other insects with DNA methylation, it was found sparsely distributed across the genome, and enriched in coding regions (Xiang *et al.*, 2010; Zemach *et al.*, 2010; Bonasio *et al.*, 2012; Libbrecht *et al.*, 2016). Among those, methylation seems to preferentially target genes with housekeeping functions, while non-methylated genes are related to more dynamic and changeable processes, such as signalling and transduction pathways (Glastad *et al.*, 2016). This common pattern in insects suggests DNA methylation is important to maintain the integrity of these fundamental cellular processes. Another pattern in insects is the general methylation impoverishment on transposable elements (TEs). This is known to be one of the main targets of DNA methylation in plants and vertebrates, as it silences this activity (Zhang *et al.*, 2006; Suzuki and Bird, 2008). This was normally the case in *T. cristinae*, although some TEs families were always enriched in methylation. One explanation for this is that these families could be very active, and thus the organisms' genome integrity would benefit from the methylation repression on them. As such, future studies interested on the role of DNA methylation in repressing TEs in insects could estimate the relationship between the hypermethylated TEs and their expression in *T. cristinae.*

At the same time, I found patterns that were contrasting between groups of insects. The methylation patterns in *T. cristinae* generally resembled those found in insect species that have incomplete metamorphosis (*i.e* "Hemimetabola" group; Krauss *et al.*, 2009; Wang *et al.*, 2014; Glastad *et al.*, 2016). Differently to the widely studied Holometabola insects (Xiang *et al.*, 2010; Bonasio *et al.*, 2012; Wang *et al.*, 2013; Cunningham *et al.*, 2015; Libbrecht *et al.*, 2016), *T. cristinae* presented elevated levels of DNA methylation, enriched in both exons and introns. As "Hemimetabola" are underrepresented in the literature about DNA methylation, these insights from *T. cristinae* contribute to reinforce the patterns contrasting these two groups. Given that

DNA methylation is more widely distributed among genes, it is possible DNA methylation plays a more important role in "Hemimetabola". In effect, to some extent these patterns are more similar to those found in vertebrates than those in Holometabola insects (Glastad *et al.*, 2016). Thus, the *T. cristinae* methylation profile, as well as that of other "Hemimetabola", could be relatively underived during the evolution of insects.

Together, the results presented here underline the relevance of studying species from different taxonomic groups in order to raise patterns, generalities and differences. Future studies should continue the effort in analysing representatives of different clades to shed light on the mechanisms by which DNA methylation variation arises in insects. More importantly, valuable insights will emerge from investigating the molecular functions of DNA methylation in different insect species, which remain largely unknown. An experiment to directly test the importance of DNA methylation in *T. cristinae* could be via administering RNA interference (RNA_i), a conserved cellular mechanism used to inactivate the expression of specific genes. Targeting the maintenance DNA methyltransferase (DNMT1) in juveniles would knock down its activity and result in demethylation of targeted tissues during development. This could be administered using the method developed by Li-Byarlay *et al.* (2013), which can treat large numbers of insects in a non-invasive way via aerosol application. With the appropriate controls, it would be possible to investigate the consequences of demethylation on gene expression, on alternative splicing, and, ultimately, on the phenotype.

This method has been applied in other insects and provided compelling results. The study silencing the DNMT3 enzyme in honeybees using RNA_i was the first to shed light on the relationship between DNA methylation and royal jelly effect on caste differentiation (Kucharski *et al.*, 2008). More recently, Bewick *et al.* (2019) demonstrated the knockdown of DMNT1 and reduction of DNA methylation compromised the reproduction and the egg viability in milkweed bugs (*Oncopeltus fasciatus*) without any effect on gene expression. That is, DNA methylation in

milkweed bugs could be more important for genome structure, integrity or other cellular processes than it is for the regulation of somatic gene expression. These examples emphasize the relevance of using manipulative analyses and experimental tests to understand the DNA methylation functions in insects. Examining the effects of DNA methylation on expression, on suppression of transposable elements and on regulatory pathways will help us understand its importance to holo and to hemimetabolous insects.

5.1.2. What is the extent and structure of DNA methylation variation in natural populations?

By studying the extent and spatial structure of natural DNA methylation, one can capture the effects of forces that are possibly acting cumulatively over many generations (Herrera *et al.*, 2016). With this in mind, in Chapter 3, I studied natural populations of *T. cristinae* varying in geographical distance and environmental factors such as climatic differences and ecotype. The extensive work on population genetics in *T. cristinae* offered a unique set up to approach these questions at epigenetic level. The results suggested considerable genome-wide DNA methylation variation between individuals, both within and between populations. My study pointed this variation is structured in geographical space, and that differences between individuals tend to increase the more distantly separated they are in physical space – a pattern that is parallel to what is found in genetic variation. In fact, genetic variation had a better power to explain DNA methylation variation than physical geographical distance, suggesting there might be some genetic control over it. There was not a significant association between genome-wide methylation variation and environment: neither with climatic variables or with ecotype. However, this result did not discard the possibility of an environmental effect in only a few localized genomic regions. The next sections provide an in-depth discussion about the effects of those factors driving methylation variation.

The results cited together indicate DNA methylation variation could accumulate in geographical space following a pattern of isolation by distance (Herrera *et al.*, 2016; Richards *et al.*, 2017), mirroring the genetic variation patterns in *T. cristinae*. To strengthen these findings, future studies should expand the number of individuals sampled from each locality to obtain a better estimate of the methylation variation within-populations and to reinforce the spatial structure patterns. Moreover, including more localities in the population survey varying in environmental factors and separated by different distances will allow us to better disentangle the effects of gene flow and environment in genome-wide patterns of methylation and in its genetic background.

5.1.3. To what extent does DNA methylation variation depend on genetic variation?

As mentioned above, I found a strong correlation between DNA methylation and genetic variation in *T. cristinae*. This was also manifested in the results from the binomial mixed models (Lea *et al.*, 2015), which suggested a significant mean heritability of methylation patterns mirroring the estimates of pairwise kinship using genetic variation (see below the discussion about heritability). This association suggests a substantial amount of DNA methylation variation in *T. cristinae* can be determined by its genetic basis. In other words, that genetic variation could control some of the methylation patterns, either by factors that *cis* or *trans*-regulate methylation state. This high correlation has been extensively reported in plants and in vertebrates (Liebl *et al.*, 2013; Schmitz *et al.*, 2013; Dubin *et al.*, 2015; Taudt *et al.*, 2016; Carja *et al.*, 2017). This finding in *T. cristinae* indicates this trend could be prevalent across eukaryotes. At the same time, it is important to note that the correlation between DNA methylation variation and genetic variation does not imply causation, and that there might be other factors covarying with both.

These conclusions were based on results from analyses considering general patterns in methylation and genetic variation (genome-wide patterns). Thus, a next step for further understanding of this association must be obtained at a finer scale, identifying the direct links between genetic variants and variation in methylation levels. Identifying these specific associations will allow us to investigate the interdependence between both parts. For example, in mammals, proximate links between methylation and genetic variation (*i.e. cis*-acting variants) are normally related to differential transcription factor binding in enhancers and/or promoters in mammals (Taudt *et al.*, 2016). In plants, these associations result from transposable elements insertions or repeats, which tend to be largely methylated (Pecinka *et al.*, 2013). Thus, such investigation in *T. cristinae* could elucidate how these associations are likely to occur in insects. This way, a study similar to the one performed at Dubin *et al.* (2015) could be performed. With a greater sample size and reduced population structure (*e.g.* using samples from the same population), genome-wide analyses (GWA) could be carried out using each SMP as a 'phenotype' to be correlated with single nucleotide polymorphisms (SNPs). As DNA methylation is present in only 2% of cytosine residues in *T. cristinae* (Chapter 2), whole-genome sequencing and an increased depth in the bisulfite sequencing would increase the probability of finding direct links between epigenetic and genetic variation (Lea *et al.*, 2017). With these analyses, one could identify some of the genetic bases of DNA methylation patterns in *T. cristinae* and the regulatory mechanisms underlying them.

5.1.4. What is the heritability of methylation variation?

Linked to the significant association between genetic and DNA methylation variation, binomial mixed models pointed to significant heritability of methylation patterns in *T. cristinae* (Lea *et al.*, 2015). This method uses a Bayesian approach to model SMPs according to a predictor

of interest, estimating heritability of methylation patterns based on the genetic random effects (estimated using pairwise genetic kinship between the samples). My results suggest the relatedness in methylation variation mirrors the genetic kinship, supporting the hypothesis about genetic control over methylation levels in *T. cristinae*. Assuming methylation variation is reset during meiosis, the results presented here imply heritability of methylation levels arise from re-establishment of the patterns in the next generation because they tag specific genetic variants (a one-to-one correspondence). It is possible that some SMPs exhibit high heritability of the patterns because of pure epigenetic inheritance, which would imply an incomplete erasure of epigenetic marks between generations One could hypothesise that if SMPs were inherited and changed in frequency in different populations, they would be behaving like SNPs – resulting in patterns that would be comparable to the pairwise genetic kinship matrix.

To date, very little is understood about epigenetic inheritance in insects. For example, it is not known whether DNA methylation reprogramming occurs in insects' gametogenesis. In Chapter 2, I discuss the finding *T. cristinae* and other insect species do not present the *de novo* DNA methyltransferase (DNMT3), which adds methyl groups to non-methylated DNA sites, only the maintenance DNA methyltransferase (DMNT1), which acts on hemi-methylated sites during DNA duplication. As such, one of the hypotheses that could be raised is that DNA methylation patterns are not erased during gametogenesis, implying maintenance of methylation status across generations. In effect, Wang *et al.* (2016) showed stable inheritance of methylation status between generations in *Nasonia* wasps, suggesting they could have been transmitted across generations via pure epigenetic inheritance. To test these hypotheses, future investigations should be carried out to elucidate the molecular mechanisms of epigenetic inheritance in insects (Fig. 1).

5.1.5. To what extent is DNA methylation variation sensitive to environmental changes?

Although there was not a significant correlation between genome-wide DNA methylation variation and environmental factors in natural populations (Chapter 3), my findings in Chapter 4 suggested this association could exist at some specific SMPs. This was obtained using binomial mixed models (Lea *et al.*, 2015), scanning for candidate SMPs that were correlated with ecotype in natural populations of *T. cristinae*. Coupled with this study, a rearing experiment simulating host shift suggested there was an association between SMPs and host plant use (*i.e.* environmental effects of host shift). The low coverage and the small sample sizes, especially in the experiment, offered some limitations to the obtention of results that had a confident statistical power using binomial mixed models. Nonetheless, my study led to some results that pointed to future perspectives to the investigation of host shift and changes in DNA methylation levels. For example, some of the putatively significant SMPs were found to be in the same genomic region in the population survey and in the experiment. In particular, one of those common regions was an exon from a major insect allergen gene, with functions associated with digestion and nutrient uptake (Randall *et al.*, 2013). I found methylation levels in this gene could be changing with host shift treatment towards the opposite direction from expected, based on the natural populations' status. With this, I suggested the environmental change could have triggered a 'non-adaptive' reaction in the insect allergen methylation levels (see Fig. 2 in Chapter 4; Ghalambor *et al.*, 2007; Nicotra *et al.*, 2015).

Overall, my research suggests there could be a significant association between DNA methylation and ecotype. In addition, it highlights the potential ability methylation has to react to an environmental change, and that it may not necessarily happen towards the 'optimum state'. In these studies, I focused on adjustments to host shift in adult individuals, a process that can be called 'acclimation' – a rapid and reversible response to environmental change. That is, my experiments did not evaluate methylation changes during development, which tend to be

stable and remain throughout an organism's lifetime (Metzger and Schulte, 2018). As such, future experiments could perform a host shift in early stages of development to assess its effect on methylation levels. In addition, these environmental effects on methylation signals should be measured across generations to determine its inheritance (Fig. 1).

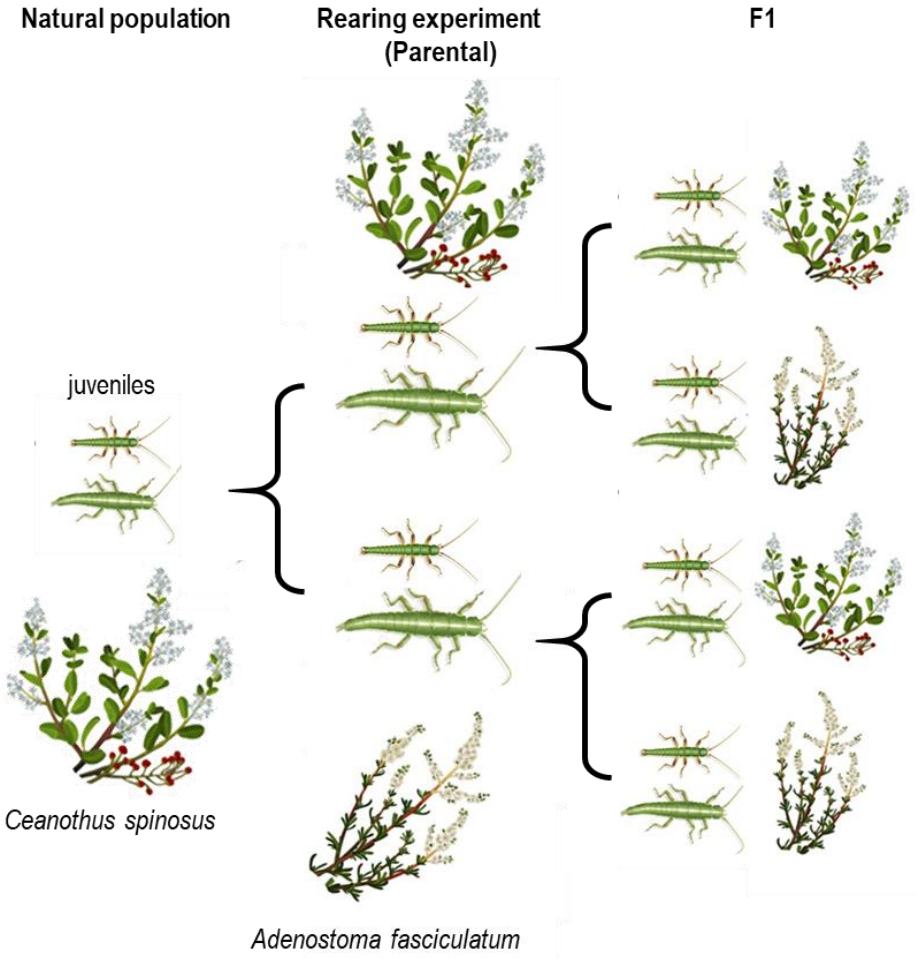


Figure 1: Design of a potential rearing and crossing experiment to be carried out in *T. cristinae*. With this design, one can evaluate (1) SMPs heritability; (2) direct effects of host shift on methylation levels in early stages of development; (3) and heritability of methylation patterns arising from the host shift effects. The procedure can be performed by collecting juveniles from natural populations where *Ceanothus* is dominant, then rearing half in the same host and the other half in *Adenostoma*. F1 eggs can be collected from each couple and split between the different host plants species. Initial *Ceanothus* natural population is an example, and can be switched to *Adenostoma.* Pictures from Rosa Marin Ribas.

5.1.6. What are the consequences of DNA methylation variation in T. cristinae?

Throughout this dissertation, I have demonstrated some evidences to the following aspects about DNA methylation variation in natural populations of *T. cristinae*:

- The patterns at species-level are characteristic of "Hemimetabola" insects;
- It is strongly correlated to genetic variation;
- There is a moderate mean heritability of methylation patterns, likely associated to their genetic background;
- It is structured in geography, likely due to its genetic background and heritability;
- It could be associated with ecotype in specific regions;
- Host shift could potentially lead to changes in some of those ecotype-associated regions.

Even though there was not any measurement of phenotype or fitness to ultimately argue about the importance of DNA methylation variation, the findings outlined above set ground to some discussion about its consequences in *T. cristinae.*

For example, one can debate about the contribution of DNA methylation effects to phenotype. My results in *T. cristinae* suggest DNA methylation variation in *T. cristinae* may not explain phenotype independently from genetic variation, given there is a strong correlation between methylation and genetic variation. That is, they imply DNA methylation variation would arise as a manifestation of the genotype, such as a phenotype. This assumption does not disregard the importance of DNA methylation variation to biological processes and to relevant changes on the phenotype. However, if it is strictly under genetic control, it would represent a proximate cause of those changes, and not the ultimate cause (Richards, 2006). However, as discussed in the previous section, the association between genetic and methylation variation was estimated only at genome-wide levels, and thus more refined analyses will be able to

elucidate to which extent SMPs depend on genetic variation. In fact, there could be regions of partial genetic control over methylation variation. In those cases, the associated genetic background would facilitate the epigenetic change (Richards, 2006): for instance, when a genetic mutation, a TE insertion, or any other structural variant occurs and creates a facilitating change to be modulated by the methylation state (Waterland and Jirtle, 2003; Pecinka *et al.*, 2013). In summary, identifying the associations between SNPs and SMPs will allow us to determine the epigenetic effects on gene expression and on the phenotype independently from genetic variation.

Moreover, by mediating phenotypic plasticity, DNA methylation might facilitate the colonization of new environments by adjusting to the new conditions (Bossdorf *et al.*, 2008). In Chapter 4, I performed an experimental host shift and measured the differences in DNA methylation variation. Although differences in the phenotype were not measured, the results suggest host shift could have resulted in DNA methylation change in *T. cristinae*. This putative change happened in the opposite direction from the 'optimal state' (*i.e.* defined by the methylation status found in the nature), which could imply the response might not necessarily happen in a fine adjustment to the new environment, but rather in a desynchronized way (Ghalambor *et al.*, 2007). It has been shown that *T. cristinae* has lower fitness when fed with *Adenostoma* (Sandoval and Nosil, 2005; Nosil and Sandoval, 2008). This way, it is possible some of the maladapted physiological reactions involved in this host shift were triggered by the changes in methylation. To test this hypothesis, future studies should not only investigate the methylation changes associated with host shift, but also assess differences in gene expression and on phenotype. In my study, I chose a natural *T. cristinae* host plant species to simulate the host shift. An interesting investigation could involve a similar experiment, but involving a plant species on which *T. cristinae* performs poorly (Larose *et al.*, 2019), and estimate the

methylation and genetic variation underlying individual's performance (*e.g.* weight gain, survival, fecundity).

5.2. Future perspectives in ecological studies in DNA methylation

Throughout this thesis, I highlighted the importance of investigating DNA methylation through an ecological lens in order to gain a holistic understanding of the functions and the evolutionary consequences of this epigenetic mechanism. With this in mind, some key issues were outlined here for future research.

DNA methylation is a complex feature, intertwined with many factors at scales that vary from molecular to ecological processes. It has been studied mainly by molecular biologists and, recently, by ecologists and evolutionary ecologists. To date, at one hand, genomic sequencing tools and molecular experiments have been applied to model organisms in the laboratory (*e.g.* van der Graaf *et al.*, 2015; Onuchic *et al.*, 2018; but see Schmitz *et al.*, 2013; Schmid *et al.*, 2018). At the other hand, DNA methylation variation has been explored at broad range in non-model organisms in their natural environment (*e.g.* Herrera and Bazaga, 2011; Richards *et al.*, 2012; Liebl *et al.*, 2013; Platt *et al.*, 2015). However, very few studies have been performed combining investigations in the common environment and in the wild (see Dubin *et al.*, 2015; Nicotra *et al.*, 2015; Groot *et al.*, 2018). By coupling both approaches (*i.e.* applying high-resolution tools in both natural population surveys and in controlled condition experiments), one can estimate natural DNA methylation variation in the wild and test the effects of some of its potential drivers in controlled conditions (*e.g.* environmental effects).

While theoretical models show DNA methylation variation has the potential to influence evolutionary dynamics (Pál and Miklós, 1999; Jablonka and Raz, 2009), its adaptive potential has rarely been empirically tested (Bossdorf *et al.*, 2008; Verhoeven *et al.*, 2016; Richards *et al.*,

2017). For example, the role DNA methylation can play in phenotypic plasticity suggests it may facilitate the response to environmental change and allow the organisms' persistence (Bossdorf *et al.*, 2010; Herrera *et al.*, 2012; Nicotra *et al.*, 2015; Foust *et al.*, 2016). Whether this phenomenon can have a meaningful ecological effect being temporally transient or if it can persist across multiple generations via a constant environmental stimulus (*i.e.* an ecological memory) or via epigenetic inheritance is still very debatable (Hagmann *et al.*, 2015). To address these questions, future studies should estimate the key role DNA methylation plays during the process of adjusting to environmental change. To this end, the DNA methylation change in response to the environmental variation must be decomposed determine how much of it arises (1) from genetic control, (2) from direct effects caused by the environment, and (3) from natural selection on methylation variation. To determine the causal links between genetic variation and DNA methylation, one could use quantitative trait locus studies in model organisms (*i.e.* both genetic, QTL, and methylation meQTL mapping) or genome-wide association studies, modelling SMPs according to genetic variation (GWAS; Taudt *et al.*, 2016). This will provide not only an estimate of the extent to which DNA methylation changes rely on the genetic background, but also to identify regions that are independent from it and their importance – which could be further analysed using targeted bisulfite sequencing or expression of candidate loci (Richards *et al.*, 2017). After this, the environmental effects on methylation can be estimated, as well as the sole contribution of methylation variation to phenotype. These SMPs must be followed for consecutive generations to estimate natural selection effects, and whether the environmental signal is associated with changes in frequency. Using laboratory-controlled conditions, one could discriminate between environmental effects and epigenetic inheritance during this process (*i.e.* if variation remains in the absence of environmental triggers). If these changes lead to phenotypic variation and influence the individuals' performance, one can finally determine whether the response to environmental change was

adaptive or non-adaptive. Ultimately, long-term evaluation of these processes in species from different ecological contexts will likely provide material to estimate the importance of DNA methylation to evolutionary processes.

Studies carried out in *A. thaliana* illustrate experiments following this framework. For example, Dubin *et al.* (2015) have identified SMPs associated with local adaptation to different temperatures; but using genome-wide association studies they have discovered these effects were largely due to genetic variants (many of which showing evidence of local adaptation themselves). Recently, Schmid *et al.* (2018) reported results from an experiment simulating rapidly changing environments in recombinant inbred lines. A reduction in methylation variation and phenotypic variation has associated with changes in SMPs frequency in consecutive generations – without significant genetic changes compared to the ancestors. This study suggests DNA methylation was subject to selection and contributed to rapid adaptive responses, although the authors could not identify the extent to which epigenetics played a role in adaptation. Studies such as these ones performed in a variety of organisms will significantly contribute to our understanding of the importance of DNA methylation to ecological and evolutionary processes.

Finally, studies should not always attempt to find an adaptive plot to DNA methylation. There is a possibility DNA methylation variation does not play a role in adaptation, but still be able to influence evolution in a neutral manner (Guerrero-Bosagna, 2017). For instance, it is known the methyl group makes a cytosine more prone to mutating into a thymine, creating a mutation bias. These transitions occur in much higher frequencies than other point mutations, and are assumed to be responsible for the CpG deficiency observed in vertebrate genomes (which are highly methylated; Simmen, 2008). In fact, biased mutations on methylated CpG sites appear to be even higher in the germ line (Kong *et al.*, 2012). Summing this fact with the enhanced epi-mutability status and lability responding to environmental triggers collectively

make DNA methylation a factor that could promote genetic variability, fuelling evolutionary processes. Empirical evidence for these phenomena are very limiting, but underline an interesting perspective on methylation variation and evolutionary potential (Feinberg and Irizarry, 2010).

Studying DNA methylation and other epigenetic mechanisms can ultimately reveal another basis underlying the organisms struggle for survival. They might shed light on missing pieces composing the phenotype (*e.g.* "the missing heritability" of complex traits; Cortijo *et al.*, 2014) and on phenomena that cannot be explained by genetic variation only. In practical senses, it might help us understand global challenges such as spread of invasive species and pests, and finally how organisms can cope with global change.

Appendix D: Ecology helps explain whether genes for cryptic coloration form a supergene or recombine

Ecology helps explain whether genes for cryptic coloration form a supergene or recombine

Romain Villoutreix[1], Clarissa F. de Carvalho[1], Víctor Soria-Carrasco[1], Dorothea Lindtke[2], Marisol De-la-Mora[1], Moritz Muschick[3], Jeffrey L. Feder[4], Zach Gompert[5], and Patrik Nosil[*1,5]

[1]*Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK*
[2]*Department of Biological Sciences, University of Calgary, Calgary AB, T2N 1N4, Canada*
[3]*Department of Fish Ecology & Evolution, EAWAG, Swiss Federal Institute for Aquatic Science and Technology, CH-6047, Kastanienbaum, Switzerland*
[4]*Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana 46556, USA*
[5]*Department of Biology, Utah State University, Utah 84322, USA*
[*]Lead contact: p.nosil@sheffield.ac.uk

Abstract

Adaptation often involves traits that are controlled by multiple genes, with specific gene combinations conferring high fitness. However, recombination breaks down favorable gene combinations. Thus, genomic regions that exhibit tight linkage and suppressed recombination among adaptive genes (i.e., 'supergenes') can promote adaptation. Putative examples of supergenes are accumulating in many organisms, hinting at taxonomic generality. However, alternatives to supergenes, such as pleiotropic effects of single loci, have rarely been assessed. Moreover, the factors favoring supergenes are often obscure. Here we address these issues by studying a supergene for cryptic coloration in *Timema* stick insects. We demonstrate that a single genetic region associated with coloration contains multiple, recombining color loci in one species, but exhibits supergene architecture in others likely due to structural changes that suppress recombination. High recombination among color genes is associated with use of host plants that exhibit fairly continuous color variation, whereas supergene architecture is associated with hosts exhibiting discontinuous colors (i.e., uniformly green leaves versus brown stems). These results led us to speculate that genetic architecture is ecologically influenced by variation in the strength of disruptive selection, a hypothesis supported by a field-transplant experiment. Our results help to explain how multi-genic variation is packaged into discrete units of diversity, such as morphs, ecotypes, and species.

Author summary

Adaptation often involves traits that are controlled by multiple genes, but recombination breaks down favorable gene combinations. Thus, genomic regions that exhibit tight linkage and suppressed recombination among adaptive genes (i.e., 'supergenes') promote adaptation. Here

we elucidate ecological factors that explain a supergene for cryptic coloration. We demonstrate that multiple, linked loci affect coloration in stick insects. We then use natural history, experimental, and genomic data to show that recombination between these loci is suppressed in some ecological circumstances (i.e., strong differences in the colors of leaves versus stems of host plants), but not in others (i.e., more continuous coloration exhibited by host plants). Our results illustrate how ecological discontinuities help package multi-genic variation into discrete units of diversity, such as morphs, ecotypes, and species.

Keywords: polygenic adaptation; chromosomal inversion; disruptive selection; structural genomic changes; ecological genomics

Introduction

It remains unclear how and why variation in polygenic traits is regularly packaged into divergent forms with few intermediates, such as discrete morphs or species [1-5]. Specifically, even if selection favors specific combinations of genes (generating linkage disequilibrium, LD, among them), recombination breaks down these combinations. Thus, discontinuous variation in polygenic traits can be difficult to evolve, at least when gene flow and recombination occurs between populations [6]. Sharp, discontinuous transitions in ecological variables are predicted to help resolve this antagonism between selection and recombination, via two complementary mechanisms [2-4]. First, such transitions may generate strong divergent selection [7-9], which maintains adaptive gene combinations more readily than weak selection. Second, such transitions could favor reduced recombination and the evolution of 'supergenes' (i.e., linked complexes of genes that segregate as major loci), for example via structural changes such as chromosomal inversions [3,4,10-14].

Evidence consistent with supergenes is accumulating in a range of organisms, largely based on multiple traits mapping to one genetic region [3,11,12,15,16]. However, this evidence is incomplete and indirect such that further studies of supergene evolution are required (Table S1 for literature review). For example, putative supergenes are often assumed to harbor multiple genetic variants that causally affect trait variation, rather than shown to do so (e.g., an alternative hypothesis is that a single gene or developmental switch has pleiotropic effects on trait variation)[12,17]. Moreover, the ecological drivers of selection are sometimes unknown, direct evidence that they favor supergene architecture is lacking, and selection strength has been inferred rather than quantified [3,11,12,18]. This is problematic because supergenes can also hinder adaptation via reduced flexibility in creating novel gene combinations, and the accumulation of deleterious mutations in regions of reduced recombination [3,11,12,15,16]. Direct estimates of the number of genetic variants affecting traits, relevant ecological variables, and selection strength are difficult to obtain [1,3,18], but are required to distinguish alternative hypotheses and to quantitatively understand the process of adaptation. We provide such estimates here using genomic data, natural history observations, and a manipulative field experiment, thereby elucidating the mechanisms underlying multi-genic adaptation and supergene evolution.

Specifically, we study wingless, herbivorous *Timema* stick insects, which rely on crypsis for protection against visual predators while resting on their host plants [19-23]. *Timema* body coloration

has thus evolved to approximate the colors of the stems and leaves of their hosts, and most species exhibit color polymorphisms that have been linked to fitness variation [19-21](e.g., green versus brown morphs that appear cryptic on leaves versus stems, respectively)(Fig. 1). It is known that color variation in *T. cristinae* segregates as a major locus on linkage group (LG hereafter) 8, named *Mel-Stripe [23,24]*. This locus spans ~10 mega-bases of sequence and exhibits suppressed recombination [24]. Although this is consistent with a supergene, this evidence alone does not rule out the alternative of a single locus with pleiotropic effects [17], nor does it indicate the number of genetic changes that contribute to trait variation or how and why recombination is suppressed [3,11,12]. Interestingly, we find here that color variation maps to a genetic region without suppressed recombination in one related species, *T. chumash*, which allowed us to quantify the number of genetic variants contributing to color variation. We then explore how and why other *Timema* species, including *T. cristinae*, exhibit suppressed recombination and supergene architecture. Although our focus is on morphs, similar processes should apply to other recognizable units of diversity, such as ecotypes or species.

Results

Variable differentiation between *Timema* color morphs. We began by studying phenotypic variation in *T. cristinae*, as well as three other species from southern California (*T. podura, T. bartmani* and *T. chumash*). All these species exhibit individuals that are green in color and others which are shades of brown, grey, or red (Figs.1c, S1-3; Table S4), and are thus known to exhibit green versus more darkly colored ('melanistic' hereafter) morphs. However, the degree of discontinuity between morphs has not been previously quantified. We used standardized photos of 1545 individuals to quantify body color in the green to blue color spectrum (a trait referred to as 'GB' hereafter), and in the red to green color spectrum (a trait referred to as 'RG' hereafter; note that *Timema* do not reflect strongly outside of the visible spectrum, OSM, Figs. S2-4, Table S6). We found that, relative to the other species and populations studied here, *T. chumash* exhibited a wider and more continuous range of color, and weaker association between GB and RG values (r^2 = 0.04, versus ~ 0.40 in the other species, Table S5). Thus, GB and RG are largely independent traits in *T. chumash*, potentially reflecting high recombination among color genes, which facilitates fine-scale genetic mapping. We thus focused our initial analyses of genotyping-by-sequencing (GBS) data on *T. chumash* (Tables S7-8), predicting that multiple genetic regions would associate with RG and GB color variation in genetic mapping analyses, with low LD among the regions.

Color variation is under multi-genic control in *T. chumash.* As predicted and in contrast to past work in *T. cristinae*, we found evidence for multi-genic control of color in *T. chumash* (Figs. 2, S5-6). We first employed a Bayesian multi-locus genome-wide association (GWA) mapping approach that accounts for LD among single-nucleotide polymorphisms (SNPs)(see Fig. 1 for details). This revealed that color maps to a ~1000 kilo-base region within the 10 mega-base *Mel-Stripe* locus of the *T. cristinae* reference genome. Notably, this region contains multiple, distinct peaks of phenotype-genotype association, generally separated from each other by several kilo-bases. Some peaks were associated with variation in only one trait (RG or GB), and accordingly the genetic correlation between RG and GB was modest (r^2 = -0.09). This provides initial evidence that a contiguous region controls color, but that multiple loci within it are involved such that control is multi-genic.

Our mapping approach further allowed us to explicitly quantify the number of genetic variants (i.e., quantitative trait nucleotides, QTN) controlling each trait, by considering how often SNPs were retained as trait-associated across different Markov chain Monte Carlo (MCMC) steps in the GWA (the proportion of such steps is termed the posterior inclusion probability, PIP hereafter, Figs. 1-2). In the case of multi-genic control with recombination among loci, the one or few SNPs that best tag each causal variant are expected to consistently be trait-associated across MCMC steps (i.e., exhibit high PIP values). Thus, PIP values across such SNPs sum to the number of total causal variants (i.e., even if causal variants are not unambiguously identified, the number of such variants can be estimated). This revealed that ~4-5 genetic variants control GB and ~3-5 control RG (Fig. 3). Thus, color is multi-genic, but not strongly so.

Also consistent with a multi-genic model rather than a single pleiotropic major effect locus, effect sizes were moderate and fairly uniformly distributed among the most strongly color-associated SNPs (Fig. 1, S5). Moreover, phenotypic color scores increasingly became more melanistic (defined by high scores for RG and low scores for GB) as the number of melanic-associated alleles an individual harbored increased (across the ten most strongly color-associated SNPs), and we did not detect evidence for strong epistasis (Figs. 2, S6). Finally, linkage disequilibrium (LD) among the top color-associated SNPs was low, indicative of recombination between them (Fig. 2). Thus, as predicted by the polygenic hypothesis, multiple linked but recombining variants affect GB and RG coloration in *T. chumash*.

The identities of the genes causally affecting color remain to be resolved. However, several genes are promising candidates (Table S9). For example, a homolog of the '*st*' gene, which causes red eye coloration in *Drosophila* [25], lies ~116 kilo-bases from a peak affecting RG. A gene with a cysteine rich flanking domain is also found near this peak, and this type of element affects yellow 'eye spot' coloration in cichlid fish [26]. Finally, another association peak is found within a protein with an UBX domain (in fact, a SNP in the third exon of this gene is unambiguously correlated with GB variation, PIP = 1). This domain is typical of ubiquitin-regulatory proteins, which are involved in the pathways causing melanic coloration in the peppered moth *Biston betularia* [27], *Heliconius* butterflies [28], and felids [29].

With our discovery of multi-genic control of color in *T. chumash*, our results suggest that the *Mel-stripe* major effect locus in *T. cristinae* represents multiple linked variants in a supergene [11,12]. Our evidence parallels that for a mimicry supergene in *Heliconius* butterflies, where multi-locus architecture in *H. melpomene* and *H. erato* segregates as a major locus in *H. numata*, due to chromosomal inversions that suppress recombination among mimicry genes [30,31]. Here, we advance understanding of supergene evolution by determining the likely cause for recombination suppression in *T. cristinae*, and then testing whether and why other *Timema* species exhibit supergene architecture.

Reduced recombination is likely due to chromosomal inversion. Analyses using population genetics and comparison of *de novo* genome assemblies of different morphs in *T. cristinae* revealed that suppressed recombination is likely due to structural genomic changes in the *Mel-Stripe* region, including a putative ~10 mega-base inversion (Fig. S7). The sizeable ~1000kb region to which color

maps in *T. chumash* coincides with one of the putative breakpoints of the inversion in *T. cristinae*. Although further cytogenetic or genomic analyses are required to definitively infer an inversion, our collective results are consistent with an inversion. Moreover, in terms of the evolutionary processes studied here, suppression of recombination is relevant no matter the precise mechanism for it.

Multiple *Timema* species exhibit suppressed recombination. We found that suppressed recombination previously reported in *T. cristinae* is also evident in *T. bartmani* and *T. podura*. For example, in these species we observed 'block-like' patterns of association on LG8 in single-locus GWA mapping, high genetic correlations between RG and GB, and strong LD in the *Mel-Stripe* region (Figs. 3-4, S8-11). These patterns are indicative of reduced recombination. Further supporting suppressed recombination, principal components analyses (PCA) of genetic variation in the ~1000kb region harboring color loci revealed distinct and color-morph-associated genetic clusters (i.e., chromosomal forms) in *T. bartmani* and *T. podura* (Fig. 4), as previously reported in *T. cristinae* [24]. In strong contrast, PCA in *T. chumash* revealed a dispersed cloud of points, rather than genetic clusters. Finally, phased genomic data revealed morph-associated haplotype blocks in *T. bartmani* and *T. podura,* again supporting reduced recombination (Fig. 4).

Although the phylogeny of *Timema* does not allow us to distinguish whether *T. chumash* lost an ancestral supergene or other species gained it [32], under either scenario our results are consistent with structural features enhancing discontinuity between color morphs. We note that phylogenetic relationships for *T. podura*, *T. bartmani*, and *T. chumash* inferred from GBS and new whole genome re-sequencing data revealed that the ~1000kb region harboring color loci tends to reflect the species tree, not grouping by the same color morph across different species (Fig. 5). This suggests the supergene alleles are not of recent origin, as reported for *T. cristinae* [24], and that they were not recently transferred between species by hybridization. We next turned to possible explanations for variation in genetic architecture and morph differentiation among species.

Host-plant coloration is associated with morph differentiation. Although the morphs of *T. chumash* form two recognizable and statistically supported clusters, they are less distinct than those in *T. cristinae* and *T. bartmani* (Fig. S1). Specifically, the color distance between morphs is *T. chumash* < *T. bartmani* < *T. cristinae* (mean Kullback Leibler distance between morphs, *T. chumash* = 14.0, *T. bartmani* = 17.4, *T. cristinae* = 29.2; posterior probabilities, *T. bartmani* > *T. chumash* = 0.87, *T. cristinae* > *T. bartmani* = 0.98, *T. cristinae* > *T. chumash* ~1.0, Fig. 6). We suspected that increased color discontinuity between morphs was associated with the use of hosts that exhibit highly discontinuous color variation [21,33]. This ecological hypothesis predicts that greater discontinuity in the colors offered by the leaves versus stems of *Timema* hosts will correspond to increased phenotypic discontinuity in color between *Timema* morphs [7-9]. We tested this prediction using standardized photos of main hosts of three *Timema* species for which we were able to collect host data (Table S10).

Consistent with the ecological hypothesis, the hosts of *T. chumash* (oak and mountain mahogany) express a wide and fairly continuous range of variation in their leaves and stems, including shades of blue, green, yellow, tan, beige, brown, and red (Fig. 6). As a result, the oak and mountain mahogany hosts of *T. chumash* displayed the lowest color difference between their leaves and stems (mean

Kullback Leibler distance = 10.1), compared to the white pine and white fir hosts of *T. bartmani* (21.8), and chamise and California lilac hosts of *T. cristinae* (39.3; posterior probabilities, *T. bartmani* > *T. chumash* hosts = 0.99, *T. cristinae* > *T. bartmani* hosts = 0.97, *T. cristinae* > *T. chumash* hosts ~1.0). These results led us to speculate that the hosts of *T. chumash* select only weakly for specific combinations of green versus melanistic coloration alleles, providing an ultimate explanation for why 'morphs' of *T. chumash* are less discrete. In contrast, the hosts of *T. bartmani* and *T. cristinae* exhibit increasingly greater color distance between their leaves and stems. These hosts could thus offer a more bi-modal and discontinuous range of colors, i.e., primarily green or brown, which could select more strongly for specific combinations of coloration alleles [7-9].

The strength of disruptive selection varies among hosts. We experimentally tested the prediction of stronger disruptive selection against intermediate coloration on hosts with greater color discontinuity, using a field-based recapture study (note that even if two morphs exist some individuals can be more intermediate in coloration than others, i.e., those nearer the center of phenotype space, and we here found classification of intermediates to be statistically repeatable, see Methods). We did so by marking and transplanting equal numbers of green, melanistic, and intermediately colored *T. chumash* to host plants comprising two treatments: (1) chamise and California lilac (hosts offering highly discrete coloration) and, (2) mountain mahogany (a host offering more continuous color variation). Consistent with prediction, we recaptured a lower proportion of intermediates in the chamise and California lilac treatment (posterior probability that survival of intermediates is greater in the mountain mahogany treatment > 0.99, multinomial-Dirichlet model, Fig. 7). Thus, we detected strong disruptive selection in the chamise and California lilac treatment ($s = -2.73$, posterior probability, pp, that $s < 0 = 0.97$; $t = -1.84$, pp $t < 0 = 0.92$, where fitness is defined as green = 1-s, intermediate = 1, and melanistic = 1-t, i.e., s or $t < 0$ implies disruptive selection, and s or $t > 0$ implies intermediate advantage). In contrast, selection was not disruptive on mountain mahogany ($s = 0.38$, pp $s < 0 = 0.14$; $t = 0.60$, pp $t < 0 = 0.03$), consistent with the wide range of color exhibited by this host. Nonetheless, selection may be weakly disruptive on other common hosts of *T. chumash* (e.g., oak), or at time periods and locations other than which our experiment was conducted. Indeed, morph differentiation, or even genetic architecture, might vary within species. In any case, for the populations studied here our results provide concordant observational and experimental support for the hypothesis that discontinuity between morphs is mediated by ecological discontinuity.

Discussion

Our results are consistent with ecological factors explaining not only trait evolution, but also the degree to which traits are packaged into discrete units of diversity. They add to other major studies in Darwin's finches where seed size distributions drive beak evolution [34,35], in stickleback where predator regimes drive bony armor evolution [36,37], and in apple maggot flies where divergent fruiting times drive differences in diapause timing between host races [38]. Our results further show how the genetic architecture of traits can change between recombining, polygenic variation and major locus (i.e., supergene) control.

Our findings help advance understanding of evolution because the plausibility and mechanisms of large or sudden evolutionary changes remains unclear [39,40]. Developmental biology provides one possible mechanism: developmental switches involving gene regulation [17,39,40]. Our results illustrate another: the conversion of polygenic variation, gradually accumulated by selection, into discrete phenotypic categories by supergene evolution. Thus, supergenes may help reconcile large evolutionary shifts and ideas concerning macro-mutation (i.e., 'hopeful monsters') with polygenic adaptation and Darwinian gradualism.

Materials and Methods

***Timema* sampling.** *Timema* were collected by shaking the branches of host plants while holding a sweep net underneath them, as in past work [21,24]. Adult (i.e., sexually mature) specimens were stored in plastic containers for immediate photographing (details below). Juvenile individuals were reared on *Ceanothus spinosus* cuttings in plastic containers until they reached adulthood, as in past work [21], and then photographed. We took digital photographs of every adult *Timema*, and then stored each specimen in an individual vial in pure ethanol for subsequent molecular work.

The data presented in this manuscript are primarily newly acquired. Specifically, new data was collected for all four species studied here: *T. bartmani, T. chumash, T. cristinae,* and *T. podura*. We also reanalyzed some data for *T. cristinae* and *T. podura*, from [21,33](Table S2). Tables S2-4 and S6-8 contain details of the *Timema* populations and samples used in the different analyses of this study.

Phenotypic measurements of *Timema* coloration from photographs. Standardized digital photographs of adult *Timema* were taken, with the exception of *T. bartmani* that develop later in the season than other species and were thus photographed at juvenile stage (note that our core conclusions are unaffected by this as they do not rely on *T. bartmani* alone, and current and past work shows color morph is highly heritable such that it persists across life history stages)[21,24,33].

All individuals were photographed with a digital Canon EOS 70D camera equipped with a macro lens (Canon EF 100mm f/2.8L Macro IS USM) and two external flashes (Yongnuo YN560-II speedlights). The images were taken with the camera set on manual, an aperture of f/14, a shutter speed of 1/250 s, a sensitivity of 100 ISO, and flashes adjusted to 1/4 power in S2 mode in an output angle corresponding to 24-mm focal length on full frame (~84° diagonal). To avoid shadows and reduce external luminosity interference, LumiQuest SoftBox LTp softboxes were attached to the flashes. In addition to the *Timema* specimens, the pictures included a ruler and a standard color chip (Colorgauge Micro, Image Science Associates LLC, Williamson, NY, USA).

Each specimen was photographed at least twice in different perpendicular positions to capture the body color without gleam or shade. The pictures were linearized and corrected for white balance, adjusting the temperature and the tint 1 based on the values obtained from the color chip neutral grey color (target #10), using ADOBE PHOTOSHOP LIGHTROOM 5.7 software (Adobe Systems Software Ireland Ltd). Due to the standardization, measurements did not vary appreciably among pictures and only minor corrections were necessary (similar procedures in past work have shown color

measurements to be highly repeatable)[21,41,42]. The pictures were adjusted for the temperature to 5950 and for the tint to +2, and exported as TIFF files.

From the standardized images we collected phenotypic measurements using the software IMAGE J 1.4.882 [43]. To quantify variation in color, we recorded mean RGB (Red, Green, Blue) values using the polygon section tool and color histogram plugin in ImageJ. For every *Timema* specimen, we measured a small area in the lateral margin of the insect's dorsal region both on thoracic and abdominal parts, and analysed the mean values between these body parts

For best interpretation of the RGB information, we processed the values as the relative difference between red and green (RG), and between green and blue (GB)[following 44]. RG channel was obtained using the relationship (R-G)/(R+G), and GB by (G-B)/(G+B), as described in the literature [44], and previously used to measure *Timema* color [33]. Although this method does not take into consideration how color is sensed by a predator, it does yield an objective quantification of color to be used in a comparative context. Thus, we obtained and analyzed two different color variables available for each insect: lateral RG and lateral GB (RG and GB hereafter). Further justifying our approach based on photos, our independent analysis of spectral reflectance finds that *Timema* specimens reflect only marginal levels of ultraviolet light, with very weak effects on stimulating photoreceptors in avian (i.e., predator) ultraviolet-sensitive systems as described in detail in the Supplementary Data (Spectra reflectance data).

Differentiation and overlap between *Timema* morphs. We used the UPGMA algorithm in hclust (from R 3.2.3)[45] to cluster *Timema* into two groups (i.e., morphs), using the Euclidean distance between every individual based on RG and GB color measurements. We then used a Bayesian approach to fit the color data for each morph to a bivariate normal distribution. We placed relatively uninformative priors on the mean vectors (normal with mu = 0 and tau = 1e-3 for both means) and for the precision matrix (Wishart with 2 degrees of freedom and a diagonal scale matrix = 0.001 $\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix). We used Markov chain Monte Carlo (MCMC) to obtain samples from the posterior distribution via the rjags (version 4.6) interface with JAGS (version 4.1.0)(1000 iteration burn-in, 5000 sampling iterations and a thinning interval of 4). We then estimated the Kullback-Leibler distances (that is, the Kullback-Leibler divergence in both directions, e.g., from morph 1 to morph 2 and morph 2 to morph 1). This was calculated over the posterior distribution of the bivariate normal parameters, and thus accounts for uncertainty in these parameters. The main text shows these results for the three *Timema* species for which we also obtained data on host-plant coloration. In all instances, the morphs of *T. chumash* are less differentiated than those of other *Timema* species (Fig. S1).

Correlation between phenotypic variables in *Timema*. The correlation between the variables RG and GB was estimated for each species. Values from individuals from different populations were pooled to obtain larger sample sizes for the analysis. Coefficient of determination (R^2) was estimated using linear models in R [45]. R^2 was also estimated using data from 2013, published in previous studies [21,33]. The statistics show a consistently high association between RG and GB in most species (Table S5). Statistics were estimated using R [45].

Genotyping-by-sequencing, alignment, and variant calling. We genotyped by sequencing a total of 1529 individuals from *T. chumash* and *T. bartmani* (deposited as NCBI BioProject XXX, Dryad repository xxx). We extracted genomic DNA of each individual from three to five legs using DNeasy Blood and Tissue Kit (Qiagen). We then generated barcoded single end DNA libraries for each individual following standard restriction-site protocols [46], as applied in several previous *Timema* studies [21,24,32,33,47-49]. These individual libraries were then distributed into pools (containing sets of different individuals). These pools were size selected for fragments of size 300-500 base pairs (including adaptors) and sequenced (one pool per lane) on a Illumina HiSeq2000 platform with V3 reagents at the National Center for Genome Research (Santa Fe, New Mexico, USA). For downstream analyses, we also used datasets of *T. cristinae* and *T. podura* from previous studies [21,33](NCBI BioProjects PRJNA284835 and PRJNA318846, Table S2 for a summary of data new to this study versus previously published data). The complete dataset used hereafter comprised the genotypes of 2181 individuals across the four *Timema* species (1529 samples new to this study plus 652 previously published samples, Table S7). Note that this number is higher than the sum of samples used for GWAS (Table S8), because it includes additional samples used for generating the consensus sequences (described in detail in the Supplementary Data). The generated sequences were used to obtain allele frequencies and genotype probabilities as described in detail in the SI (Genotyping-by-sequencing, alignment, and variant calling). Posterior genotype probabilities we obtained were used for all multi-locus GWA mapping analyses in GEMMA [50]. GENABEL [51], which was used for single-locus GWA, cannot handle genotype posterior probabilities and requires called genotypes. Thus, for analyses using GENABEL we called genotypes from genotype probabilities (GP) using the following thresholds: GP <=0.5: homozygote for reference allele; 0.5 < GP <1.5: heterozygote; GP >= 1.5: homozygote for alternate allele. Posterior genotype probabilities we obtained were used for all multi-locus GWA mapping analyses in GEMMA 0.94 [50]. GENABEL v1.8.0 [51], which was used for single-locus GWA, cannot handle genotype posterior probabilities and requires called genotypes. Thus, for analyses using GENABEL we called genotypes from genotype probabilities (GP) using the following thresholds: GP <=0.5: homozygote for reference allele; 0.5 < GP <1.5: heterozygote; GP >= 1.5: homozygote for alternate allele.

Detection of chromosomal inversion in *T. cristinae*. Several lines of evidence were used to delineate the approximate breakpoints for a putative large inversion in *T. cristinae* that is associated with green versus melanistic color morphs (Fig. S7). We focused on two scaffolds on LG8 (702.1 and 128) where a large number of contiguous SNPs were associated with color, suggestive of an inversion or a region of otherwise reduced recombination [24,49].

We started by using a comparative alignment of *de novo* genome assemblies from melanistic and green *T. cristinae* morphs to constrain the possible bounds of the putative inversion. Both genome assemblies combined data from standard fragment libraries, mate-pair libraries, and Dovetail Chicago libraries. These assemblies, along with the comparative alignment, were described in detail in [49]. The alignment between melanistic morph scaffold 702.1 and green morph scaffold 1575 indicated that these genomes were co-linear along scaffold 702.1 up to base pair 10,032,025 (the end of the alignment between these two scaffolds). Likewise, we found that scaffold 128 from the melanistic genome aligned to scaffold 4214 from the green genome, and that these two scaffolds were co-linear beyond the boundary of the GWA signal in *T. cristinae* at ~6 megabases on scaffold 128. Thus, we

fixed the breakpoints for the putative inversion between ~10 megabases on melanistic scaffold 702.1 and ~6 megabases on melanistic scaffold 128. This region corresponds to, but is slightly narrower than, the broad region of elevated GWA signal for color from the single SNP GWA analysis (see Fig. S7). Within this region, the green morph genome comprises many small scaffolds, preventing clear identification of the inversion based on these data alone. We think that the reason for poorer assembly in this region for the green morph was that the individual used for the *de novo* assembly was heterozygous for the green and melanistic haplotypes (and thus for the inversion), creating difficulty with the assembly. Our approach moving forward was thus as follows.

We fit a hidden Markov model (HMM) based on patterns of LD across scaffolds 702.1 and 128 to explicitly test for and better resolve the bounds of the putative inversion, using the *R* (version 3.4.2) package *HiddenMarkov* (version 1.8.11)[52]. In *T. cristinae* individuals homozygous for the brown morph haplotype, we would expect normal/high LD for SNPs on either side of the breakpoint when sequences are aligned to the melanistic morph reference genome. In contrast, for individuals homozygous for the green haplotype, LD should be lower when SNPs span the inversion breakpoint, as such SNPs are not actually physically near each other and thus have a greater opportunity for recombination to reduce LD. The pattern described above is the specific signal we thus next searched for, as described in detail in the Supplementary Data (Detection of putative chromosomal inversion in *T. cristinae*).

Multi-locus genome-wide association mapping in *T. chumash* with GEMMA. As in previous work [21,24,32] we used the software GEMMA 0.94 [50] for multi-locus GWA mapping in *T. chumash*. This method accounts for linkage disequilibrium among SNPs and is thus well suited for localizing genotype-phenotype associations within the genome, as was the goal in *T. chumash*. Briefly, we used GEMMA to implement Bayesian sparse linear mixed models (BSLMMs) using a multiple-SNP Bayesian approach to model the genetic architecture of color variation while accounting for genetic relatedness among individuals. In this method, the effects of SNPs are modeled as a mixture of two distributions: (1) those that individually have infinitesimal effects ('polygenic distribution') and, (2) those with measurable (i.e., 'sparse') effects. This approach provides posterior inclusion probabilities (PIPs, also called γ parameter) for each SNP, which represent the fraction of MCMC iterations that the SNP was retained as having a measurable effect. PIPs thus reflect the weight of evidence that an individual SNP is associated with the trait of interest. As described in detail below, PIPs also form the basis for quantitatively estimating the number of causal variants affecting a trait.

We estimated PIP values for BSLMMs applied separately to RG and GB values, calculated across 5 independent MCMC runs per trait (prior to GWAS, we corrected the color measurements for differences between sexes by extracting the residuals using sex as an independent variable in linear models). For each chain, we ran 3,000,000 iterations with a recording pace of one record state in every 100 steps and discarded the first 1,000,000 iterations as burn-in. We excluded SNPs with a minor allele frequency (MAF) less than one percent.

Estimating number of variants affecting color from the GEMMA model. We obtained Bayesian estimates of the number of genetic loci (i.e., quantitative trait nucleotides, or QTN) within *Mel-Stripe* that were associated with GB and with RG color traits for *T. bartmani*, *T. chumash* and *T. podura*.

This number represents the number of causal variants affecting each trait, and is estimated via the sum of PIPs in a region [50](Fig. 1). In the case of multi-genic control with recombination among loci, the one or few SNPs that best tag each causal variant are expected to consistently be trait associated across MCMC steps (i.e., exhibit high PIP values). Thus, PIPs across such SNPs sum to the number of total causal variants. In contrast, in the case of suppressed recombination for example via inversion, many SNPs with very low (but non-zero) PIPs are expected because different SNPs can readily tag the causal variants (i.e., many SNPs carry redundant information). This will lead to PIP values of multiple SNPs across a genetic region that sum near one.

We estimated the number of QTN in a way that accounted for uncertainty in individual SNP-trait associations as measured by the PIPs from GEMMA [53]. For each species and color trait, we drew samples from the posterior distribution of the number of QTN in *Mel-Stripe* by sampling a binary indicator variable (1 = QTN for color, 0 = not QTN for color) for each SNP in that region based on its PIP. The sum of SNPs in the region sampled as QTN was then taken as a posterior sample for the number of QTN in *Mel-Stripe*. We repeated this procedure 10,000 times for each species and color trait to obtain posterior distributions for the number of QTN, which we summarized based on their medians and 95% ETPIs (i.e., the 2.5th and 97.5th quantiles of the distribution). We also estimated the genetic correlations between GB and RG as described in the Supplementary Data (Genetic correlation between GB and RG).

Linkage disequilibrium between color-associated SNPs in *T. chumash* multi-locus GWA. We quantified linkage disequilibrium (LD) among the SNPs most strongly associated with color in *T. chumash* based on the multi-locus GWA mapping results from GEMMA (see above). Pairwise LD was quantified as the squared Pearson correlation between genotypes at each pair of SNPs with high posterior inclusion probabilities (PIPs)[as in 54]. Specifically, we considered SNPs with PIPs greater than 0.4 in the region showing strong associations (5-6 mega-base region on scaffold 128). This revealed that LD was generally low in this region, and was not accentuated for trait-associated SNPs (Fig. 2).

Distribution of effect sizes, dominance, and epistasis. To further characterize how closely the genetic architecture of color in *T. chumash* is polygenic and predominantly additive we considered the distribution of phenotypic effect sizes across the most strongly color-associated SNPs from the multi-locus mapping, and tested for epistasis between these SNPs. A polygenic model predicts a fairly uniform distribution of effect sizes (small to moderate effects for most SNPs), and little or no epistasis. Our results are largely consistent with these predictions (Figs. 2, 3, S5-6, and details below), as described in detail in the Supplementary Data (Distribution of effect sizes, dominance, and epistasis).

Single locus genome-wide association (GWA) mapping with GENABEL. Following past work [21,33] we used GENABEL v1.8.0 [51] to perform single locus GWA mapping analysis in *T. bartmani*, *T. chumash*, and *T. podura*. This method does not account for LD among SNPs and is thus well suited for visualizing larger 'blocks' of genotype-phenotype association within the genome, as might occur in regions of reduced recombination.

Briefly, transformed genetic probabilities were filtered using the GENABEL quality control function. Excluded from analysis were SNPs with MAF less than or equal to 1%, individuals with extreme heterozygosity at a false discovery rate <1%, and individuals with identity by state (IBS) >= 0.95, (calculated on a randomly selected subset of 2000 SNPs). Analyses were run both with and without control for population structure, and gave qualitatively comparable results. We stress that we largely analyzed samples collected in the same locality, and that the core point of the single locus mapping was to visualize block-like patterns of association, not to detect causal loci affecting color. Thus, for our purposes potential population structure is less problematic than in studies aiming to find casual variants.

Association results taking population structure into account were obtained using the GENABEL egscore function. This function implements the method of [55] and extracts principal components of a kinship matrix (here IBS indices) calculated using a randomly selected subset of 2000 SNPs (excluding those from LG8 and SNPs not associated to a linkage group). The principal components are then used as covariates in the GWA linear models. Results are displayed in the form of Manhattan plots. These graphics show the association score (expressed as $-\log_{10}(pvalue)$) of every SNP tested along their physical position in the *T. cristinae* genome. Gaps between scaffolds are not represented in these graphics.

Our results revealed that at the scale of LG8, *T. chumash* exhibits a peak of association (which actually represents several distinct peaks when zoomed in further on scaffold 128), *T. bartmani* a narrow 'block' of association, and *T. podura* a wide block of association (Figs. S8-S11). Accordingly, *T. chumash* exhibits much lower LD in the *Mel-Stripe* region than do the other species (Fig. 3).

Linkage disequilibrium between color-associated SNPs in single-locus GWA. We calculated LD among all SNPs in the *Mel-Stripe* locus [following 49] for the three *Timema* species. Pairwise LD was quantified as the squared Pearson correlation between genotypes at each pair of SNPs in this region for each species. We then summarized the distribution of LD across the region by the median and 95% quantile across all pairwise LD estimates.

Principal component analysis (PCA) on genotypes. Posterior genotypes probabilities were obtained and used for all PCA analysis. PCA was run on all SNPs located between position 5Mbp and 6Mbp on scaffold 128 with the prcomp function in R [45]. The position of individual's on principal component axes was extracted and plotted using custom scripts.

Structure based on phased genomic data. We obtained phased haplotypes for *T. bartmani*, *T. chumash* and *T. podura* using fastPHASE 1.4.8 [56]. We used the same GBS data that was also used for GWA mapping for these species. Following past work [24], we computed phred-scaled genotypes likelihoods for SNPs with a maximum of 15% missing values (85% of samples need to have at least a read to estimate a genotype likelihood) using a custom perl script (bcf2gl.pl) and we formatted these genotype likelihoods into a fastPHASE input file using another custom per script (mkfastphaseinp.pl). We ran fastPHASE allowing for 20 random starts (-T 20 option), a maximum of 35 EM iterations (-C35 option), a lower limit for K detection of 10, an upper limit of K detection of 20 and an interval for K detection of 2 (-KL 10 -KU20 -Ki 2 options), scanning for genotype errors using a 4 parameter

model (-em4 option) and bracketing genotypes with a posterior probability below 0.9 (-q 0.9 option). fastPHASE output consisted of 2 haplotype files, one for each DNA strand.

We then used the linkage model in structure (version 2.3.4)[57] to test for haplotype blocks in the 5-6 megabase pair region of scaffold 128 associated with color in *T. chumash*. Haplotype blocks in this region would provide evidence of chromosomal variants with restricted recombination associated with distinct color pattern alleles in *T. bartmani* and *T. podura*. As in past work with *T. cristinae* [24], we fit the linkage admixture model for the phased genotype data from fastPHASE for all SNPs on linkage group 8 (see the preceding section). We used the correlated allele frequencies model with the number of groups (source populations) set to $k = 2$ (to reflect two putative chromosomal variants). We ran five MCMC chains for each species, each with a burnin of 200,000 iterations followed by 200,000 sampling iterations. We used the local ancestry probabilities from the site-by-site (linkage) analysis to identify SNPs that were likely homozygous for ancestry from the same source population (chromosomal variant) or heterozygous for ancestry. We called ancestry/source in cases where the posterior probability of ancestry of a given type was > 0.5 (other cases were considered ambiguous). We only expect these ancestry assignments to be meaningful/informative in the putative chromosomal variants.

Phylogenetics. We used variants from both GBS and whole genome re-sequencing (WGS) data to infer genome-wide and color-associated genetic region (on LG8) trees. We subsampled our extensive GBS dataset to include the ten individuals with the highest number of mapped reads per species (*T. bartmani*, *T. chumash* and *T. podura*) and morph (green and melanistic), resulting in a dataset of 60 individuals. Because the GBS dataset did not include numerous SNPs that overlapped between species in the color-associated region, we generated WGS data to confirm the results from the GBS data. Specifically, WGS data was obtained for a total of 48 individuals from *T. bartmani* (green and melanistic), *T. chumash* (green) and *T. podura* (melanistic, all deposited as NCBI BioProject XXX). Details on the used samples, DNA extraction, sequencing, alignment, and variant calling are provided in the Supplementary Data. We used a custom Perl script to generate multiple alignments from the genotypes with the highest likelihood and coding heterozygotes as IUPAC ambiguities. For the genome-wide inferences, we produced alignments concatenating 25,000 variants randomly taken from across all scaffolds assigned to linkage groups. The alignments for the region associated with color (5 Mbp to 6 Mbp on scaffold 128 of LG8) comprised 305 (GBS) and 1923 variants (WGS). RAxML implements the Lewis ascertainment bias correction [58], but it requires at least one unambiguous sample (i.e. homozygote) for each allele for a position to be recognized as variable. Thus, a number of positions were excluded, resulting in alignments of size 151 (GBS, color region), 12,498 (GBS, genome wide), 1173 (WGS, color region), and 13,682 (WGS, genome wide). For each alignment, we inferred maximum-likelihood (ML) trees using RAxML 8.2.11 [59] from a randomized stepwise addition order parsimony starting tree. We used a GTR substitution model with a GAMMA model of rate heterogeneity and the Lewis ascertainment bias correction ("-m ASC_GTRGAMMA --asc_corr=lewis"). We used the rapid bootstrapping approach ("-f a")[60] with the number of bootstrap replicates automatically determined using the extended majority-rule consensus tree bootstrapping criterion ("-N autoMRE")[61]. Plots were generated using R packages ape 5.1 [62], phytools 0.6-60 [63] and phangorn [64]. Alignments, trees, and code are available in Dryad repository XXX.

Phenotypic measurements of plant coloration from photographs. We quantified the coloration of the main host plants of *T. chumash, T. bartmani,* and *T. cristinae* using digital photographs of the hosts (we were unable to get photographs of the hosts of *T. podura*), using the same photographic procedures described above for *Timema* specimens. Host-plant cuttings were collected in 2015 from the localities listed in Table S10, and kept in a cooler until they were photographed. For the plant tissues, we measured a standard area of a circle with 1-millimeter diameter. This area was chosen because it fits the lateral margin size of all specimens studied (including the *T. bartmani* nymphs). In addition, this allowed us to measure samples of broad leaves and individual needles using the same standard. For samples with broad leaves (i.e., C (California lilac): *Ceanothus spinosus*; MM (Mountain Mahogany): *Cercocarpus sp.*; Q (Oak): *Quercus sp.*), we recorded the RGB values for the upper (adaxial) and lower (abaxial) leaf surfaces in different samples, and for the stem. For plants with needle-like leaves (i.e., A (Chamise): *Adenostoma fasciculatum*; WF (White Fir): *Abies concolor*; P (Pine): *Pinus sp.*), we recorded one measurement if the surface was uniform in color (i.e., P), or two if the colors varied in the upper and lower surfaces (i.e., WF). Host plant parts were categorized as stems or leaves. As we did for estimating differentiation between *Timema* morphs, we then estimated the Kullback-Leibler distance between plant parts in both directions (e.g., from stems to leaves and leaves to stems). See Table S10 in the Supplementary Data (Phenotypic measurements of plant coloration from photographs) for details about the host plant samples used in this study.

Manipulative field experiment. We tested experimentally the prediction of stronger disruptive selection (i.e., stronger selection against intermediate coloration) on hosts associated with greater differentiation of *Timema* morphs. We did so by marking and transplanting green, melanistic, and intermediately colored *T. chumash* to two treatments: (1) hosts associated with highly discrete morphs (*Adenostoma* and *Ceanothus* respectively, A/C hereafter) versus, (2) a host associated with less discrete morphs (mountain mahogany, MM hereafter). We used *T. chumash* because this species exhibits the most continuous range of color such that reasonable numbers of intermediately colored individuals could be collected to have their survival assayed (alongside with clearly green or melanistic individuals). The rationale for the choice of these hosts / treatments is provided in the Supplementary Data (Manipulative field experiment – host and treatment rationale).

The experimental *T. chumash* were collected from *Cercocarpus* in the vicinity of the locality Horse Flats 5 (HF5, N 34 15.584, W 118 6.254). A total of 602 individuals were collected between May 9 and May 11, 2018. These were kept alive in plastic containers and moved to laboratory space on the campus of the University of California, Santa Barbara.

On May 12, 2018 we scored 120 of these individuals into three phenotypic categories as follows. To represent the green category, forty of the brightest and darkest green specimens were selected to represent one extreme of the green-melanistic continuum. Thus, the colors of these chosen individuals resemble the discrete variation found in green morphs of *T. cristinae* and the other polymorphic species analyzed in this study. To represent the intermediate category, we selected forty individuals with green-yellow, yellow, brown-yellow tones, and green-blue tones, as these collectively depict the transition from green to brown (melanistic) colors. For this category, green-yellow coloration was present in the majority of individuals. To represent the melanistic category, we chose forty individuals with the darkest brown and red coloration.

We estimated the repeatability of this scoring to be 96% (95% ETPIs from a Bayesian beta-binomial model with a Jeffreys prior = 88-99%, this model has an analytical solution), by scoring 50 individuals twice, where only two scoring errors were made (21 individuals scored green both times, 15 individuals scored intermediate both times, 12 individuals scored melanistic both times, 2 individuals scored as intermediate once and as green once). Representative specimens of each category are shown in Figure 7 of the main text.

To ensure we could distinguish our experimental animals from naturally occurring ones, we marked each individual on the abdomen with a fine-tipped sharpie pen, as in past work [19,41]. The marks were thus not visible when the insects were naturally resting on their host plants. Each category of color (green, intermediate, melanistic) received a differently colored mark, facilitating accurate rescoring of color in recaptured specimens. As our experimental design involved two blocks (details below) we alternated which color mark was assigned to which category (block 1: greens marked with a blue pen, intermediates marked with a green pen, melanistics marked with a red pen; block 2: greens marked with a green pen, intermediates marked with a red pen, melanistics marked with a blue pen).

On May 13th, 2018 we transplanted the marked specimens back onto host plant individuals at the locality they were collected from. This was done in two blocks, where each block contained each treatment (MM and A/C), using a single plant individual of each host species. Equal numbers of green, intermediate, and melanistic individuals were released on each treatment and block (i.e., 20 individuals of each category on each treatment and block, total n = 120). The location of each experimental plant was as follows: block 1, MM N 34 15.584, W 118 6.254, A/C N 34 15.599 W 118 6.256; block 2 MM N 34 15.682 W 118 6.127, A/C N 34 15.631 W 118 6.216). Experimental plants were chosen to be separated from other plants by ‘bare ground’ (sandy or gravelly regions not containing plants), forming an ‘experimental island’. Past has shown that dispersal across such bare ground is near absent [41,65-68].

We were interested in rapid changes in the frequency of each color category because past studies in *Timema* have documented adaptive divergence between experimental populations within a week upon transplantation to new environments, and because adult and penultimate instar *Timema* tend to live for only one to three weeks in the field, with bird predation being a major source of selective mortality [41,65,66,68]. Thus, on May 15th, 2018 we recaptured the surviving individuals using visual surveys and sweep nets. In total, the number of recaptured individuals of each category and treatment was as follows (see also Fig. 7). On MM we recaptured 8, 13, and 5 individuals that were green, intermediate, and melanistic, respectively. On A/C we recaptured 8, 2, and 6 individuals that were green, intermediate, and melanistic, respectively. Past mark-recapture work has shown this protocol is highly effective at recapturing the overwhelming majority of surviving individuals [41,65-68].

The recapture data were analyzed as follows. We fit a Bayesian multinomial-Dirichlet model to these data using the rjags interface with JAGS (JAGS version 4.1.0, rjags version 4.6, R version 3.2.3)[69]. Specifically, recapture counts were assumed to follow a multinomial distribution with a vector *w* of length three, that gives the relative fitnesses of the three color-categories. These relative fitnesses can be rescaled (e.g., relative to the fitness of any one color category) to aid interpretation of the results.

We placed an uninformative Jeffreys Dirichlet prior on this vector (all shape parameters set to 0.5). We considered two models, one where the two blocks had independent *w* vectors and one where they were constrained to be the same. Posterior distributions were obtained by running three MCMC chains each with a 1000 iteration burnin, 9000 sampling iterations and thinning intervals of 3. The constrained model was preferred by DIC (DIC = 36.81 for the constrained model versus 54.99 for the unconstrained model), and thus we focus on results from that model. With that said, for both models the posterior probability (pp) that intermediate morphs had a higher relative fitness on MM than C/A was > 0.99. We defined the relative fitnesses (w) of the color morphs as: $w_{green} = 1\text{-}s$, $w_{intermediate} = 1$, and $w_{melanistic} = 1\text{-}t$ (as in Eq. 1.25c in [70]), with details provided in the Supplementary Data (Manipulative field experiment – relative fitnesses estimation).

Acknowledgements. The work was funded by a grant from the European Research Council (NatHisGen R/129639, https://erc.europa.eu/) and a fellowship from the Royal Society of London to PN. V. Soria-Carrasco was supported by a Leverhulme Trust Early Career Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank T. Reimchen and D. Ayala for discussion and comments on previous versions of the manuscript, J. Stapley for the use of her spectrophotometer, and T. Oakley for lab space. We thank Ó. Mira Pérez, L. Lloyd, J. Thakrar, F. Whiting and A. Štambuk for assistance with lab work. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged, as well as access to the High Performance Computing Facilities, particularly to the Iceberg and ShARC HPC clusters, from the Corporate Information and Computing Services at the University of Sheffield.

Author contributions. RV, CFC, ZG, and PN conceived the project. RV, CFC, VS, DL, MM, and PN collected data. RV, CFC, VS, MD, and ZG led data analysis, aided by all authors. All authors contributed to writing.

Competing interests. The authors declare no competing interests.

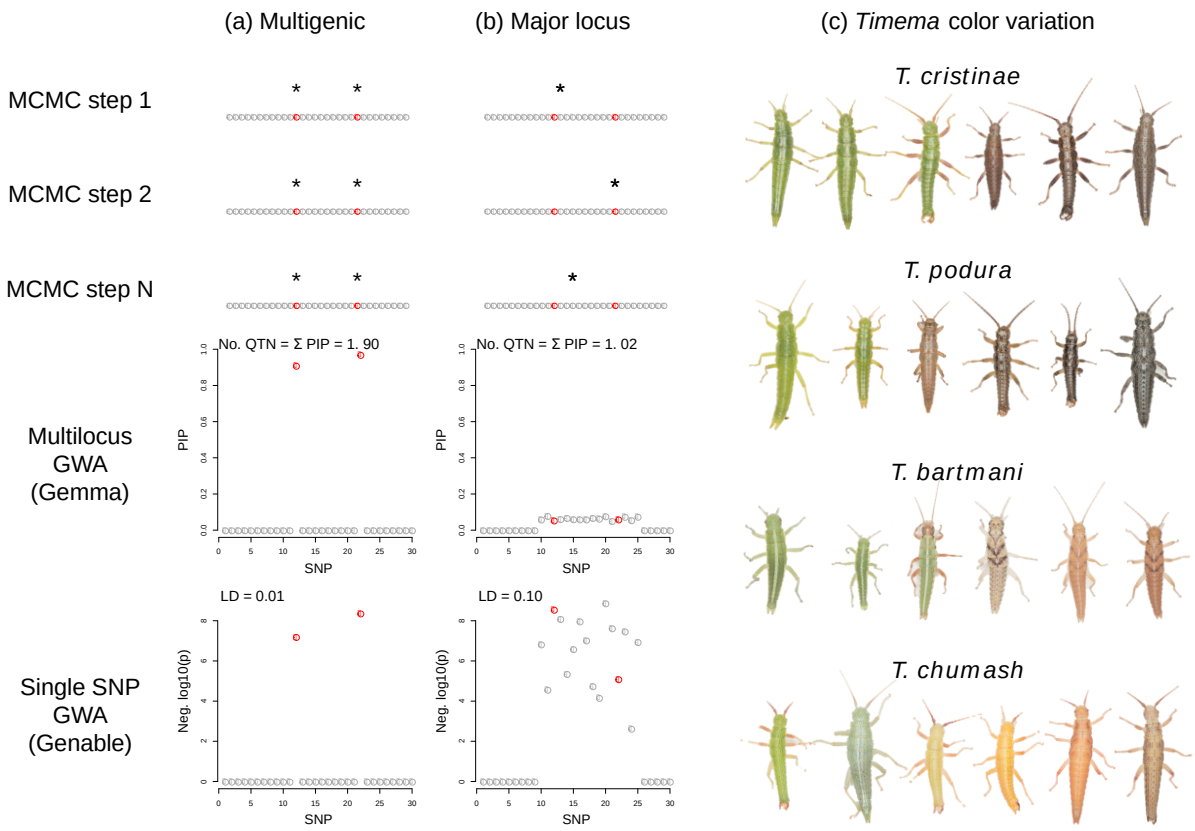


Figure 1. Predicted and observed patterns for multi-genic versus major (e.g., single non-recombining locus) genetic control of color. Red dots in (a) and (b) represent causal variants affecting color. Asterisks represent single nucleotide substitutions (SNPs) that are retained as trait-associated in each different Markov chain Monte Carlo (MCMC) step in multi-locus genome-wide association (GWA) mapping (this controls for linkage disequilibrium (LD) among SNPs). The proportion of steps that a SNP is retained is the posterior inclusion probability (PIP). In the case of multi-genic control with recombination among loci, the one or few SNPs that best tag each causal variant are expected to consistently be trait-associated across MCMC steps (i.e., exhibit high PIP values). Thus, PIP values across such SNPs sum to the number of total causal variants (i.e., provide an estimate of the number of quantitative trait nucleotides (QTN) contributing to trait variation). In contrast, in the case of suppressed recombination, many SNPs with low (but non-zero) PIPs are expected because different SNPs can readily tag the causal variants (i.e., SNPs carry redundant information). This leads to PIP values summing near one. Also shown are expected patterns for single SNP GWA that does not account for LD. Photographs of representative samples of the four *Timema* species studied here are shown in (c). These are the same types of photos from which color data was collected for GWA mapping.

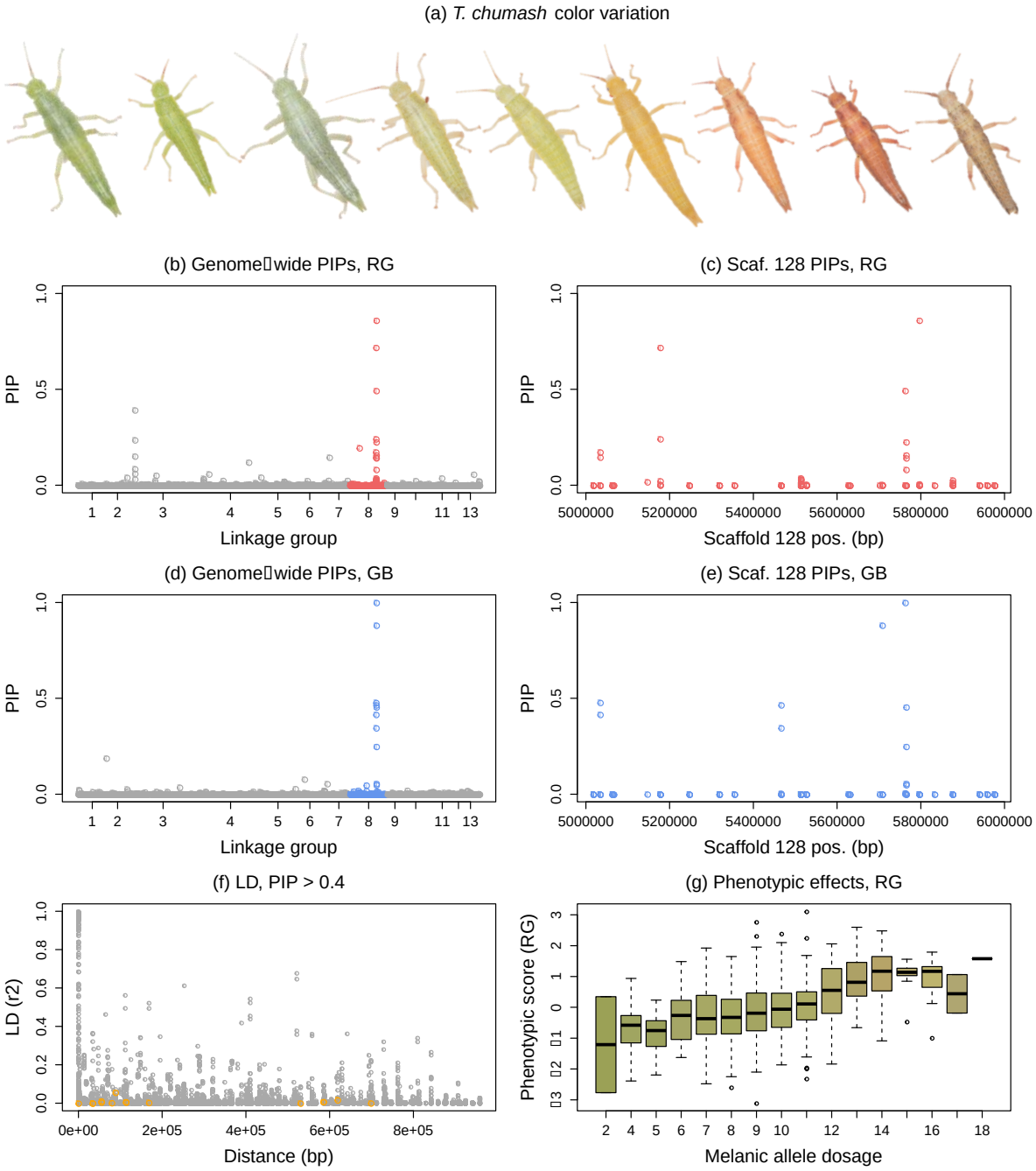


Figure 2. Evidence for multi-genic control of color variation in *T. chumash*. (a) Photographs of representative samples of *T. chumash.* Panels (b) to (e) show results from multi-locus genome wide association (GWA) mapping in GEMMA at two different genomic scales (genome wide and for the scaffold showing the bulk of strong associations, i.e., scaffold 128 on LG8). PIP = posterior inclusion probability (see Fig. 1 for further explanation). Red-green (RG) color variation is shown with red dots and green-blue (GB) color variation by blue dots. pos = position, where bp = base pair. Panel (f) shows linkage disequilibrium (LD, measured as the squared correlation coefficient, r^2) between color-associated single nucleotide polymorphisms (SNPs) in *T. chumash*, shown as a function of the base pair (bp) distances between such SNPs. Orange dots are SNPs with PIPs > 0.4. LD between these

SNPs is low, and no higher than between other SNPs (i.e., those with PIPs < 0.4, grey dots) in this genomic region. Panel (g) depicts phenotypic scores for RG as a function of melanic allele dosage across the ten top color-associated SNPs. The results use rounded melanic dosages (to the nearest integer), with boxplots showing the distribution of phenotypic scores for each melanic allele dosage bin. Here, the color of the boxes is the mean color in hexadecimal code (for further results see Figs. S5-6).

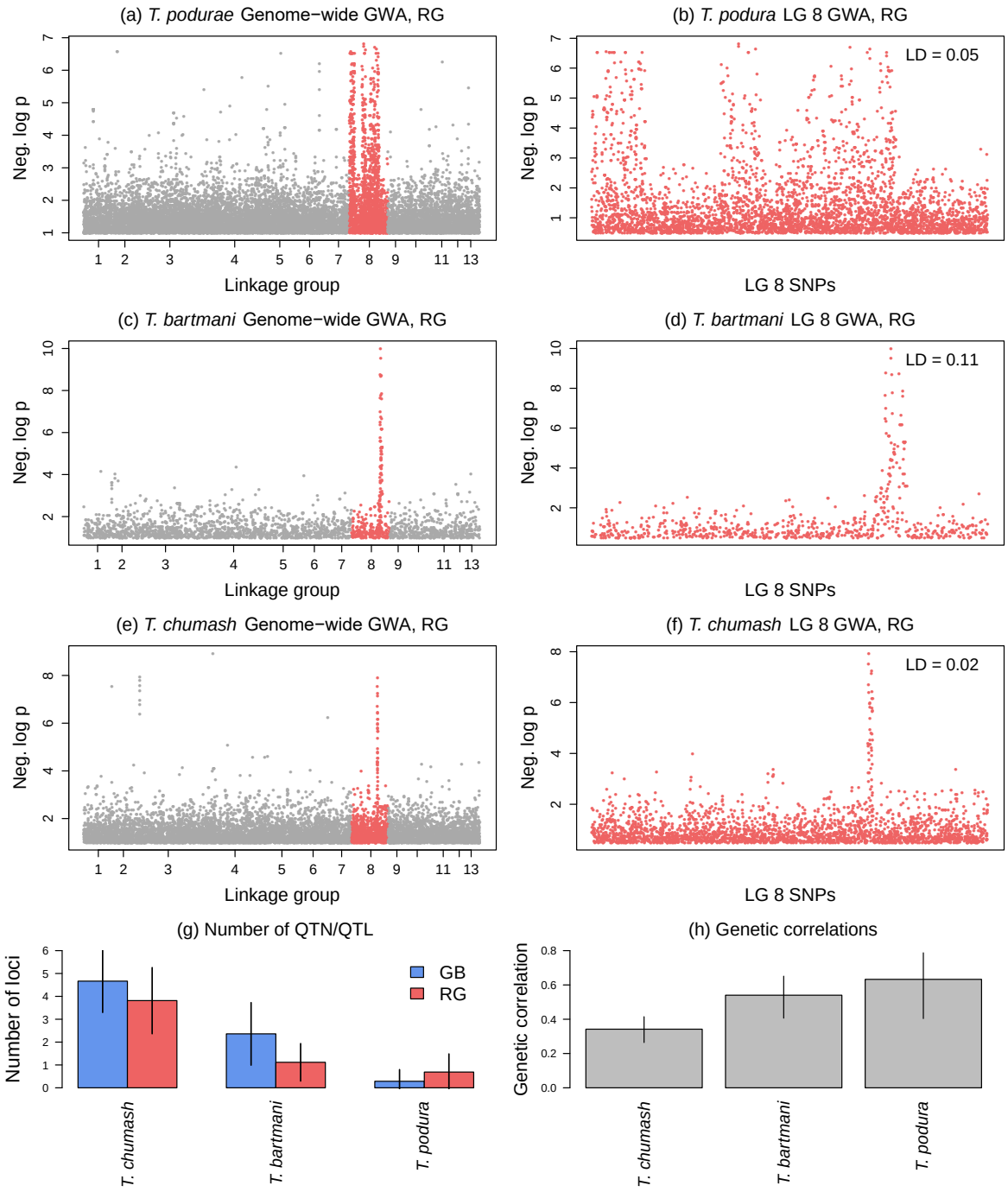


Figure 3. Evidence for variation in genetic architecture of color among *Timema* species, including

major locus control in species other than *T. chumash.* Panels (a)-(d) show results of single-locus genome wide association (GWA) mapping of red-green (RG) color variation in three *Timema* species at two genomic scales (genome wide and for the single linkage group (LG8) showing the bulk of associations), without correction for population structure in GENABLE. For analogous results with correction for population structure, and for the green-blue (GB) trait see Figures S9-10. The y-axis shows the negative $\log_{10}$ *P-value* (Neg. log p) for each test that a single-nucleotide polymorphism (SNP) is associated with color variation. At the scale of LG8, *T. chumash* exhibits a peak of association (which actually represents several distinct peaks when zoomed in further on scaffold 128, see Fig. 2), *T. bartmani* a narrow block of association, and *T. podura* a wide block of association. For details on linkage disequilibrium see Figure S11. Panel (e) shows the number of genetic variants (i.e., quantitative trait nucleotides, QTN) estimated to affect RG and GB in each species. Bars are medians and vertical lines show the 95% ETPIs. Panel (f) shows estimates of the genetic correlation between RG and GB in each species. Bars are Pearson correlation coefficients and vertical lines show the 95% confidence intervals.

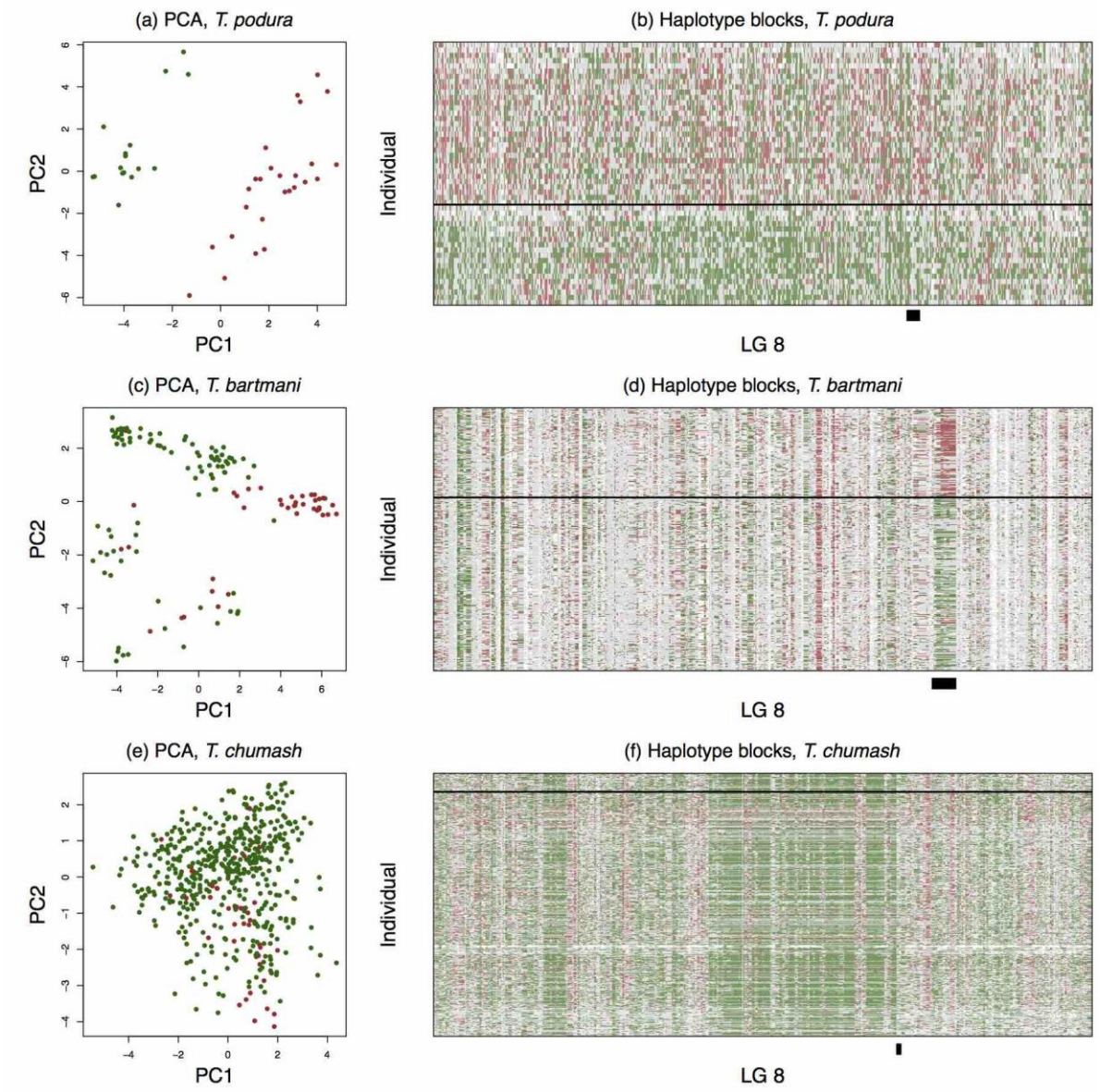


Figure 4. Statistical summary of genetic variation in *T. podura* (a), *T. bartmani* (c) and *T. chumash* (e) based on principal component analysis of SNPs on genome scaffold 128 between 5 and 6 megabase (Mbps). Each point represents an individual stick insect, and points are colored green or brown to denote cluster membership from hierarchical clustering based on RG and GB color scores. Panels (b), (d), and (f) summarize haplotype ancestry/blocks across linkage group (LG) 8 for each of these species. Each row corresponds with an individual and columns denote SNPs along LG 8. Black boxes delineate the 5-6 Mbps pair region of scaffold 128 associated with color in *T. chumash*. Plots are colored to reflect homozygous green chromosomal variants (green), homozygous melanic chromosomal variants (brown), heterozgyous for chromosomal variants (gray) or uncertain (posterior probability of an ancestry/chromosomal variant less than 0.5; white). These assignments are only meaningful within chromosomal variants and for species that have them (i.e., *T. bartmani* and *T. podura*).

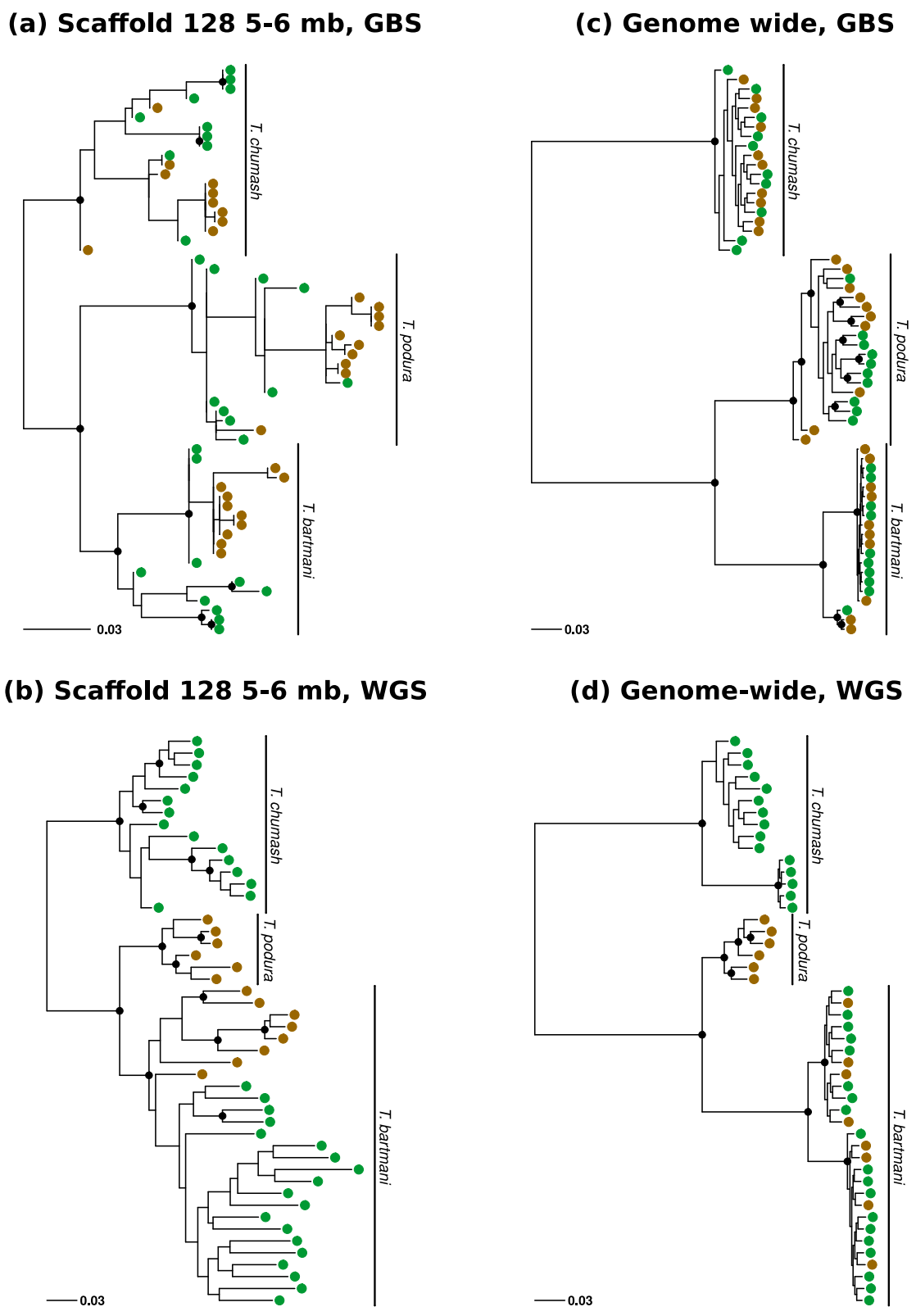


Figure 5. Phylogenetic trees for *T. podura*, *T. bartmani*, and *T. chumash* color morphs. Trees were estimated based on SNPs in the color-associated region of LG 8 (5-6 megabases on scaffold 128) (a,

b) or concatenated SNPs from across the genome (c, d). Trees are shown for SNPs from either GBS data (a, c) or from whole genome sequence data (WGS; b, d). Colored dots denote color morphs (green versus melanic = brown). Trees are rooted with *T. chumash*. Black points on internal nodes denote >70% bootstrap support. In all cases, stick insects group by species rather than by color morph across species (there is evidence of grouping by morph within species in the trees from the scaffold 128, 5-6 megabases SNP data set).

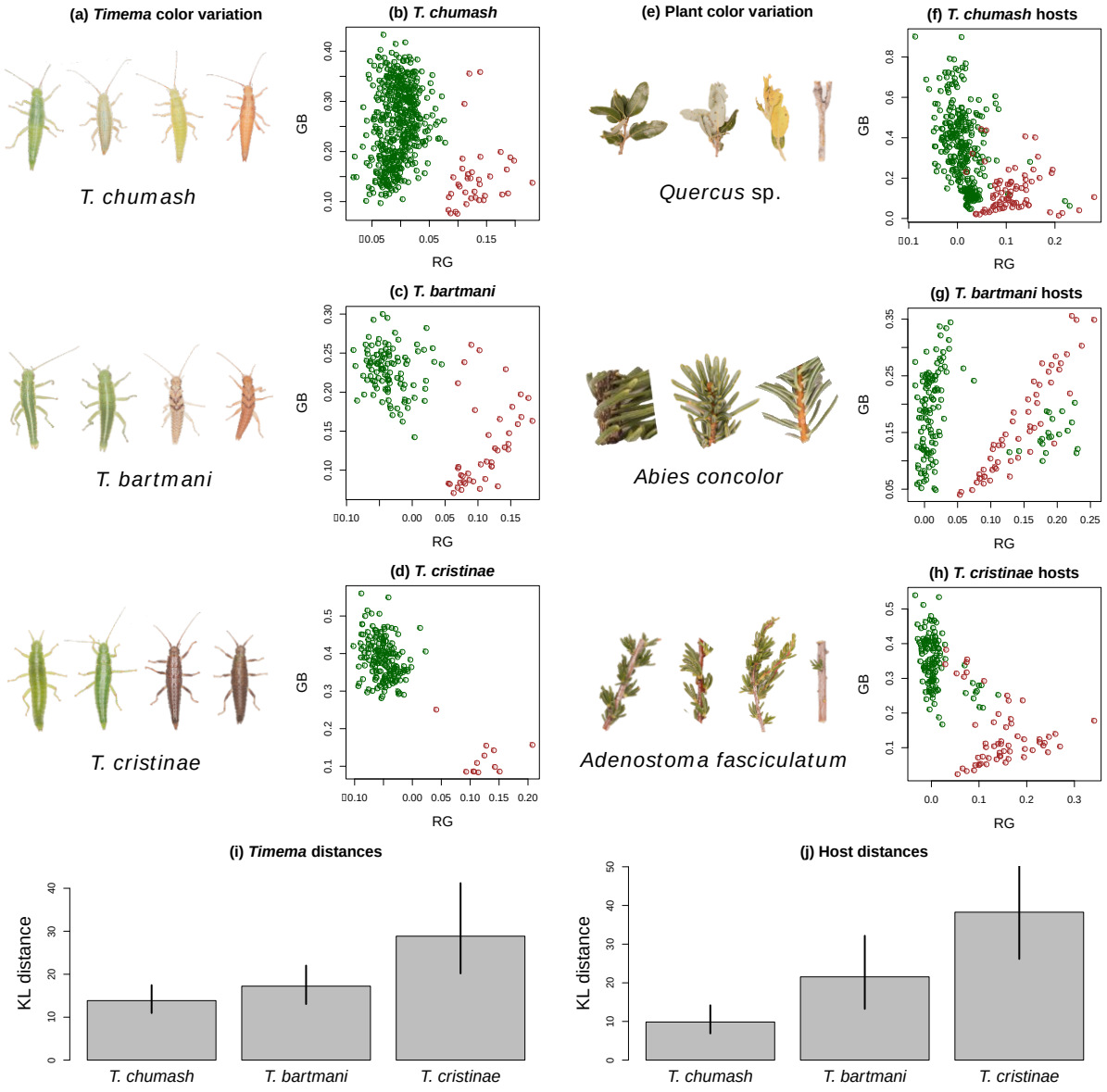


Figure 6. Overlap in coloration between green and melanistic *Timema* morphs, and the leaves versus stems of the host plants they are found upon; for *T. chumash* this is oak (*Quercus sp.*) and mountain mahogany (*Arctostaphylos sp.),* for *T. bartmani* this is white pine (*Pinus flexilis*) and white fir (*Abies concolor*), for *T. cristinae* this is California lilac (*Ceanothus spinosus*) and chamise (*Adenostoma fasciculatum*), and we were unable to obtain data for *T. podura*. Panel (a) shows representative

Timema photos and panels (b) to (d) empirical data, where the green and brown dots are clusters corresponding to green versus melanistic morphs. Panel (e) shows representative photos of the latter host pair listed above for each species. Note that the plant coloration data depicted in panels (f) to (h) is from all the hosts listed above, not just those illustrated in panel (e). For host plants, the green and brown dots are data from leaves versus stems, respectively. The bottom two panels (i) and (j) show the mean Kullback Leibler (KL) distance between morphs and host tissues as bars, with vertical lines representing 95% ETPIs. RG = red-green spectrum, GB = green-blue spectrum.

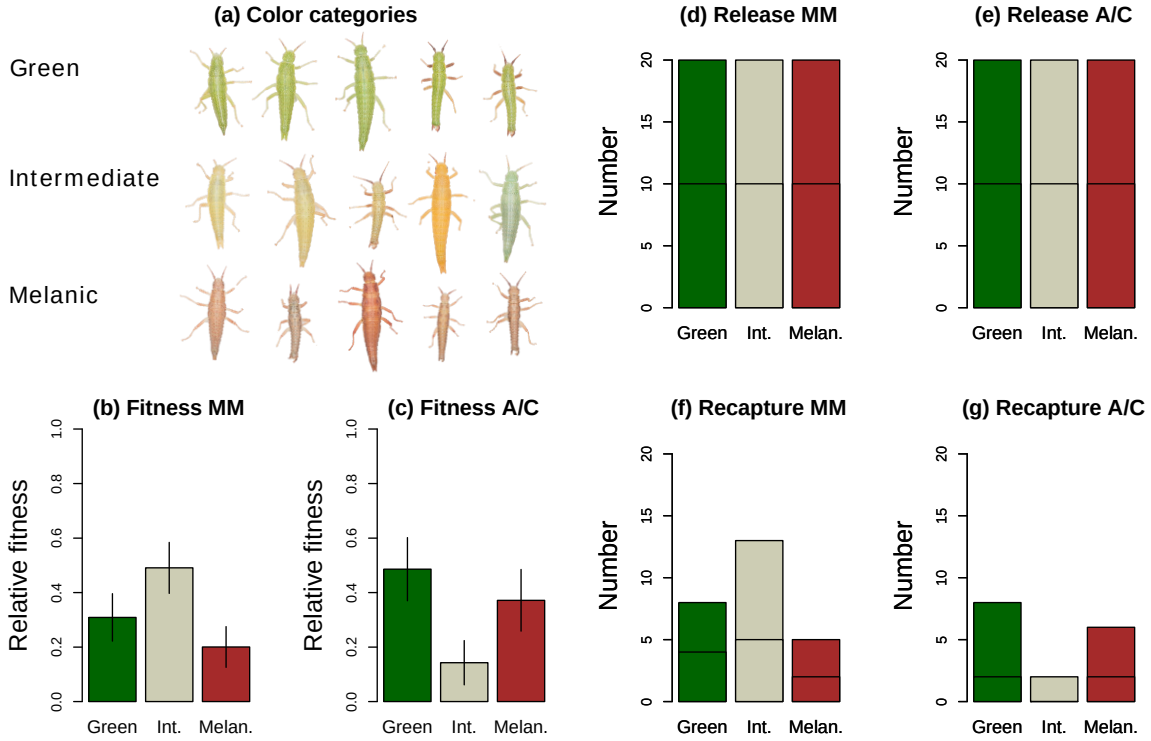


Figure 7. Results of the transplant experiment in *T. chumash* testing for stronger disruptive selection on chamise and California lilac (*Adenostoma* and *Ceanothus*, respectively, abbreviated A/C) than on mountain mahogany (MM). Panel (a) shows representatives of the green, melanistic, and intermediate coloration in *T. chumash*. Panels (b) and (c) show relative fitness in each treatment. Bars are means and standard deviations of the posterior (analogous to standard errors). The raw number of individuals released and recaptured are shown in panels (d) to (g), where the horizontal line in each bar distinguishes numbers of individuals from each of the two experimental blocks.

Literature cited

1. Rockman MV (2012) THE QTN PROGRAM AND THE ALLELES THAT MATTER FOR EVOLUTION: ALL THAT'S GOLD DOES NOT GLITTER. Evolution 66: 1-17.
2. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, et al. (2014) Genomics and the origin of species. Nature Reviews Genetics 15: 176-192.
3. Schwander T, Libbrecht R, Keller L (2014) Supergenes and Complex Phenotypes. Current Biology 24: R288-R294.
4. Ford EB (1971) Ecological Genetics: Chapman and Hall; 3rd Revised edition edition.
5. Reimchen TE (1979) SUBSTRATUM HETEROGENEITY, CRYPSIS, AND COLOR POLYMORPHISM IN AN INTER-TIDAL SNAIL (LITTORINA-MARIAE). Canadian Journal of Zoology 57: 1070-1085.
6. Felsenstein J (1981) Skepticism towards Santa Rosalia, or why are there so few kinds of animals? Evolution 35: 124-138.
7. Merilaita S, Tuomi J, Jormalainen V (1999) Optimization of cryptic coloration in heterogeneous habitats. Biological Journal of the Linnean Society 67: 151-161.
8. Bond AB, Kamil AC (2006) Spatial heterogeneity, predator cognition, and the evolution of color polymorphism in virtual prey. Proceedings of the National Academy of Sciences of the United States of America 103: 3214-3219.
9. Houston AI, Stevens M, Cuthill IC (2007) Animal camouflage: compromise or specialize in a 2 patch-type environment? Behavioral Ecology 18: 769-775.
10. Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. Genetics 173: 419-434.
11. Charlesworth D (2016) The status of supergenes in the 21st century: recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. Evolutionary Applications 9: 74-90.
12. Thompson MJ, Jiggins CD (2014) Supergenes and their role in evolution. Heredity 113: 1-8.
13. Llaurens V, Whibley A, Joron M (2017) Genetic architecture and balancing selection: the life and death of differentiated variants. Molecular Ecology 26: 2430–2448.
14. Lowry DB, Willis JH (2010) A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. Plos Biology 8.
15. Kupper C, Stocks M, Risse JE, dos Remedios N, Farrell LL, et al. (2016) A supergene determines highly divergent male reproductive morphs in the ruff. Nature Genetics 48: 79-+.
16. Lamichhaney S, Fan GY, Widemo F, Gunnarsson U, Thalmann DS, et al. (2016) Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). Nature Genetics 48: 84-+.
17. Nijhout HF (2003) Polymorphic mimicry in Papilio dardanus: mosaic dominance, big effects, and origins. Evolution & Development 5: 579-592.
18. Wellenreuther M, Bernatchez L (2018) Eco-Evolutionary Genomics of Chromosomal Inversions. Trends in Ecology & Evolution 33: 427-440.
19. Sandoval CP (1994) Differential Visual Predation on Morphs of Timema-Cristinae (Phasmatodeae, Timemidae) and Its Consequences for Host-Range. Biological Journal of the Linnean Society 52: 341-356.

20. Sandoval CP, Nosil P (2005) Counteracting selective regimes and host preference evolution in ecotypes of two species of walking-sticks. Evolution 59: 2405-2413.
21. Comeault AA, Flaxman SM, Riesch R, Curran E, Soria-Carrasco V, et al. (2015) Selection on a Genetic Polymorphism Counteracts Ecological Speciation in a Stick Insect. Current Biology 25: 1-7.
22. Sandoval CP (1994) The effects of relative geographical scales of gene flow and selection on morph frequencies in the walking-stick *Timema cristinae*. Evolution 48: 1866-1879.
23. Nosil P, Villoutreix R, de Carvalho CF, Farkas TE, Soria-Carrasco V, et al. (2018) Natural selection and the predictability of evolution in Timema stick insects. Science 359: 765-770.
24. Lindtke D, Lucek K, Soria-Carrasco V, Villoutreix R, Farkas TE, et al. (2017) Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. Molecular Ecology 26: 6189-6205.
25. Tearle RG, Belote JM, McKeown M, Baker BS, Howells AJ (1989) CLONING AND CHARACTERIZATION OF THE SCARLET GENE OF DROSOPHILA-MELANOGASTER. Genetics 122: 595-606.
26. Maan ME, Sefc KM (2013) Colour variation in cichlid fish: Developmental mechanisms, selective pressures and evolutionary consequences. Seminars in Cell & Developmental Biology 24: 516-528.
27. van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, et al. (2016) The industrial melanism mutation in British peppered moths is a transposable element. Nature 534: 102-+.
28. Nadeau NJ, Pardo-Diaz C, Whibley A, Supple MA, Saenko SV, et al. (2016) The gene cortex controls mimicry and crypsis in butterflies and moths. Nature 534: 106-+.
29. Schneider A, Henegar C, Day K, Absher D, Napolitano C, et al. (2015) Recurrent Evolution of Melanism in South American Felids. Plos Genetics 11.
30. Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, et al. (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477: 203-U102.
31. Joron M, Papa R, Beltran M, Chamberlain N, Mavarez J, et al. (2006) A conserved supergene locus controls colour pattern diversity in Heliconius butterflies. Plos Biology 4: 1831-1840.
32. Riesch R, Muschick M, Lindtke D, Villoutreix R, Comeault AA, et al. (2017) Transitions between phases of genomic differentiation during stick-insect speciation. Nature Ecology and Evolution 1: 0082.
33. Comeault AA, Carvalho CF, Dennis S, Soria-Carrasco V, Nosil P (2016) Color phenotypes are under similar genetic control in two distantly related species of Timema stick insect. Evolution 70: 1283-1296.
34. Grant BR, Grant PR (2008) Fission and fusion of Darwin's finches populations. Philosophical Transactions of the Royal Society B-Biological Sciences 363: 2821-2829.
35. Grant PR, Grant BR (2002) Unpredictable evolution in a 30-year study of Darwin's finches. Science 296: 707-711.
36. Reimchen TE (1980) Spine deficiency and polymorphism in a population of *Gasterosteus aculeatus* - an adaptation to predators. Canadian Journal of Zoology-Revue Canadienne De Zoologie 58: 1232-1244.
37. Reimchen TE (1995) Predator-induced cyclical changes in lateral plate frequencies of Gasterosteus. Behaviour 132: 1079-1094.

38. Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, et al. (2010) Widespread genomic divergence during sympatric speciation. Proceedings of the National Academy of Sciences 107: 9724-9729.
39. Chouard T (2010) Evolution: Revenge of the hopeful monster. Nature 463: 864-867.
40. Dietrich MR (2003) Richard Goldschmidt: hopeful monsters and other 'heresies'. Nature Reviews Genetics 4: 68-74.
41. Nosil P, Crespi BJ (2006) Experimental evidence that predation promotes divergence in adaptive radiation. Proceedings of the National Academy of Sciences of the United States of America 103: 9090-9095.
42. Comeault AA, Soria-Carrasco V, Gompert Z, Farkas TE, Buerkle CA, et al. (2014) Genome-Wide Association Mapping of Phenotypic Traits Subject to a Range of Intensities of Natural Selection in Timema cristinae*. American Naturalist 183: 711-727.
43. Abràmoff MD, Magalhães PJ, Ram SJ (2004) Image Processing with ImageJ. Biophotonics International 11: 36–42.
44. Endler JA (2012) A framework for analysing colour pattern geometry: adjacent colours. Biological Journal of the Linnean Society 107: 233-253.
45. Team RDC (2013) R: A Language and Environment for Statistical Computing. Vienna, Austria.
46. Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, et al. (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. Molecular Ecology 21: 2991-3005.
47. Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, et al. (2014) Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. Science 344: 738-742.
48. Nosil P, Gompert Z, Farkas TE, Comeault AA, Feder JL, et al. (2012) Genomic consequences of multiple speciation processes in a stick insect. Proceedings of the Royal Society B: Biological Sciences 279: 5058-5065.
49. Nosil P, Villoutreix R, de Carvalho CF, Farkas TE, Soria-Carrasco V, et al. (2018) Natural selection and the predictability of evolution in Timema stick insects. Science 359: 765-+.
50. Zhou X, Carbonetto P, Stephens M (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models. Plos Genetics 9.
51. Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM (2007) GenABEL: an R library for genorne-wide association analysis. Bioinformatics 23: 1294-1296.
52. Harte D (2017) HiddenMarkov: Hidden Markov Models. R package version 1.8-11. Statistics Research Associates, Wellington. URL: http://www.statsresearch.co.nz/dsh/sslib/.
53. Lucas LK, Nice CC, Gompert Z (2018) Genetic constraints on wing pattern variation in Lycaeides butterflies: A case study on mapping complex, multifaceted traits in structured populations. Molecular Ecology Resources 18: 892-907.
54. Vos PG, Paulo MJ, Voorrips RE, Visser RGF, van Eck HJ, et al. (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. Theoretical and Applied Genetics 130: 123-135.
55. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904-909.

56. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 78: 629-644.
57. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164: 1567-1587.
58. Lewis PO (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50: 913-925.
59. Stamatakis A (2014) RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics 13: 1312-1313.
60. Stamatakis A, Hoover P, Rougemont J (2008) A Rapid Bootstrap Algorithm for the RAxML Web Servers. Systematic Biology 57: 758-771.
61. Pattengale N, Alipour M, Bininda-Emonds O, Moret B, Stamatakis A (2009) How Many Bootstrap Replicates Are Necessary? In: Research in Computational Molecular Biology, Lecture Notes in Computer Science. Springer Berlin / Heidelberg. pp. pp. 184–200.
62. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289-290.
63. Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3: 217-223.
64. Schliep KP (2011) phangorn: phylogenetic analysis in R. Bioinformatics 27: 592-593.
65. Nosil P (2004) Reproductive isolation caused by visual predation on migrants between divergent environments. Proceedings of the Royal Society B-Biological Sciences 271: 1521-1528.
66. Nosil P, Gompert Z, Farkas TE, Comeault AA, Feder JL, et al. (2012) Genomic consequences of multiple speciation processes in a stick insect. Proceedings of the Royal Society B-Biological Sciences 279: 5058-5065.
67. Sandoval C (2000) Persistence of a walking-stick population (Phasmatoptera : Timematodea) after a wildfire. Southwestern Naturalist 45: 123-127.
68. Gompert Z, Comeault AA, Farkas TE, Feder JL, Parchman TL, et al. (2014) Experimental evidence for ecological selection on genome variation in the wild. Ecology Letters 17: 369-379.
69. Plummer M (2016) rjags: Bayesian Graphical Models using MCMC. R package version 4-6. https://CRAN.R-project.org/package=rjags.
70. Ewens WJ (2004) Mathematical population genetics. I. Theoretical introduction. Interdisciplinary applied mathematics, 27.

Supplementary Data for:

Ecology helps explain whether genes for cryptic coloration form a supergene or recombine

Romain Villoutreix, Clarissa F. de Carvalho, Víctor Soria-Carrasco, Dorothea Lindtke, Marisol De-la-Mora, Moritz Muschick, Jeffrey L. Feder, Zach Gompert, and Patrik Nosil

This supplementary material includes
Supplementary methods
Supplementary results
Supplementary tables S2-S10
Supplementary figures S1-S11
Supplementary references

Literature review of the effects of selection and reduced recombination in the evolution of multi-genic adaptation

A literature review of the effects of selection and reduced recombination in the evolution of multi-genic adaptation was conducted. However, it was not represented in this dissertation for simplicity (Table S1 was not included here).

Study design and samples used

Table S2 details whether the data from *Timema* specimens used in this study were newly acquired, or re-analysis of previously published data [56-58].

Table S2. General description of the sampling and study design.

Species	Phenotypic overlap	Spectral reflectance	GWAS and genetic structure	Whole genome phylogenetics
T. cristinae	New data	New and published data	Published data	N/A
T. podura	New data	New data	Published data	New data
T. bartmani	New data	New data	New data	New data
T. chumash	New data	New data	New data	New data

'Published data' refers to data collected in 2013 and reported in references [56-58]. 'New data' is novel to the current study, where in the case of phenotypic measurements all data were collected in a standardized fashion in 2015. For details on individual components, including populations used, sample sizes, etc. see Tables S4, S5, and S7.

Sampling locations

Table S3 contains details of the sampling localities, where samples of *Timema* sp. and plants were collected. Abbreviations of these population codes were used throughout this text.

Table S3. Details about the *Timema* populations used in this study and host plants found in the sites.

Popcode	Host plant	Latitude	Longitude	Altitude	Description
BALD	C, Q	34.22108	-117.668	1172	Mount Baldy
BC	Q	36.06	-121.57	614	Big Creek
BM	WF	33.837	-116.75	2288	Black Mountain
BMCG3	WF, Q	33.83124	-116.741	2261	Black Mountain Camp Ground 3
BMTB	Q	33.83	-116.78	1857	Black Mountain Trailside Boulder
BS	C	33.82	-116.79	1614	Bay Spring
DZR	A	33.86	-116.84	1306	Diamond Zen Ranch
FH	A	34.52	-119.8	742	Far Hill
GR10.43	MM, Q	34.22505	-117.68	1332	Glendora Ridge Mile 10.43
GR8.06	Q, MM	34.22	-117.71	1370	Glendora Ridge Mile 8.06
HF4	C, Q	34.26536	-118.098	1429	Horse Flats 4
HF6	Q	34.26695	-118.117	1262	Horse Flats 6
HFDPD	M,Q	34.34081	-118.016	1819	Horse Flats Daniel Paul Duran
HFRB	Q	34.25822	-118.105	1407	Horse Flats Red Box Picnic Area
HFRS	MM	34.35558	-118.012	1793	Horse Flats Rosenita Saddle
HFTP	C	34.34355	-117.983	1808	Horse Flats Three Points Parking
JL	IC, P, WF, WP	34.16	-116.9	1974	Jenk's Lake
NH	C	34.515	-119.8	825	Near Hill
PCT	WF	33.83944	-116.738	2372	Pacific Coast Trail
SM	Q	SM	37.019	561	Summit Mt. Madonna

Host-plant abbreviations are as follows. A: *Adenostoma fasciculatum*, C: *Ceanothus spinosus*, IC: *Calocedrus decurrens*, M: *Arctostaphylos sp.*, MM: *Cercocarpus sp.*, P: *Pinus sp.*, Q: *Quercus sp.*, WF: *Abies concolor*, WP: *Pinus flexilis*. Popcode. = population code. N-ind= number of individuals used in the study.

Timema sampling

Table S4 presents details about the samples used for acquisition of phenotypic data. Digital photographs of these samples were taken and color variables were measured. See appropriate Materials and Methods section for detailed information.

Table S4. Details of the *Timema* samples used to extract color variables for phenotyping.

Species	Popcode.	Host plant	N-ind	Year
T. bartmani	JL	IC, P, WF, WP	150	2015
T. chumash	GR8.06	Q, MM	541	2015
T. cristinae	FH	A	602	2013*
T. cristinae	FH	A	190	2015
T. podura	BS	C	42	2013*
T. podura	BS	C	4	2015
T. podura	BMTB	Q	6	2015
T. podura	DZR	A	10	2015

Host-plant abbreviations are as follows. A: *Adenostoma fasciculatum,* C: *Ceanothus spinosus,* IC: *Calocedrus decurrens,* MM: *Cercocarpus sp.,* P: *Pinus sp.,* Q: *Quercus sp.,* WF: *Abies concolor,* WP: *Pinus flexilis*. Lat. = latitude. Long. = longitude. Popcode. = population code. N-ind= number of individuals used in the study.Year = year when the samples were collected. *Samples from 2013 refer to the data from [57,58], and individuals collected 2015 are new to this study.

Correlation between phenotypic variables in Timema

Table S5 presents the correlation between the variables RG and GB, estimated for each species. A linear model was used to estimate the coefficient of determination (r^2). See appropriate Materials and Methods section for further information.

Table S5. Correlation between RG and GB values per species.

Species	Popcode	N-ind	Year	r^2	Pearson
T. bartmani	JL	150	2015	0.44	-0.66
T. chumash	GR8.06	541	2015	0.04	-0.15
T. cristinae	FH	602	2013*	0.61	-0.78
T. cristinae	FH	190	2015	0.56	-0.75
T. podura	BS	42	2013*	0.60	-0.78
T. podura	BS, DZR, BMTB	20	2015	0.20	-0.45

Statistics were estimated using R [59]. Popcode = population codes from each species. N-ind= total number of individuals used in the analysis. Year = year when the samples were collected; * values from 2013 samples correspond to already published data [57,58], and are represented here for comparison. r^2 = coefficient of determination. Pearson= Pearson correlation coefficient.

Differentiation and overlap between Timema *morphs*

Figure S1 presents the results of UPGMA algorithm in hclust (from R 3.2.3)[59] to cluster *Timema* into two groups (i.e., morphs) based on the RG and GB color measurements on samples collected in 2015. See appropriate Materials and Methods section for information on clustering method.

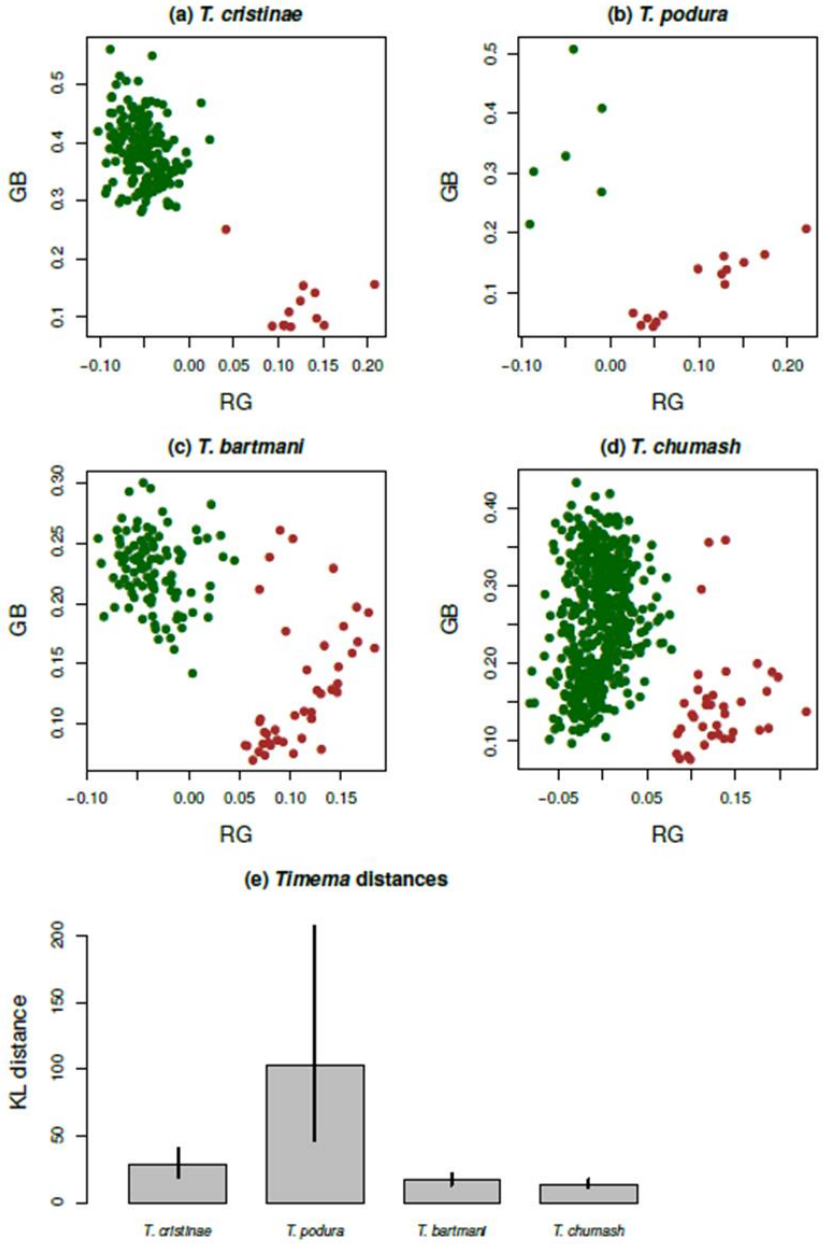


Figure S1. Overlap in coloration between green and melanistic *Timema* morphs (green and brown dots, respectively), for all four polymorphic species studied here.
The bottom panel shows the mean Kullback-Leibler (KL) distance between morphs as bars, with vertical lines representing 95% ETPIs. RG = red-green spectrum, GB = green-blue spectrum.

Spectra reflectance data

We focused the core analyses in the main text on two traits (RG and GB) that represent color variation within the human visible spectrum. This was done under the assumption that these variables accurately represent color variation in *Timema*, and the perception of it by their avian predators. We thus directly tested this assumption, including quantification of ultra-violet reflectance, by recording the spectral reflectance of several individuals from the *Timema* species we studied (Table S6). Our results, described below in detail, largely validate our assumption, justifying the use of RG and GB values quantified from photographs.

We collected spectral data using a USB2000 Fibre Optic Spectrometer (Ocean Optics Inc.) equipped with a 400-micron reflection probe (R400-7-SR) and a pulsed Xenon lamp (PX-2) with an output spectrum of 220-750nm. The spectra were measured at a 45^{o} angle with an integration time of 50 milliseconds, the boxcar width adjusted to 5, and averaging across 20 scans. These measurements then were corrected for nonlinearity, stray light, and electric dark using the OceanView software (Ocean Optics). The reflectance was measured relative to a Spectralon >99% white reflectance standard provided by the manufacturer (WS-1). For each individual, we recorded two reflectance spectra: one from the dorsal anterior part of the body (comprising thorax and head) and the second from the dorsal posterior part (abdomen). We interpolated the raw reflectance in the light spectrum between 300-700nm, corrected the negative values to zero and applied triangular smoothing with a distance of 10 nm using the software AVICOL [60]. Finally, we estimated the mean reflectance as that averaged between the dorsal and abdominal measurements in R.

We assigned individuals to different morphs as described in detail above. We then estimated the mean reflectance across the measured spectrum for each species and morph. As there were only two individuals from *T. cristinae* (both green, Table S6), here we used spectra from 10 individuals from a previous study [58] to generate spectra curves for comparison. The raw spectra was processed using the same procedures cited above. The green morph presents a clear peak of reflectance at medium wavelengths (i.e., between 495–570 nm, green spectra)(Fig. S2). In contrast, the melanistic morph exhibits growing reflectance values towards higher wavelengths in the visible spectrum, being richer from long to middle wavelengths (i.e., brown)[61](Fig. S3). All specimens presented low reflectance at ultraviolet wavelengths (between 300-400 nm), with average reflectance below 6%. In the literature, authors tend to disregard ultraviolet spectra with less than 10% reflectance [61-64]. Thus, *Timema* reflect mainly in the visible spectrum, as do leaves and bark [61], rather than in the ultraviolet spectrum.

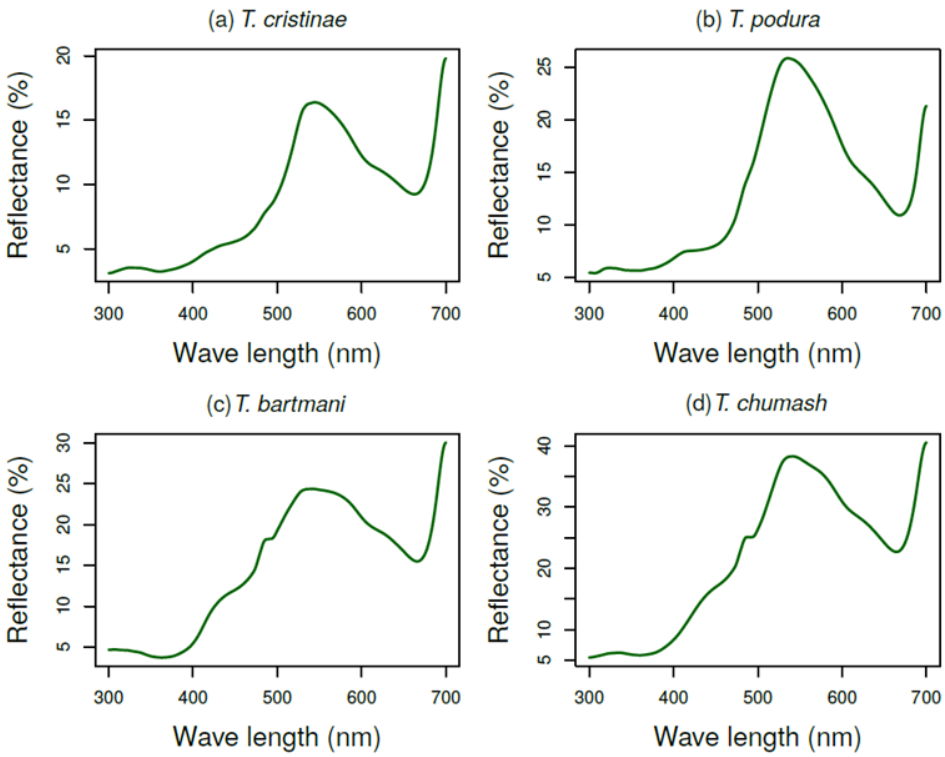


Figure S2. Reflectance curves per species for green morphs.

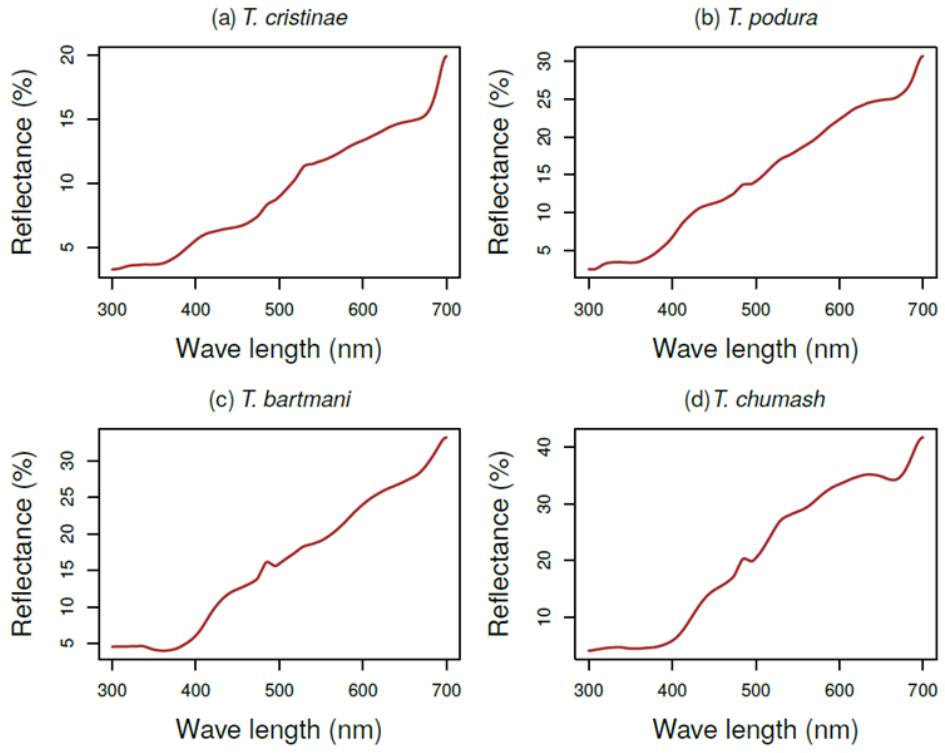


Figure S3. Reflectance curves per species for melanistic morphs.

Aluminous stimuli are apprehended by different photoreceptors based on their different sensitivities (termed 'quantum catch'). Here, the avian visual system was used because birds are a major predator of *Timema*, and likely the main predator of *Timema* during late life-history stages [65]. Birds present four types of photoreceptors: cones that capture long-wave stimuli (LW, red); mediumwave (MW, green); shortwave (SW, blue); and ultraviolet (UV)[66]. Therefore, colors seen by birds are a result of the interaction of three primary colors plus ultraviolet, stimulating different types of cells in the retina.

Quantum catch for the four photoreceptors was calculated using the file provided by Endler & Mielke [67], with average photoreceptors' sensitivities of birds with ultraviolet-sensitive type of photoreceptor (UVS), given most passerines exhibit this system [68,69]. We used the mean reflectance for each nanometer on specimens of the green morph and melanistic individuals, which were analysed separately. The analyses were conducted using the *pavo* R package [70]. The results showed MW is the most stimulated cone when green individuals are the object and LW is the most stimulated for the melanistic morph (Fig. S4). The UV cone was very marginally stimulated in comparison, suggesting that the UV wavelength range does not contribute strongly the the image perceived by birds.

In summary, our results show that *Timema* mostly reflect colors in the visible spectrum, and that this range is what is seen by birds, their predators. There is marginal reflectance in the ultraviolet range, but it is low enough to likely be largely biologically insignificant for the final color perceived. Hence, our work focused on analyses of digital photographs is justified and biologically relevant.

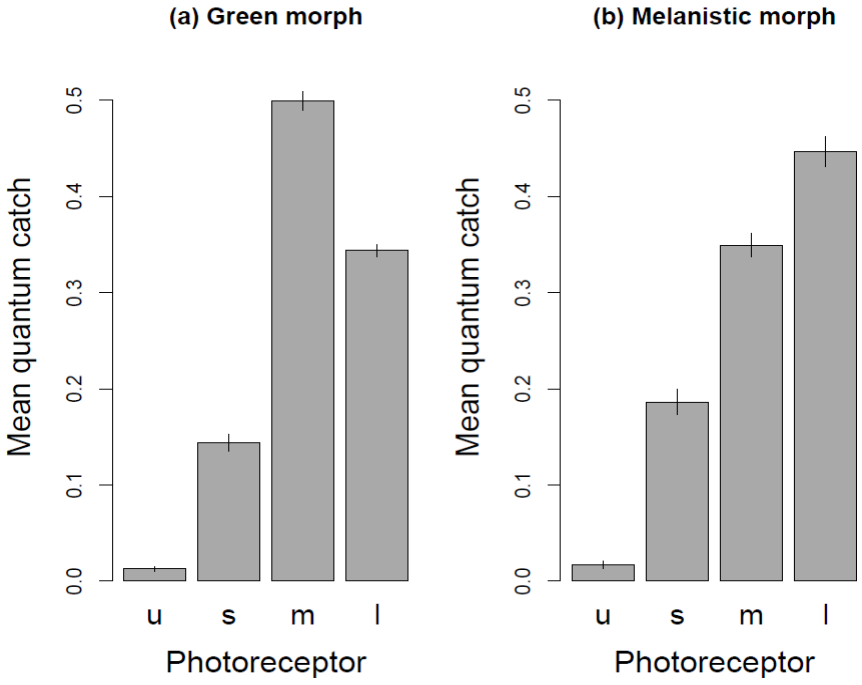


Figure S4. Quantum catch of colors at each photoreceptor of average avian ultraviolet-sensitivity system (UVS)[67]. The different morphs of *Timema* were analysed separately, with their reflectance values averaged at each nanometer. The analysis was conducted at *pavo* R package [70].

u – ultraviolet photoreceptor, *s* – small wavelength photoreceptor (blue), *m* – medium wavelength photoreceptor (green), and *l* – long wavelength photoreceptor (red).

Table S6. Details about specimens used for spectra reflectance measurements per population and per morph.

Species	Location	Host	Year	Morph	N-ind
T. bartmani	JL	WF, WP	2015	green	4
				melanistic	5
T. chumash	GR8.06	MM, Q	2015	green	4
			2015	melanistic	6
T. cristinae	FH	A	2013*	green	5
				melanistic	5
			2015	green	2
T. podura	BMTB	Q	2015	green	2
	BS	C		melanistic	2

Year = year when the measurements were collected. *Spectra information from 2013 refer to the data from [57,58]; and individuals collected 2015 are new to this study.

Genotyping-by-sequencing, alignment, and variant calling.

See appropriate Materials and Methods section for information on DNA extraction, library preparation and sequencing. Following sequencing, we first quality-filtered the reads with a custom Perl script that removed reads with a minimum average phred-scale quality score below 20, trimmed bases with a phred quality score below 20 from the end of the reads, and removed reads with a length of less than 25 bp after trimming. As in previous studies [71,72], we demultiplexed the data using custom Perl scripts that identify and remove the in-line barcodes, including those that were 1 bp away due to synthesizing or sequencing errors, and remove the following six base pairs of the EcoRI cut site and the adapters at the 3' end when present. These scripts then relabel the sequences with the corresponding individual identifiers, and save the reads to separate files for each individual. Sequences lacking barcodes, or those shorter than 16 bp after parsing, were discarded. After these steps, we obtained a total of 1,952,524,371 DNA sequences with an average length of 84 bp (95% equal-tail confidence interval (ETCI)=82-85 bp). The mean number of reads per individual was 895,243 (95% ETCI=389,082-1,896,058), with some differences among species: *T. bartmani*: 622,512 (95% ETCI=331,156-854,020), *T. chumash*: 694,491 (95% ETCI=210,901 1,146,745), *T. cristinae*: 1,365,174 (95% ETCI=617,454-2,140,699), *T. podura*: 2,434,359 (95% ETCI=1,291,065-3,178,544)(see Table S7 for details).

In order to enhance the alignment of reads for species other than *T. cristinae* and, most importantly, avoid discarding in downstream steps variants tagged as multi-allelic when compared to the *T. cristinae* genome (e.g., variants with only two alleles in a population, but both different from the reference allele on the *T. cristinae* genome), a multi-step process was followed to create a consensus reference sequence for each species. This involved the following steps. First, we aligned all reads to the *T. cristinae* reference genome 1.3c2 (NCBI WGS PGFK01000000), with BOWTIE2 version 2.2.9

[73] with the local model and the '--very-sensitive-local' preset (-D 20 -R 3 -N 0 -L 20 -i S,1,0.50). SAMTOOLS version 1.3.1 [74] was used to sort and index alignments. Second, we called variants using SAMTOOLS mpileup and BCFTOOLS call version 1.3.1 using the original consensus caller, excluding all alignments with a phred-scale mapping quality score below 20, and requiring the probability of the data to be less than 0.05 under the null hypothesis that all samples were homozygous for the reference allele to call a variant. Variants with reads for fewer than 25% of the individuals, a quality score of less than 20, or a depth of more than 10 times the number of individuals were excluded. We generated a consensus fasta sequence for each species using BCFTOOLS consensus with the species-specific bcf files, with variants produced in the previous step and the *T. cristinae* reference genome.

Subsequently, we aligned the reads of each species to its specific consensus reference (or the *T. cristinae* genome in the case of *T. cristinae*) and called variants and estimated genotype likelihoods using SAMTOOLS and BCFTOOLS as above. The raw files with all the variants were then subset to include only the individuals phenotyped (i.e., those to be used for downstream GWA analyses, see Tables S4, S7, S8). We then discarded individuals with fewer than 100,000 mapped reads and filtered out variants that had reads for fewer than 50% of the individuals, a quality score below 20, a depth greater than 10 times the number of individuals, more than two alleles, or a minor allele frequency lower than 1%. This generated datasets with a mean number of SNPs of 72,144 (range=19,236-104,955), a mean depth across samples of XXXx (range=XX-XXXx), a mean depth per individual and SNP of 4x (range=3.6-5.4x), and a mean gappiness of 19.1% (range=9.8-25.8%) (Table S8 for details). These numbers are for all SNPs, including those not assigned to one of the 13 linkage groups used for GWA mapping.

We used a hierarchical Bayesian method implemented in the program alleleEst version 0.1 (deposited on bitbucket/Dryad DOI XXX) to co-estimate allele frequencies, genotype probabilities, and genetic diversity from the genotype likelihoods previously inferred with BCFTOOLS [75,76]. This model assumes Hardy-Weinberg and linkage equilibrium and accounts for uncertainty due to low-coverage data and sequencing errors. For each dataset we obtained three independent MCMC chains of 10,000 steps, saving samples every 10^{th} step. We then removed the first 5,000 steps as burnin, concatenated the runs, and estimated mean genotype posterior probabilities from the joint distributions of 1500 samples.

Table S7. Summary of individuals and reads used for building consensus reference sequences.

Species	Population	No samples	No reads	No mapped	Percent mapped
T. bartmani	PCT, BMCG3, JL	735	457,545,991	339,644,687	74.23
T. chumash	BS, HF6, GR8.06, BALD, HFDPD, HFRS, HFRB, HF4, GR10.43, HFTP	794	551,425,570	384,475,844	69.72
T. cristinae *	FH	602	821,834,860	789,388,267	96.05
T. podura	BS	50	121,717,950	86,768,833	71.29

**T. cristinae* is included for reference (*57*).

Table S8. Summary of the genetic data used for GWA analyses.

Species	Pop code	No. ind.	No SNP	Depth SNP	Depth ind	Depth SNP and ind	Gap
T. bartmani	JL	132	19236	481.2 [215.0-1053.0]	70127 [42557-96087]	3.6 [0.0-10.0]	9.8
T. chumash	GR8.06	531	92242	1628.4 [527.0-3688.0]	282868 [93596-465666]	3.1 [0.0-11.0]	25.8
T. podura	BS	42	104955	227.4 [42.0-411.0]	568292 [428989-687106]	5.4 [0.0-17.0]	21.7

The following information is showed: Species, Pop code=population code, No. ind= number of individuals, No. SNP= number of variants, Depth SNP= mean depth per variant across all individuals, Depth ind=mean depth per individual across all variants (i.e. number of aligned reads), Depth SNP and ind=mean depth per variant and individual, and Gap= proportion of missing genotypes (i.e. gappiness). 95% ETCI is represented in brackets for each mean in the depth statistics. Data form *T. bartmani* and *T. chumash* are newly acquired; *T. podura* data were reanalyzed from [57].

Genetic correlations between GB and RG

We estimated genetic correlations between RG and GB color traits for the three newly studied species here, namely *T. bartmani*, *T. chumash*, and *T. podura*. For each species, we first obtained genomic estimated breeding values (GEBVs) from the BSLMM in GEMMA (version 0.94.1)[77]. We did this for the same GBS data sets described above. We obtained posterior distributions of the BSLMM parameters (e.g., PIPs and regression coefficients) by running five MCMC simulations, each with a 1 million step burn-in, 5 million sampling steps and a thinning interval of 100. We then generated GEBVs using the *-predict 1* option, which generates predictions based on the inferred polygenic effects and the SNP-trait associations (i.e., from the PIPs and regression coefficients). Genetic correlations, along with the 95% confidence intervals for these, were then calculated for each species in *R* as the Pearson correlation between RG and GB GEBVs.

Distribution of effect sizes, dominance, and epistasis

We generated two summaries of the effect size distribution for the RG and GB color traits in *T. chumash*. We based our results on the output from the GEMMA analysis. First, we simply plotted the distribution of effect sizes for the 10 SNPs with the highest PIP within the indel region that harbors the top color-associated SNPs (see below for details on the indel). While simple, this approach neglects uncertainty in SNP-color associations. Thus, as a complementary approach, we repeatedly sampled sets of SNPs based on their PIPs and used the sets of SNPs to compute an effect size distribution. This was done based on 1000 vectors of sampled SNPs. Results from both approaches were similar and suggest that multiple SNPs had similar and non-trivial effect sizes. Thus, genetic control of these traits is not dominated by single variants of large effect (Fig. S5). With that said, there was a single variant with approximately twice the effect of any other variant for the GB trait.

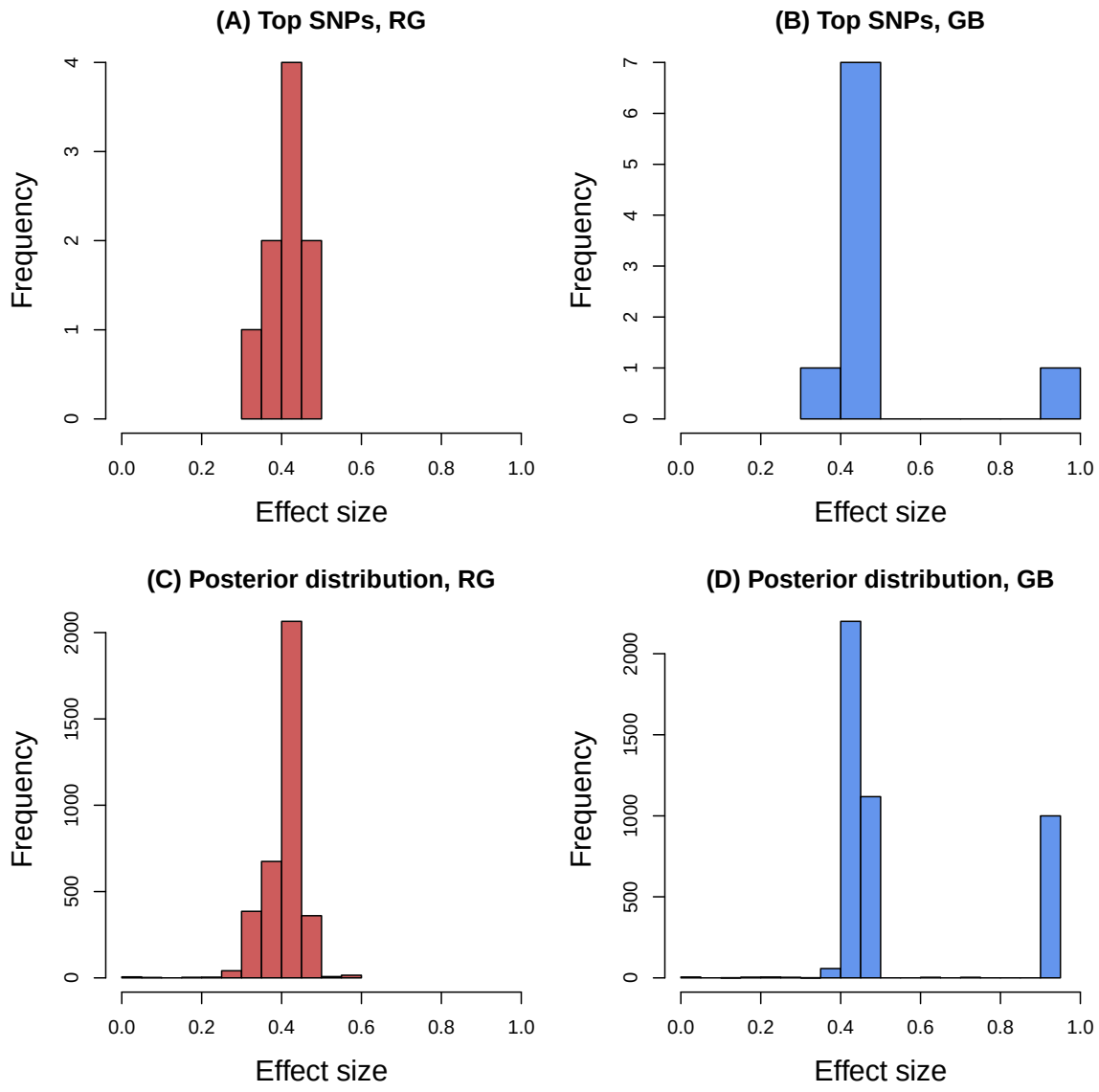


Figure S5. Histograms show the effect size distribution for RG and GB traits based on the GEMMA analyses. Results in A and B are based on the 10 SNPs with the highest PIPs, whereas results in C and D are based on 1000 samples of SNPs based on the PIPs. Absolute values of effect sizes are shown.

We used two approaches to test for epistasis among the SNPs most associated with color in *T. chumash*. We first took a heuristic approach where we examined the relationship between genotype and phenotype (RG and GB color traits) at the SNPs most associated with color. Specifically, we focused on the 10 SNPs from scaffold 128 with the highest PIPs from the GEMMA analysis for RG and GB. For both traits, the top 10 SNPs fell within the indel region discussed below. We used the phenotypic mean of individuals with each genotype at each of the 10 loci to identify the allele with a positive effect on the melanistic coloration (i.e., higher phenotypic scores for RG and lower scores for GB). We then computed the melanic allele dosage for each individual and color trait, which we defined as the sum of the number of melanic alleles across the 10 loci. We did this both based on the

posterior mean genotype at these loci (resulting in a continuous melanic allele dosage score) and with rounded mean genotype scores (resulting in integer valued melanic allele dosage scores). A value of 0 means an individual was homozygous for green alleles at the 10 SNPs and a value of 20 means the individual was homozygous for melanic alleles at the 10 SNPs.

Plots of RG or GB color scores as a function of melanic allele dosage were visually consistent with mostly additive effects across the 10 loci most associated with each trait (Fig. S6). That is, there was a roughly linear increase in color score with an increasing number of melanic alleles (i.e., strong epistasis would cause marked departures from linearity). Quantitatively, linear models of RG or GB color versus melanic allele dosage explained 19.6% and 26.6% of the variation in the color scores, respectively.

As a complementary and more formal approach, we used the program MAPIT [78] to test whether any of the SNPs in the color-associated region (~5-6Mbp on LG8) for either color trait exhibited non-zero marginal epistatic interactions summed across all other SNPs. This approach does not identify specific pairwise or higher order epistatic interactions, but rather tests the null hypothesis that each SNP only exhibits additive effects by testing for a non-zero epistasis variance component (i.e., no epistasis). Thus, there is only one result per SNP. We used the standard model (i.e. without any covariates) with the 199 SNPs in the color-associated region. We used the Hybrid approach for computing *p-values*, which uses an approximate method based on a normal test followed by the Davies exact method for *p-values* below the threshold of 0.05. Thus, we identified 7 SNPs for RG and 9 SNPs for GB with *p-values* < 0.05. However, when we applied Bonferroni correction to account for multiple comparisons (i.e. using a *p-value* threshold below 2.5×10^{-4}), we did not find any evidence of significant epistasis. Collectively, the results are consistent with a fairly additive, polygenic model, where dominance and epistasis do not strongly contribute to either color trait.

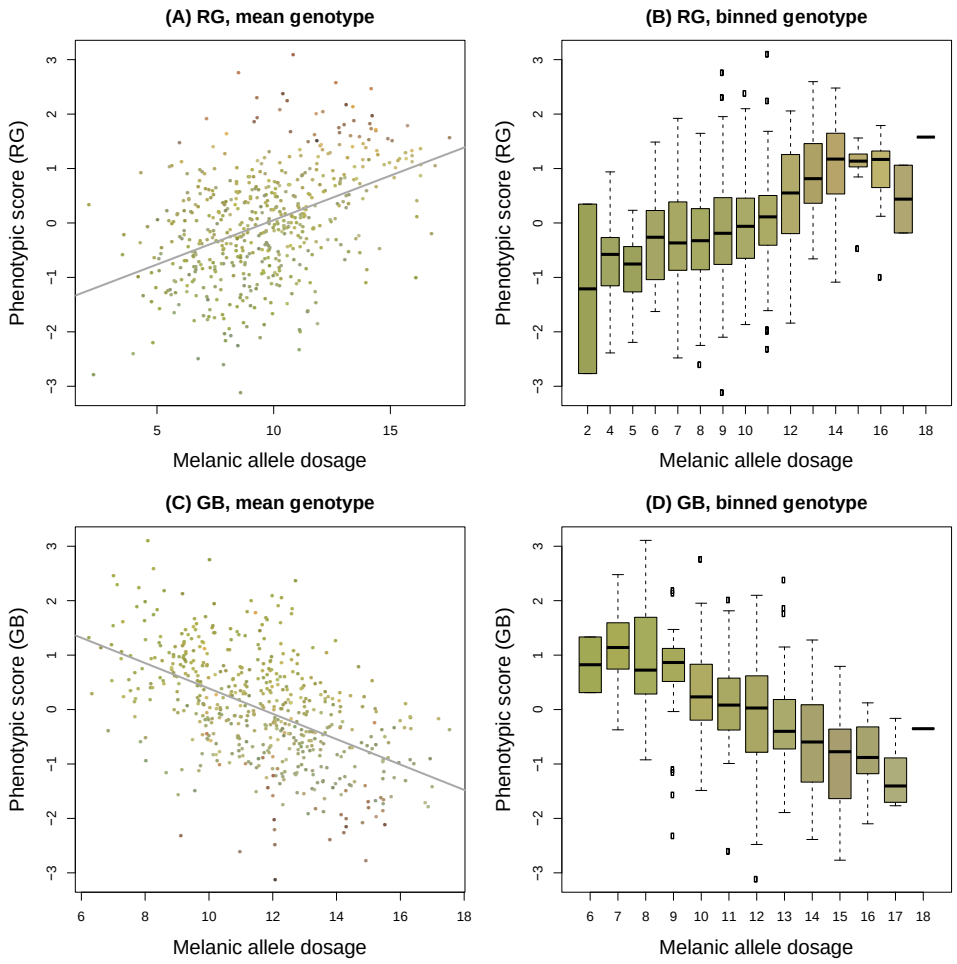


Figure S6. Plots depict phenotypic scores for RG (A, B) and GB (C, D) as a function of melanic allele dosage. Panels A and B show scatterplots with melanic allele dosage as a continuous metric based on the posterior mean genotype estimates. Colored points denote individuals and are colored based on the observed colors of individuals converted to hexadecimal code. Solid lines are from linear regression models (RG, β melanic_dosage = 0.16, $P < 0.0001$, $r^2 = 0.196$; GB, β melanic_dosage = -0.23, $P < 0.0001$, $r^2 = 0.267$). Panels C and D used rounded melanic dosages (to the nearest integer) with boxplots showing the distribution of phenotypic scores for each melanic allele dosage bin. Here, the color of the boxes is the mean of the real color also in hexadecimal code**.**

Functional annotation

We performed structural annotation with Braker1 version 1.9 [79], a pipeline for unsupervised genome annotation that only requires RNA-Seq data aligned to a genome assembly. Braker1 uses GeneMark-ET version 4.32 [80] to generate *ab initio* gene predictions from unsupervised training using RNA-Seq data, which are then used by AUGUSTUS version 3.2.2 [81] along with RNA-Seq reads to generate final, more accurate predictions. Repetitive and low complexity regions can cause the prediction of false positive gene structures. Thus, prior to annotating genes, we annotated and masked the genome for repeats and Transposable Elements (TEs). We built a *de novo* library of repeats using RepeatModeler version 1.0.8 [82] with the *T. cristinae* genome draft 1.3c2 and

combined it with a curated library of TEs developed from the previous genome draft 0.1 [83]. We used vsearch version 2.3.0 [84] to merge both libraries, clustering sequences after sorting by length (--cluster_fast), searching both strands (--strand both), rejecting clusters when identity was below 0.8 (--id 0.8), and using the identity definition of CD-HIT (--iddef 0).

Next, we soft-masked the *T. cristinae* genome draft 1.3c2 with RepeatMasker version 4.0.6 [85] using the slow search with the NCBI search engine ('-xsmall -s -e ncbi'). Subsequently, we aligned the 454 RNA reads from [86] to the masked genome using STAR version 020201 [87] with the basic 2-pass mapping, mapping all reads in the first step, and discarding alignments with a ratio of mismatches greater than the 5% of the mapped length (--twopassMode Basic --twopass1readsN -1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.05). Finally, we used Braker1 to generate gene predictions from the soft-masked genome *T cristinae* genome draft 1.3b2 and the RNA-Seq alignments. This resulted in 57,039 predicted genes (including 790 splice forms) and 164,290 coding DNA sequences (CDS). The mean quality score was 0.61 (95% interval = 0.1-1.0), and 35,315 genes had a quality score equal or greater than 0.5 (61.9%).

Functional annotation was carried out with InterProScan version 5.20-59.0 [88]. We scanned Braker1 predicted proteins against 15 signature databases: CDD 3.14, Coils 2.2.1, Gene3D 3.5.0, HAMAP 201605.11, PANTHER 10.0, Pfam 30.0, PRINTS 42.0, PIRSF 3.01, ProDom 2006.1, ProSiteProfiles 20.119, ProSitePatterns 20.119, SFLD 1, SMART 7.1, SUPERFAMILY 1.75, and TIGRFAM 15.0. The scan of the 57,039 proteins yielded 188,374 hits distributed as follows: 7,754 for CDD, 7,248 for Coils, 24,668 for Gene3D, 260 for HAMAP, 42,355 for PANTHER, 25,745 for Pfam, 11,991 for PRINTS, 494 for PIRSF, 158 for ProDom, 17,227 for ProSiteProfiles, 7960 for ProSitePatterns, 60 for SFLD, 18084 for SMART, 23,689 for SUPERFAMILY, and 681 for TIGRFAM. We found 25,529 predicted proteins had at least one match to any of the databases, 20,340 had matches with at least one InterPro accession, and 14,008 had matches with at least one Gene Ontology (GO) term associated. We found 2218 unique GO terms: 938 for biological process, 315 for cellular component, and 965 for molecular function.

All annotations have been deposited in Dryad DOI:XXX and are also available from the Nosil Lab website (http://nosil-lab.group.shef.ac.uk). Functional info for the genes found in region harboring color-associated loci in *T. chumash* is provided in Table S9.

Table S9. Genes in the color-associated region on scaffold 128, and predicted functions with expanded text.

Start	End	Strand	Attributes	Expanded function
4976413	4980950	-	ID=g6221;	
4987216	4987551	+	ID=g6222;	
4987677	4988048	+	ID=g6223;	
5016443	5018622	-	ID=g6224;	
5114755	5115141	+	ID=g6225;	
5116798	5119312	-	ID=g6226;	
5121576	5121953	+	ID=g6227;	
5122679	5122975	+	ID=g6228;	
5126950	5136425	-	ID=g6229; Dbxref=Gene3D:G3DSA:3.80.10.10, InterPro:IPR000483, InterPro:IPR001611, InterPro:IPR003591, InterPro:IPR026906, InterPro:IPR032675, PANTHER:PTHR24373, Pfam:PF13306, Pfam:PF13855, Prosite:PS51450, SMART:SM00082, SMART:SM00365, SMART:SM00369, Superfamily:SSF52058; Ontology_term=GO:0005515;	Leucine-rich repeat domain superfamily; Cysteine-rich flanking region, C-terminal; Leucine-rich repeat; Leucine-rich repeat, typical subtype; Leucine rich repeat 5; Leucine-rich repeat domain superfamily; Slit related leucine-rich repeat neuronal protein; Leucine rich repeats; Leucine-rich repeat; Leucine rich repeat profile; Leucine rich repeat C-terminal domain; Leucine-rich repeat, SDS22-like subfamily; Leucine-rich repeats, typical (most populated) subfamily; protein binding
5152428	5157256	-	ID=g6230; Dbxref=Coils:Coil;	coiled-coil conformation;
5157774	5158319	-	ID=g6231; Dbxref=InterPro:IPR006111, PANTHER:PTHR10773, PANTHER:PTHR10773:SF13; Ontology_term=GO:0003677, GO:0003899, GO:0006351;	Archaeal RpoK/eukaryotic RPB6 RNA polymerase subunit; DNA-Directed RNA polymerases I, II, and III subunit RPABC2; DNA binding; DNA-directed 5'-3' RNA polymerase activity; transcription, DNA-templated;
5184661	5185176	-	ID=g6232; Dbxref=PANTHER:PTHR24559, PANTHER:PTHR24559:SF174, Superfamily:SSF56672;	DNA/RNA polymerases;
5185658	5185972	-	ID=g6233; Dbxref=Gene3D:G3DSA:3.10.10.10, PANTHER:PTHR10178, PANTHER:PTHR10178:SF302, Superfamily:SSF56672;	DNA/RNA polymerases; DNA/RNA polymerases superfamily;
5197448	5197771	+	ID=g6234;	
5225248	5225613	+	ID=g6235;	
5243884	5245011	+	ID=g6236; Dbxref=Coils:Coil;	coiled-coil conformation;
5250750	5251898	+	ID=g6237; Dbxref=Coils:Coil;	coiled-coil conformation;
5276588	5276926	+	ID=g6238;	
5293645	5303045	-	ID=g6239; Dbxref=InterPro:IPR013525, PANTHER:PTHR19241, PANTHER:PTHR19241:SF305, Pfam:PF01061; Ontology_term=GO:0016020;	ABC-2 type transporter; ABC transporter-like; ATP-binding cassette transporter; membrane;
5324713	5325243	-	ID=g6240; Dbxref=PANTHER:PTHR10492;	Uncharacterized;
5333918	5334813	+	ID=g6241;	
5358947	5359237	+	ID=g6242;	
5359360	5362923	-	ID=g6243;	

5381021	5400431	-	ID=g6244; Dbxref=Gene3D:G3DSA:3.40.50.300, InterPro:IPR003439, InterPro:IPR003593, InterPro:IPR017871, InterPro:IPR027417, PANTHER:PTHR19241, PANTHER:PTHR19241:SF305, Pfam:PF00005, Prosite:PS00211, Prosite:PS50893, SMART:SM00382, Superfamily:SSF52540; Ontology_term=GO:0005524, GO:0016887;	P-loop containing nucleotide triphosphate hydrolases; ABC transporter-like; AAA+ ATPase domain; ABC transporter, conserved site; P-loop containing nucleoside triphosphate hydrolase; ABC transporter, G1; ABC transporter; ABC transporters family signature; ATP-binding cassette, ABC transporter-type domain profile; AAA - ATPases associated with a variety of cellular activities; P-loop containing nucleoside triphosphate hydrolases superfamily; ATP binding; ATPase activity;
5407757	5414160	-	ID=g6245;	
5416724	5417032	+	ID=g6246;	
5465238	5465549	+	ID=g6247;	
5465758	5466078	+	ID=g6248;	
5489905	5490357	-	ID=g6249;	
5518960	5520639	+	ID=g6250;	
5534243	5537814	-	ID=g6251; Dbxref=PANTHER:PTHR11697, PANTHER:PTHR11697:SF102;	General transcription factor 2-related zinc finger protein;
5555794	5558984	-	ID=g6252;	
5568346	5580271	-	ID=g6253; Dbxref=InterPro:IPR013525, PANTHER:PTHR19241, PANTHER:PTHR19241:SF305, Pfam:PF01061; Ontology_term=GO:0016020;	PiggyBac transposable element-derived protein; ATP-Binding cassette transporter; ABC-2 type transporter; membrane;
5580783	5582827	-	ID=g6254; Dbxref=InterPro:IPR029526, PANTHER:PTHR28576, Pfam:PF13843;	PiggyBac transposable element-derived protein; Transposase IS4;
5584921	5586256	+	ID=g6255; Dbxref=Gene3D:G3DSA:3.40.50.300, InterPro:IPR027417, PANTHER:PTHR19241, PANTHER:PTHR19241:SF305, Superfamily:SSF52540;	P-loop containing nucleotide triphosphate hydrolases; P-loop containing nucleoside triphosphate hydrolase; ATP-Binding cassette transporter; P-loop containing nucleoside triphosphate hydrolase;
5588391	5591849	-	ID=g6256; Dbxref=Gene3D:G3DSA:3.40.50.300, InterPro:IPR003439, InterPro:IPR027417, PANTHER:PTHR19241, PANTHER:PTHR19241:SF305, Pfam:PF00005, Superfamily:SSF52540; Ontology_term=GO:0005524, GO:0016887;	P-loop containing nucleotide triphosphate hydrolases; ABC transporter-like; P-loop containing nucleoside triphosphate hydrolase; ATP-Binding cassette transporter; ABC transporter; P-loop containing nucleoside triphosphate hydrolases superfamily; ATP binding; ATPase activity;
5601973	5603314	+	ID=g6257; Dbxref=InterPro:IPR025476, PANTHER:PTHR10492, Pfam:PF14214;	Helitron helicase-like domain; Helitron helicase-like domain at N-terminus;
5603705	5603995	+	ID=g6258; Dbxref=Gene3D:G3DSA:3.40.50.300, InterPro:IPR010285, InterPro:IPR027417, PANTHER:PTHR10492, Pfam:PF05970, Superfamily:SSF52540; Ontology_term=GO:0000723, GO:0003678, GO:0006281;	P-loop containing nucleotide triphosphate hydrolases; DNA helicase Pif1-like; P-loop containing nucleoside triphosphate hydrolase; PIF1-like helicase; P-loop containing nucleoside triphosphate hydrolases superfamily; telomere maintenance; DNA helicase activity; DNA repair;

5610883	5611272	+	ID=g6259;	
5611885	5612193	+	ID=g6260;	
5614232	5618783	+	ID=g6261; Dbxref=PANTHER:PTHR19446;	Reverse transcriptases;
5619302	5620456	+	ID=g6262;	
5620969	5621256	-	ID=g6263; Dbxref=InterPro:IPR000477, Pfam:PF00078, Prosite:PS50878;	Reverse transcriptase domain; Reverse transcriptase (RNA-dependent DNA polymerase); Reverse transcriptase (RT) catalytic domain;
5622187	5622810	+	ID=g6264;	
5622898	5623449	-	ID=g6265;	
5623511	5628401	-	ID=g6266;	
5640809	5642490	-	ID=g6267;	
5656256	5657291	-	ID=g6268;	
5658190	5667389	-	ID=g6269;	
5689220	5716135	-	ID=g6270; Dbxref=Gene3D:G3DSA:1.10.1070.11, Gene3D:G3DSA:1.25.40.70, Gene3D:G3DSA:2.60.40.150, Gene3D:G3DSA:3.10.20.90, Gene3D:G3DSA:3.30.1010.10, InterPro:IPR000008, InterPro:IPR000341, InterPro:IPR000403, InterPro:IPR001263, InterPro:IPR002420, InterPro:IPR011009, InterPro:IPR015433, InterPro:IPR016024, InterPro:IPR018936, InterPro:IPR029071, PANTHER:PTHR10048, PANTHER:PTHR10048:SF14, Pfam:PF00454, Pfam:PF00613, Pfam:PF00792, Pfam:PF00794, Prosite:PS00915, Prosite:PS00916, Prosite:PS50290, Prosite:PS51545, Prosite:PS51546, Prosite:PS51547, SMART:SM00142, SMART:SM00145, SMART:SM00146, Superfamily:SSF48371, Superfamily:SSF49562, Superfamily:SSF54236, Superfamily:SSF56112; Ontology_term=GO:0005488, GO:0005515, GO:0016301, GO:0046854, GO:0048015;	Phosphatidylinositol 3-/4-kinase, catalytic domain superfamily; C2 domain superfamily; Phosphatidylinositol 3-kinase Catalytic Subunit; Chain A, domain 1; Phosphatidylinositol 3-kinase Catalytic Subunit; Chain A, domain 4; C2 domain; Phosphatidylinositol 3-kinase Ras-binding (PI3K RBD) domain; Phosphatidylinositol 3-/4-kinase, catalytic domain; Phosphoinositide 3-kinase, accessory (PIK) domain; Phosphatidylinositol 3-kinase, C2 domain; Protein kinase-like domain superfamily; Phosphatidylinositol kinase; Armadillo-type fold; Phosphatidylinositol 3/4-kinase, conserved site; Ubiquitin-like domain superfamily; Phosphatidynositol kinase; Phosphatidynositol 3-kinase 1; Phosphatidylinositol 3- and 4-kinase; Phosphoinositide 3-kinase family, accessory domain (PIK domain); Phosphoinositide 3-kinase C2; PI3-kinase family, ras-binding domain; Phosphatidylinositol 3- and 4-kinases signature1; Phosphatidylinositol 3- and 4-kinases signature 2; Phosphatidylinositol 3- and 4-kinases family; PIK helical domain; Phosphatidylinositol 3-kinase Ras-binding (PI3K RBD) domain; Phosphatidylinositol 3-kinase C2 (PI3K C2) domain; Phosphatidylinositol 3-kinase, C2 domain; Phosphoinositide 3-kinase, accessory (PIK) domain; Phosphatidylinositol 3-/4-kinase, catalytic domain; Armadillo-type fold; C2 domain (Calcium/lipid-binding domain, CaLB) superfamily; Ubiquitin-like domain superfamily; Protein kinase-like domain superfamily; binding; protein binding; kinase activity; Phosphatidylinositol phosphorylation; Phosphatidylinositol-mediated signaling;
5721107	5721613	+	ID=g6271;	

5725166	5735216	-	ID=g6272; Dbxref=Gene3D:G3DSA:3.10.20.90, InterPro:IPR003113, InterPro:IPR015433, InterPro:IPR029071, PANTHER:PTHR10048, PANTHER:PTHR10048:SF14, Pfam:PF02192, Prosite:PS51544, SMART:SM00143, Superfamily:SSF54236; Ontology_term=GO:0046854, GO:0048015;	Phosphatidylinositol 3-kinase Catalytic Subunit; Chain A, domain 1; Phosphatidylinositol 3-kinase adaptor-binding (PI3K ABD) domain; Phosphatidylinositol kinase; Ubiquitin-like domain superfamily; Phos[hatidynositol kinase; Phosphatidylinositol 3-kinase 1; PI3-kinase family, p85-binding domain; Phosphatidylinositol 3-kinase adaptor-binding (PI3K ABD) domain; Phosphatidylinositol 3-kinase adaptor-binding (PI3K ABD) domain; Ubiquitin-like domain superfamily; phosphatidylinositol phosphorylation; phosphatidylinositol-mediated signaling;
5743180	5743479	-	ID=g6273;	
5747533	5749772	+	ID=g6274;	
5757211	5780398	+	ID=g6275; Dbxref=Gene3D:G3DSA:1.25.10.10, Gene3D:G3DSA:3.10.20.90, InterPro:IPR001012, InterPro:IPR011989, InterPro:IPR018997, InterPro:IPR029071, PANTHER:PTHR23153, PANTHER:PTHR23153:SF38, Pfam:PF00789, Pfam:PF09409, Prosite:PS50033, SMART:SM00166, SMART:SM00580, Superfamily:SSF143503, Superfamily:SSF54236; Ontology_term=GO:0005515;	Leucine-rich Repeat Variant; Phosphatidylinositol 3-kinase Catalytic Subunit; Chain A, domain 1; UBX domain; Armadillo-like helical; PUB domain; Ubiquitin-like domain superfamily; UBX domain; PUB; UBX ; UBX; PUG; PUG domain-like superfamily; Ubiquitin-like domain superfamily; protein binding;
5783737	5805427	-	ID=g6276; Dbxref=InterPro:IPR008806, InterPro:IPR013197, PANTHER:PTHR12949, Pfam:PF05645, Pfam:PF08221; Ontology_term=GO:0003677, GO:0003899, GO:0006351;	RNA polymerase III Rpc82, C -terminal; RNA polymerase III subunit RPC82-related, helix-turn-helix; DNA binding; DNA-directed 5'-3' RNA polymerase activity; transcription, DNA-templated;
5818284	5823210	+	ID=g6277; Dbxref=InterPro:IPR004307, PANTHER:PTHR10057, Pfam:PF03073; Ontology_term=GO:0016021;	TspO/MBR-related protein; integral component of membrane;
5849032	5849349	+	ID=g6278; Dbxref=InterPro:IPR009057, Superfamily:SSF46689; Ontology_term=GO:0003677;	Homeobox-like domain superfamily; Homeodomain-like superfamily; DNA binding;
5882354	5882695	-	ID=g6279;	
5908379	5909206	+	ID=g6280;	
5999700	6008081	+	ID=g6281;	
6014155	6014883	+	ID=g6282;	
6038833	6039150	-	ID=g6283;	
6094553	6095044	-	ID=g6284;	
6144534	6145832	+	ID=g6285;	
6146817	6147110	-	ID=g6286;	
6171844	6172125	-	ID=g6287;	
6183805	6184326	-	ID=g6288;	
6184413	6185135	-	ID=g6289;	

Detection of putative chromosomal inversion in T. cristinae.

The following paragraphs describe the selection of homozygous individuals and detailed methods for inversion break points detection. See appropriate Materials and Methods section for further details on the detection of chromosomal inversion in *T. cristinae*.
Specifically, as in [56] we used PCA and K-means clustering to assign color genotypes to each individual. PCA was performed on the centered and standardized genotype matrix for each species; we standardized genotypes by dividing by $\sqrt{[pi\ (1\text{-}pi)]}$, where *pi* is a Bayesian estimate of the allele frequency given a binomial sampling distribution and a beta prior with α and β equal to 1. We next used k-means clustering to group individuals based on their scores for the first two PCs. We assumed six clusters (which assumes three haplotypes/alleles found in all homozygous and heterozygous combinations) and used the Hartigan-Wong algorithm with 100 starts and a maximum of 100 iterations for clustering [89]. This was done with the *kmeans* function in the *R MASS* package (version 7.3.47)[90].

We then divided scaffolds 702.1 and 128 into non-overlapping 10 kilo-base (kb) windows. We took sets of seven 10 kb windows at a time, and for each group (i.e., melanistic homozygotes and green homozygotes) we calculated the mean LD between all the SNPs in the first three and last three windows in the set of seven (i.e., two 30 kb windows separated by a 10 kb window). We measured LD as the coefficient of determination calculated from the genotype estimates. Again, we would expect mean LD to be lower in the green homozygotes if the breakpoint occurred within the set of seven 10 kb windows, and particularly so if it was in the middle 10 kb window. To capture this, we calculated what we refer to as the standardized difference in LD between melanistic and green homozygotes at $\Delta LD_i = (LD^{mel}_i - LD^{green}_i)/(LD^{mel}_i + LD^{green}_i)$, where LD^{mel}_i and LD^{green}_i are the mean LD for window set *i* for melanistic and green homozygotes, respectively. Note that this metric is analogous to a signed coefficient of variation in LD between groups. We then computed this statistic in sliding sets of seven 10 kb windows with 10 kb window shifts.

We fit a HMM to the ΔLDi metrics using the R package *HiddenMarkov,* version 1.8.11 [91] to fit the models, but modified the *Mstep* function to allow for these fixed parameter values. Doing so allowed us to focus on hidden states of interest for detecting the inversion. We assumed a Gaussian error distribution with means set to the empirical mean of the ΔLDi vector ('normal' state) and to the 95th quantile of the ΔLDi distribution ('high' or breakpoint state). Standard deviations for both states were set at 80% of the empirical standard deviation. We used the Baum-Welch algorithm with 500 iterations and a tolerance of 0.0001 to estimate the transition matrix between hidden states and the Viterbi algorithm to estimate the hidden states themselves [92].

We found several clustered regions of the high LD state on scaffold 702.1, within the possible bounds of the inversion given by the comparative alignment (a smaller region of the high LD state was found outside of this region, and was ignored). We used the combination of these clustered regions to define the likely location of the 'left' breakpoint of the putative inversion between 11.69 and 12.90 mb on scaffold 702.1. The 'right' bound was similarly defined on scaffold 128 between 4.98 and 6.19 mb.

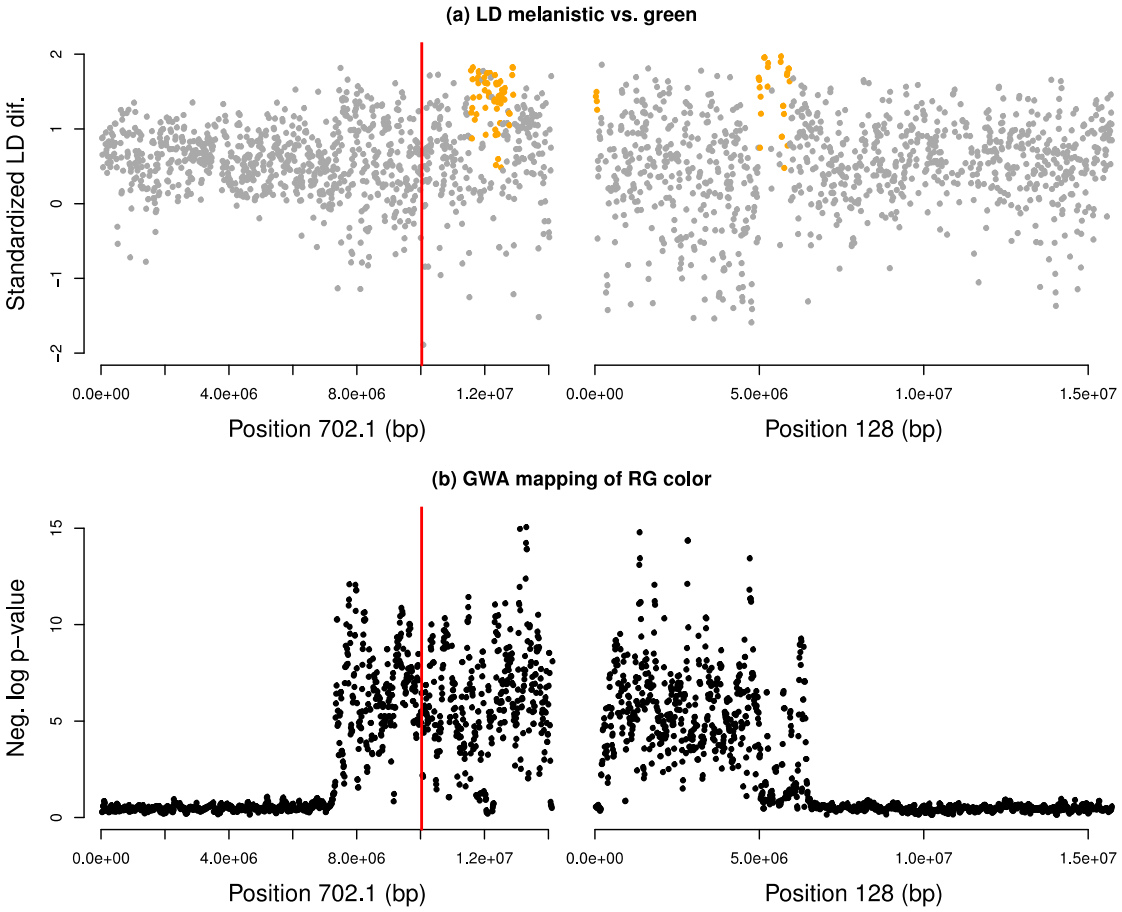


Figure S7. Evidence for chromosomal inversion. Panel (a) shows a standardized metric of the difference in linkage disequilibrium (LD) between *T. cristinae* individuals homozygous for the melanistic or green haplotype. The position of the putative inversion is bounded by information from a comparative whole genome alignment (vertical red line). Orange points denote regions of elevated differences in LD within this region based on a two-state Hidden Markov model and define the breakpoints for the putative inversion. The breakpoints are included in a region of the genome that is highly associated with color variation in single SNP GWA analyses (panel (b); results are shown for RG).

Figures S8-S11 present GWA results for the three newly studied *Timema* species. See the appropriate Materials and Methods section for details on GENABEL analyses.

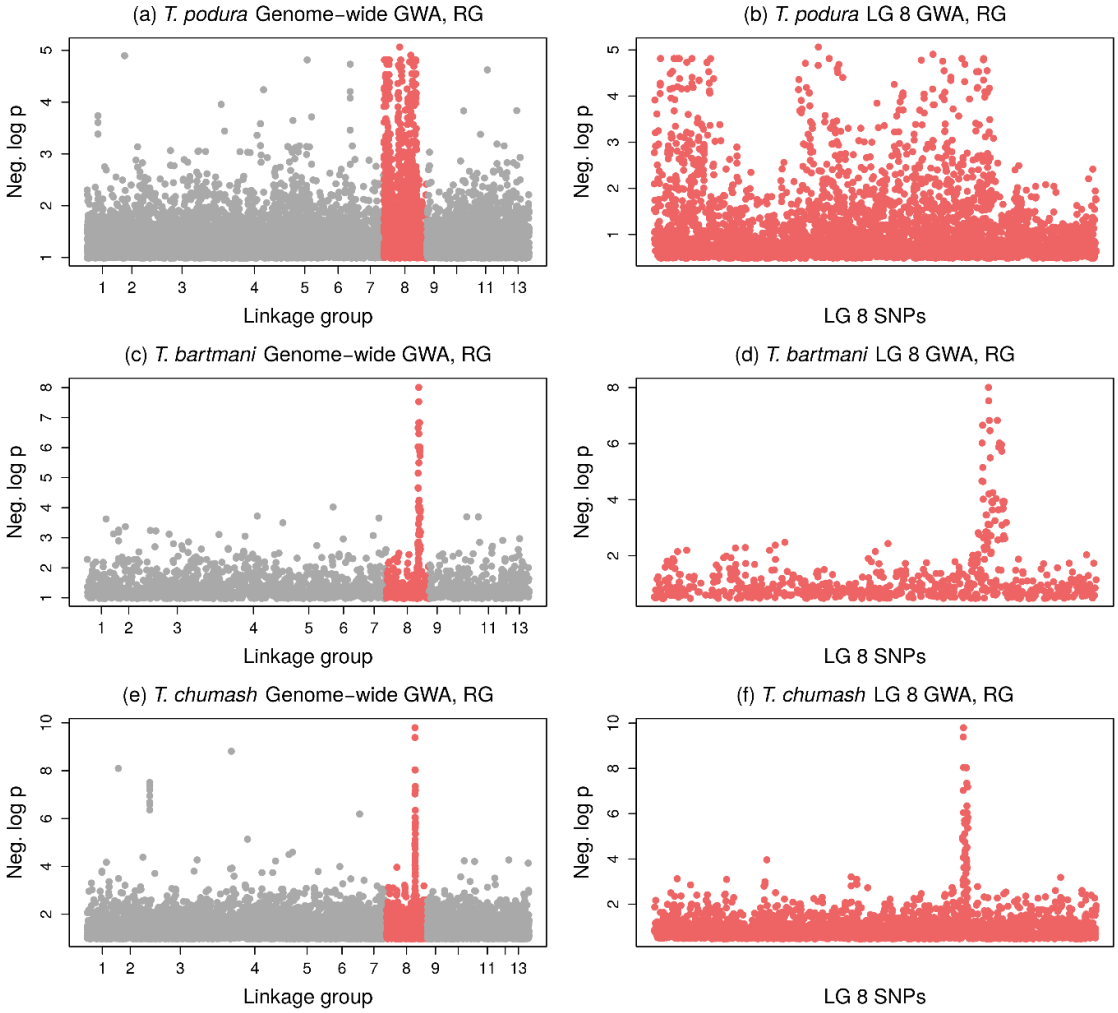


Figure S8. Single-locus genome wide association (GWA) mapping of red-green (RG) color variation in three *Timema* species, at two genomic scales (genome wide and for the single linkage group (LG8) showing the bulk of associations), with correction for population structure. The y-axis shows the negative $\log_{10}$ *P-value* (Neg. log p) for each test that a single-nucleotide polymorphism (SNP) is associated with color variation. At the scale of LG8, *T. chumash* exhibits a peak of association (which actually represents several distinct peaks when zoomed in further on scaffold 128, see Fig. 2), *T. bartmani* a narrow 'block' of association, and *T. podura* a wide block of association. For details on linkage disequilibrium, genetic correlations between traits, etc. see Figures of the main text.

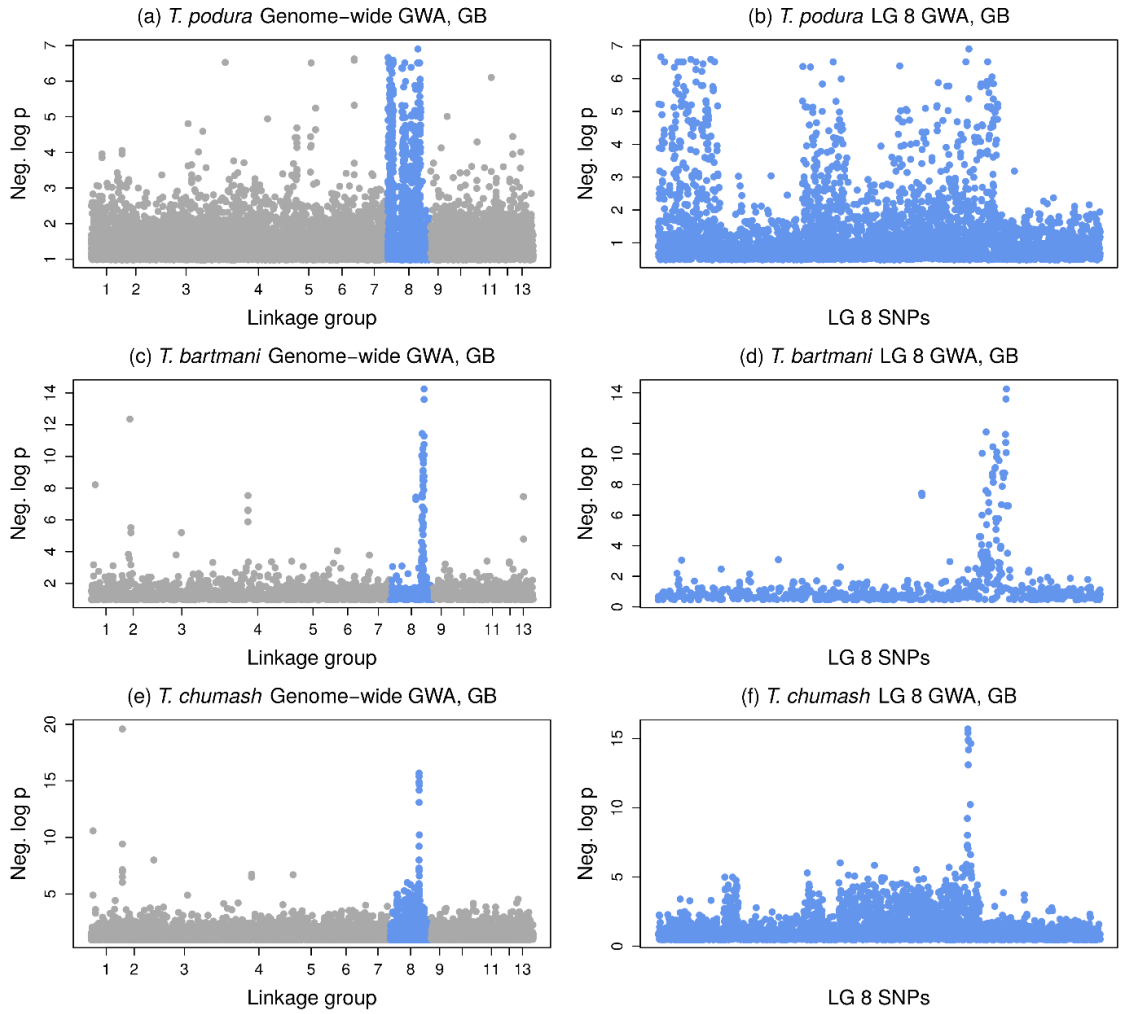


Figure S9. Single-locus genome wide association (GWA) mapping of green-blue (GB) color variation in three *Timema* species, at two genomic scales (genome wide and for the single linkage group (LG8) showing the bulk of associations), without correction for population structure. The y-axis shows the negative $\log_{10}$ *P-value* (Neg. log p) for each test that a single-nucleotide polymorphism (SNP) is associated with color variation. At the scale of LG8, *T. chumash* exhibits a peak of association (which actually represents several distinct peaks when zoomed in further on scaffold 128, see Fig. 2), *T. bartmani* a narrow 'block' of association, and *T. podura* a wide block of association. For details on linkage disequilibrium, genetic correlations between traits, etc. see Figures of the main text.

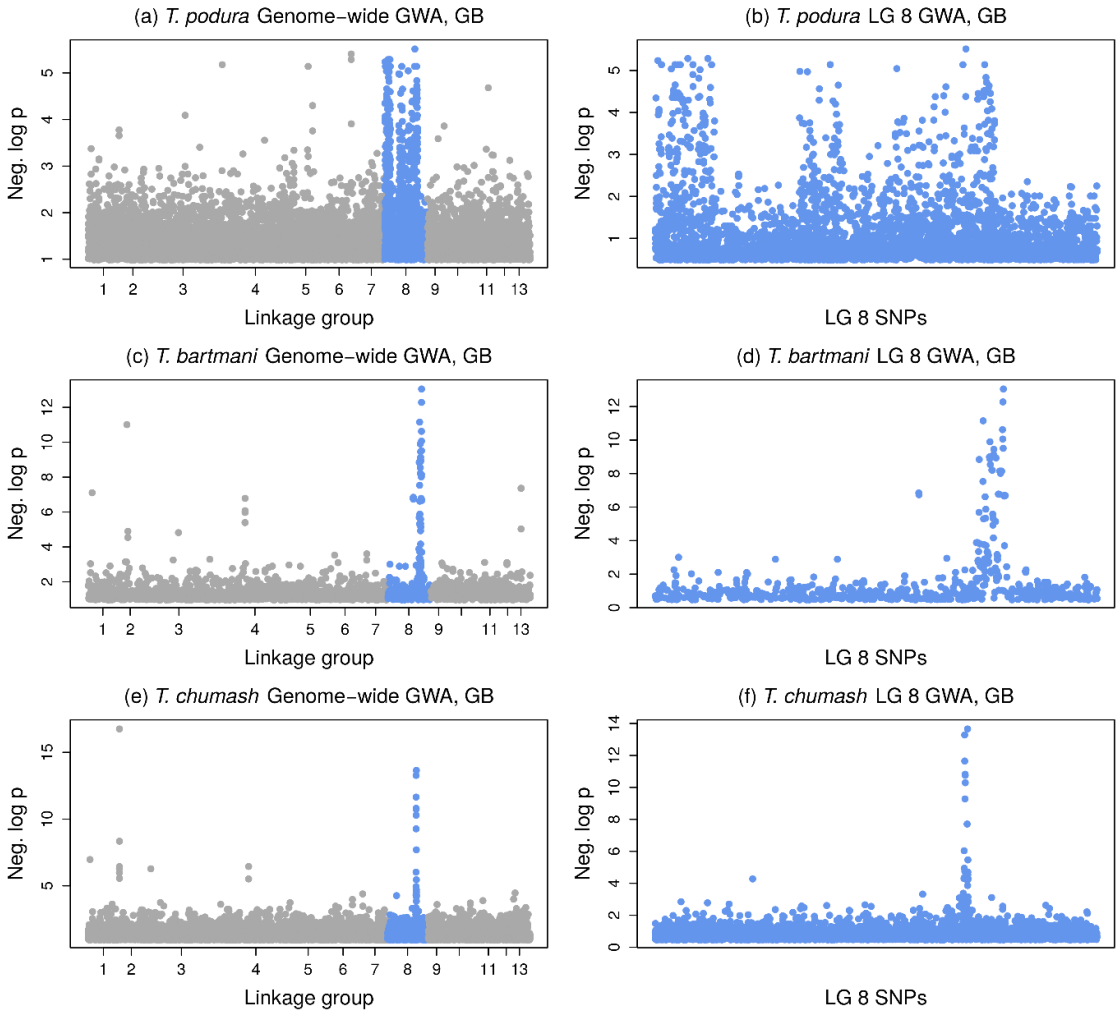


Figure S10. Single-locus genome wide association (GWA) mapping of green-blue (GB) color variation in three *Timema* species, at two genomic scales (genome wide and for the single linkage group (LG8) showing the bulk of associations), without correction for population structure. The y-axis shows the negative $\log_{10}$ *P-value* (Neg. log p) for each test that a single-nucleotide polymorphism (SNP) is associated with color variation. At the scale of LG8, *T. chumash* exhibits a peak of association (which actually represents several distinct peaks when zoomed in further on scaffold 128, see Fig. 2), *T. bartmani* a narrow 'block' of association, and *T. podura* a wide block of association. For details on linkage disequilibrium, genetic correlations between traits, etc. see Figures 2 and 5 of the main text.

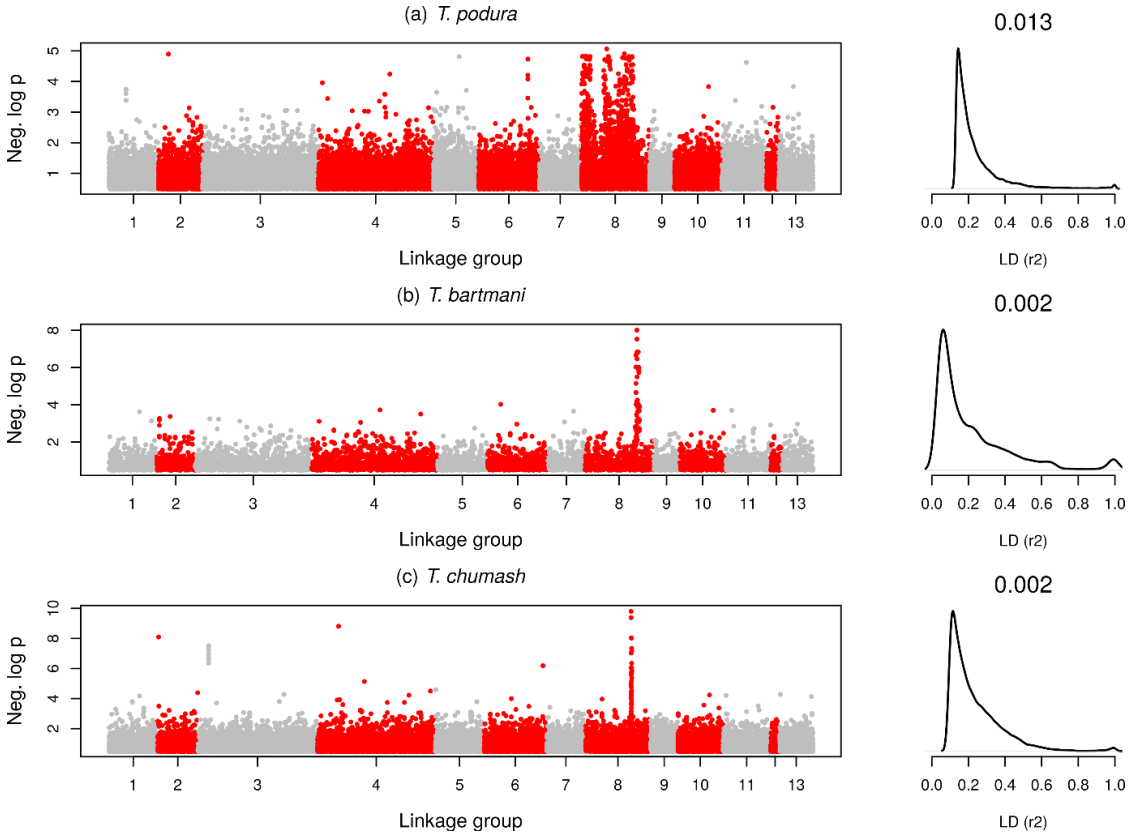


Figure S11. Results of single-locus genome wide association (GWA) mapping of red-green (RG) color variation in the three polymorphic *Timema* species used in this study (left-hand panel of each row), and linkage disequilibrium (LD) analyses (right-hand panel of each row). Results shown do correct for population structure. The y-axis in the left-hand panel shows the negative $\log_{10}$ *P-value* (Neg. log p) for each test that a single-nucleotide polymorphism (SNP) is associated with color variation. The right-hand panels show density plots of the distribution of pairwise LD (measured as r^2) within the *Mel-Stripe* locus for the 5% of pairs of SNPs with the highest LD. Median LD for all pairs of SNPs within *Mel-Stripe* are given above each density plot.

Phylogenetics. We used both GBS and whole genome re-sequencing (WGS) data to infer trees. For GBS analyses, we aligned reads of a subset of 60 individuals of *T. bartmani*, *T. chumash* and *T. podura* (10 of each morph for each species, see main text) to the *T. cristinae* reference genome 1.3c2 using BOWTIE2 and called variants with SAMTOOLS mpileup and BCFTOOLS (see 'Genotyping-by-sequencing, alignment, and variant calling' above for details). As before, we filtered out variants that had reads for fewer than 50% of the individuals, a quality score below 20, a depth greater than 10 times the number of individuals, more than two alleles, or a minor allele frequency lower than 1%. This resulted in 208,770 variants, which were subsequently subsampled for downstream analyses (see main text).

Whole genome re-sequencing data was obtained for a total of 48 individuals from *T. bartmani*, *T. chumash* and *T. podura* sampled in 2015 (8 melanistic *T. bartmani* from populations JL, PCT and hosts IC, JP, WF, WP; 19 green *T. bartmani* from populations JL, PCT and hosts IC, JP, PP, WF, WP; 15 green *T. chumash* from populations BS, GR8.06 and hosts C, MM, Q; 6 melanistic *T. podura* from populations BS, PCT and hosts C,WF, WP; Host-plant abbreviations are as follows. C: *Ceanothus spinosus*, IC: *Calocedrus decurrens*, M: *Arctostaphylos sp.*, MM: *Cercocarpus sp.*,

P: *Pinus sp.*, Q: *Quercus sp.*, WF: *Abies concolor*, WP: *Pinus flexilis*). We extracted genomic DNA for these individuals from 3 to 5 legs using Quiagen's DNeasy Blood and Tissue Kit. We shipped the genomic DNA on dry ice to the Welcome Trust for Human Genetics at the University of Oxford, who prepared multiplexed whole genome resequencing libraries from it (allowing one to assign reads to a particular individual even after pooling libraries on a lane). Libraries were pooled and sequenced along with samples for another project on three lanes of a HiSeq 4000. We obtained de-multiplexed paired-reads for every individual from Oxford genomics.

We aligned the DNA sequence reads to version 1.3c2 of *T. cristinae* genome using the bwa mem algorithm (version 0.7.10-r789)[93]. For the alignments, we set the minimum seed length to 20, the band width to 100, and the internal seed search option (-k) to 1.3. We used a minimum score for output of 30. We then compressed, sorted and indexed the alignments with samtools (version 1.5)[74,93]. Next, we removed PCR duplicates from the alignments using PicardTools MakrDuplicates (version 2.1.1)(**http://broadinstitute.github.io/picard)**. We used GATK's HaplotypeCaller (GATK version 3.5-0-g36282e4) for variant calling [94]. We did this in two steps by first generating individual g.vcf files and then performing joint variant calling with the GenotypeGVCFs command. We ran the HaplotypeCaller with a prior probability of heterozygosity of 0.001, a minimum mapping base quality of 30, and with the 'aggressive' PCR-error correction model. We filtered the initial set of SNP variants identified by GATK using a series of custom perl scripts. Specifically, we removed SNPs with a mean coverage of <1X per individual, with fewer than four reads supporting the non-reference allele, with mapping qualities <40, Phred-scaled *p-values* from Fisher's Exact Test for strand bias of > 60, and with rank-sum test statistics (absolute values) more extreme than 8, 12.5 and 8, for the mapping quality, read position, and ratio of variant confidence tests, respectively. This left us with 3,297,072 SNPs for downstream analyses.

Phenotypic measurements of plant coloration from photographs

Table S10 presents the host-plant samples used in this study. See appropriate Materials and Methods section for details.

Table S10. Details about the host plant samples used in this study.

Host code	**Host plant**	**Population code**	**No samples**
A	*Adenostoma*	DZR	102
C	*Ceanothus*	NH	86
MM	*Cercocarpus*	GR	98
Q	*Quercus*	SM	99
Q	*Quercus*	GR	89
Q	*Quercus*	BC	99
WF	*Abies*	BM	101
WP	*Pinus*	JL	87

Manipulative field experiment

The following paragraphs provide further details on aspects of the manipulative field experiment. For an overall description of the experiment, see the appropriate Materials and Methods section.

Host and treatment rationale:

The rationale for the choice of hosts is as follows. *C. spinosus* is a core host of *T. cristinae*, and is thought to impose strong disruptive selection between leaves and stems [58]. However, we could not use this plant species because it is not found within the *T. chumash* species range. *T. chumash* is found in southern California, at times in sympatry or parapatry with *T. podura [71]*. The first treatment thus specifically focused on two hosts that are found in southern California (*Ceanothus leucodermis* and *Adenostoma fasciculatum*), the combination of which is thought to impose strong divergent selection for green versus melanistic coloration in *T. podura* (but fitness of intermediates was not measured)[57]. Accordingly, each replicate in this treatment used one plant individual of each of these plant species, where the individuals were touching each other. Note that *T. chumash* regularly uses *Ceanothus leucodermis* in the wild [57,71], and can survive on *Adenostoma fasciculatum* [95].

The second treatment of *Cercocarpus* was chosen because it exhibits relatively weak differentiation in plant coloration between stems and leaves (see main text), and it's use by *T. chumash* is associated with weak differentiation between morphs. Moreover, the moderate, shrubby size of *Cercocarpus* individuals mirrors that of *Ceanothus* and *Adenostoma*, and is amenable to mark-recapture experimentation. In contrast, large oak trees (*Quercus* spp.) are not amenable to such experimentation.

Relative fitnesses estimation:

We defined the relative fitnesses (w) of the color morphs as: w_green = 1-s, w_intermediate = 1, and w_melanistic = 1-t (as in Eq. 1.25c in [96]). We then computed posterior estimates of s and t from the relative fitness data. Thus, s or t < 0 implies disruptive selection, whereas s or t > 0 implies intermediate advantage. Note that as the relative fitnesses are ratios and must be >= 0, positive values of s and t must be less than 1, but negative values can be much, much larger. The results for the mountain mahogany treatment were s = 0.3827, 95% ETPIs = -0.4555, 0.7543, pp s < 0 = 0.1367 and t = 0.6039, 95% ETPIs = -0.0017, 0.8711, pp t < 0 = 0.0253. The results for the chamise and California lilac treatment were s = -2.7363, 95% ETPIs = -20.3366, 0.0130, pp s < 0 = 0.9740 and t = -1.8357, 95% ETPIs = -15.6886, 0.3212, pp t < 0 = 0.9248).

References

56. Lindtke D, Lucek K, Soria-Carrasco V, Villoutreix R, Farkas TE, et al. (2017) Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. Molecular Ecology 26: 6189-6205.
57. Comeault AA, Carvalho CF, Dennis S, Soria-Carrasco V, Nosil P (2016) Color phenotypes are under similar genetic control in two distantly related species of Timema stick insect. Evolution 70: 1283-1296.
58. Comeault AA, Flaxman SM, Riesch R, Curran E, Soria-Carrasco V, et al. (2015) Selection on a Genetic Polymorphism Counteracts Ecological Speciation in a Stick Insect. Current Biology 25: 1-7.
59. Team RDC (2013) R: A Language and Environment for Statistical Computing. Vienna, Austria.

60. Gomez D (2006) AVICOL, a program to analyse spectrometric data. Last update October 2011. Free executable available.
61. Endler JA (1993) THE COLOR OF LIGHT IN FORESTS AND ITS IMPLICATIONS. Ecological Monographs 63: 1-27.
62. Finger E, Burkhardt D (1994) BIOLOGICAL ASPECTS OF BIRD COLORATION AND AVIAN COLOR-VISION INCLUDING ULTRAVIOLET RANGE. Vision Research 34: 1509-1514.
63. Cuthill IC, Bennett ATD, Partridge JC, Maier EJ (1999) Plumage reflectance and the objective assessment of avian sexual dichromatism. American Naturalist 153: 183-200.
64. Schmidt V, Schaefer HM, Winkler H (2004) Conspicuousness, not colour as foraging cue in plant-animal signalling. Oikos 106: 551-557.
65. Nosil P, Crespi BJ (2006) Experimental evidence that predation promotes divergence in adaptive radiation. Proceedings of the National Academy of Sciences of the United States of America 103: 9090-9095.
66. Bennett ATD, Cuthill IC (1994) Ultraviolet Vision in Birds: What Is Its Function? . Vision Research 34: 1471–1478.
67. Endler JA, Mielke PW (2005) Comparing entire colour patterns as birds see them. Biological Journal of the Linnean Society 86: 405-431.
68. Odeen A, Hastad O (2003) Complex distribution of avian color vision systems revealed by sequencing the SWS1 opsin from total DNA. Molecular Biology and Evolution 20: 855-861.
69. Odeen A, Hastad O, Alstrom P (2011) Evolution of ultraviolet vision in the largest avian radiation - the passerines. Bmc Evolutionary Biology 11.
70. Maia R, Eliason CM, Bitton PP, Doucet SM, Shawkey MD (2013) pavo: an R package for the analysis, visualization and organization of spectral data. Methods in Ecology and Evolution 4: 906-913.
71. Riesch R, Muschick M, Lindtke D, Villoutreix R, Comeault AA, et al. (2017) Transitions between phases of genomic differentiation during stick-insect speciation. Nature Ecology and Evolution 1: 0082.
72. Nosil P, Villoutreix R, de Carvalho CF, Farkas TE, Soria-Carrasco V, et al. (2018) Natural selection and the predictability of evolution in Timema stick insects. Science 359: 765-+.
73. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357-U354.
74. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
75. Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, et al. (2014) Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. Molecular Ecology 23: 4555-4573.
76. Gompert Z, Lucas LK, Nice CC, Buerkle CA (2013) GENOME DIVERGENCE AND THE GENETIC ARCHITECTURE OF BARRIERS TO GENE FLOW BETWEEN LYCAEIDES IDAS AND L-MELISSA. Evolution 67: 2498-2514.
77. Zhou X, Carbonetto P, Stephens M (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models. Plos Genetics 9.
78. Crawford L, Zeng P, Mukherjee S, Zhou X (2017) Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. Plos Genetics 13.

79. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767-769.
80. Lomsadze A, Burns PD, Borodovsky M (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Research 42.
81. Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24: 637-644.
82. Smit AFA, Hubley R (2008) RepeatModeler Open-1.0. 2008-2015 (version 1.0.8). http://www.repeatmasker.org/RepeatMasker-open-4-0-1.tar.gz.
83. Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, et al. (2014) Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. Science 344: 738-742.
84. Rognes T, Flouri T, Nichols B, Quince C, Mahe F (2016) VSEARCH: a versatile open source tool for metagenomics. Peerj 4.
85. Smit AFA, Hubley R, Green P (2010) Repeatmasker Open 3.0. <**Error! Hyperlink reference not valid.**.
86. Comeault AA, Sommers M, Schwander T, Buerkle CA, Farkas TE, et al. (2012) De novo characterization of the Timema cristinae transcriptome facilitates marker discovery and inference of genetic divergence. Molecular Ecology Resources 12: 549-561.
87. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.
88. Jones P, Binns D, Chang HY, Fraser M, Li WZ, et al. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30: 1236-1240.
89. Hartigan JA, Wong MA (1979) Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society Series C (Applied Statistics) 28: 100-108.
90. Venables WN, Ripley BD (2002) Modern Applied Statistics with S; Chambers J, Eddy W, Härdle W, Sheather S, Tierney L, editors. New York, NY: Springer New York.
91. Harte D (2017) HiddenMarkov: Hidden Markov Models. R package version 1.8-11. Statistics Research Associates, Wellington. URL: http://www.statsresearch.co.nz/dsh/sslib/.
92. Baum LE, Petrie T, Soules G, Weiss N (1970) A MAXIMIZATION TECHNIQUE OCCURRING IN STATISTICAL ANALYSIS OF PROBABILISTIC FUNCTIONS OF MARKOV CHAINS. Annals of Mathematical Statistics 41: 164-&.
93. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
94. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297-1303.
95. Nosil P, Sandoval CP (2008) Ecological niche dimensionality and the evolutionary diversification of stick insects. PLoS One 3: e1907.
96. Ewens WJ (2004) Mathematical population genetics. I. Theoretical introduction. Interdisciplinary applied mathematics, 27.

References

Abràmoff, M. D., Magalhães, P. J. and Ram, S. J. (2004) 'Image Processing with ImageJ', *Biophotonics International*, 11, pp. 36–42.

Alexa A, Rahnenfuhrer J (2018). topGO: Enrichment Analysis for Gene Ontology. R package version 2.34.0.

Agosta, S. J. (2006) 'On ecological fitting, plant-insect associations, herbivore host shifts, and host plant selection', *Oikos*, 114(3), pp. 556–565.

Akalin, A. *et al.* (2012) 'MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles', *Genome Biology*, 13(10).

Altschul, S. F. *et al.* (1997) 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res*, 25(17), pp. 3389–3402.

Ardura, A. *et al.* (2017) 'Epigenetic signatures of invasive status in populations of marine invertebrates', *Scientific Reports*. Nature Publishing Group, 7(January), pp. 1–10.

Arraj, J. A. and Marinus, M. G. (1983) 'Phenotypic reversal in dam mutants of Escherichia coli K-12 by a recombinant plasmid containing the dam+ gene', *Journal of Bacteriology*, 153(1), pp. 562–565.

Arsenault, S. V., Hunt, B. G. and Rehan, S. M. (2018) 'The effect of maternal care on gene expression and DNA methylation in a subsocial bee', *Nature Communications*. Springer US, 9(1), p. 3468.

Becker, C. *et al.* (2011) 'Spontaneous epigenetic variation in the Arabidopsis thaliana methylome', *Nature*, 480(7376), pp. 245–249.

Bernays, E. and Graham, M. (1988) 'On the Evolution of Host Specificity in Phytophagous', *Ecology*, 69, pp. 886–892.

Bewick, A. J. *et al.* (2017) 'Evolution of DNA methylation across insects', *Molecular Biology and Evolution*, 34(3), pp. 654–665.

Bewick, A. J. *et al.* (2019) 'Dnmt1 is essential for egg production and embryo viability in the large milkweed bug, Oncopeltus fasciatus', *Epigenetics & Chromatin*. BioMed Central, 12(1), pp. 1–14.

Bird, A. (2007) 'Perceptions of epigenetics', *Nature*, 447(7143), pp. 396–398.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120.

Bonasio, R. *et al.* (2012) 'Genome-wide and caste-specific DNA methylomes of the ants camponotus floridanus and harpegnathos saltator', *Current Biology*, 22(19), pp. 1755–1764.

Bossdorf, O. *et al.* (2010) 'Experimental alteration of DNA methylation affects the phenotypic plasticity of ecologically relevant traits in Arabidopsis thaliana', *Evolutionary Ecology*, 24(3), pp. 541–553..

Bossdorf, O., Richards, C. L. and Pigliucci, M. (2008) 'Epigenetics for ecologists', *Ecology Letters*, 11(2), pp. 106–115.

Brown, J. H. . and Kodric-Brown, A. (1979) 'Convergence , Competition , and Mimicry in a Temperate Community of Hummingbird- Pollinated Flowers', 60(5), pp. 1022–1035.

Cardoso, C. *et al.* (2019) 'Domain structure and expression along the midgut and carcass of peritrophins and cuticle proteins analogous to peritrophins in insects with and without peritrophic membrane', *Journal of Insect Physiology*. Elsevier, 114(February), pp. 1–9.

Carja, O. *et al.* (2017) 'Worldwide patterns of human epigenetic variation', *Nature Ecology and Evolution*. Springer US, 1(10), pp. 1577–1583. doi: 10.1038/s41559-017-0299-z.

Charmantier, A. *et al.* (2008) 'Adaptive phenotypic plasticity in response to climate change in a wild bird population.', *Science*, 320(5877), pp. 800–803.

Chung, H. *et al.* (2014) 'A single gene affects both ecological divergence and mate choice in Drosophila.', *Science (New York, N.Y.)*, 343(6175), pp. 1148–51.

Clarke, R. T., Rothery, P., & Raybould, A. F. (2002). Confidence limits for regression relationships between distance matrices: Estimating gene flow with distance. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*(3), 361–372.

Cokus, S. J. *et al.* (2008) 'Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning', *Nature*, 452(7184), pp. 215–219.

Comeault, A. A. *et al.* (2012) 'De novo characterization of the *Timema cristinae* transcriptome facilitates marker discovery and inference of genetic divergence', *Molecular Ecology Resources*, 12(3), pp. 549–561.

Comeault, A. A. *et al.* (2015) 'Selection on a Genetic Polymorphism Counteracts Ecological Speciation in a Stick Insect', *Current Biology*, 25(15), pp. 1975–1981.

Comeault, A. A. *et al.* (2016) 'Color phenotypes are under similar genetic control in two distantly related species of Timima stick insect', *Evolution*, pp. 1283–1296.

Comeault, A. a *et al.* (2014) 'Genome-wide association mapping of phenotypic traits

subject to a range of intensities of natural selection in Timema cristinae.', *The American naturalist*, 183(5), pp. 711–27.

Conover, D. O., Duffy, T. A. and Hice, L. A. (2009) 'The Covariance between Genetic and Environmental Influences across Ecological Gradients', *Annals of the New York Academy of Sciences*, 1168(1), pp. 100–129.

Cortijo, S. *et al.* (2014) 'Mapping the epigenetic basis of complex traits', *Science*, 343(6175), pp. 1145–1148.

Crevillén, P. *et al.* (2014) 'Epigenetic reprogramming that prevents transgenerational inheritance of the vernalized state', *Nature*, 515(7528), pp. 587–590..

Cubas, P., Vincent, C. and Coen, E. (1999) 'An epigenetic mutation responsible for natural variation in floral symmetry', *Nature*, 401(6749), pp. 157–161.

Cunningham, C. B. *et al.* (2015) 'The genome and methylome of a beetle with complex social behavior, nicrophorus vespilloides (coleoptera: Silphidae)', *Genome Biology and Evolution*, 7(12), pp. 3383–3396.

Darwin, C. (1859). On the origin of species. John Murray, London, UK

Deichmann, U. (2016a) 'Epigenetics : The origins and evolution of a fashionable topic', *Developmental Biology*, 416, pp. 249–254.

Deichmann, U. (2016b) 'Why epigenetics is not a vindication of Lamarckism – and why that matters', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*. Elsevier Ltd, 57, pp. 80–82.

Dimond, J. L. and Roberts, S. B. (2016) 'Germline DNA methylation in reef corals: Patterns and potential roles in response to environmental change', *Molecular Ecology*, 25(8), pp. 1895–1904.

Dobin, A. *et al.* (2012) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21.

Dobzhansky, T. (1937). Genetics and the Origin of Species. Columbia university press, New York.

Dostálová, A. *et al.* (2011) 'The midgut transcriptome of Phlebotomus (Larroussius) perniciosus, a vector of Leishmania infantum: comparison of sugar fed and blood fed sand flies', *BMC Genomics*, 12(1).

Dubin, M. J. *et al.* (2015) 'DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation', *eLife*, 4(MAY), pp. 3–5.

Duncan, E. J., Gluckman, P. D. and Dearden, P. K. (2014) 'Epigenetics, plasticity, and evolution: How do we link epigenetic change to phenotype?', *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 322(4), pp. 208–220.

Ehrlich, P. R. . and Raven, P. H. . (1964) 'Butterflies and Plants : A Study in Coevolution', *Evolution*, 18(4), pp. 586–608.

Elias, I., & Lagergren, J. (2007). Constrained hidden Markov models for population-based haplotyping. *BMC Bioinformatics*, *8*, 8–89.

Farkas, T. E. *et al.* (2013) 'Evolution of camouflage drives rapid ecological change in an insect community', *Current Biology*. Elsevier Ltd, 23(19), pp. 1835–1843.

Feil, R. and Fraga, M. F. (2012) 'Epigenetics and the environment: Emerging patterns and implications', *Nature Reviews Genetics*. Nature Publishing Group, 13(2), pp. 97–109.

Feinberg, A. P. and Irizarry, R. a (2010) 'Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease.', *Proceedings of the National Academy of Sciences of the United States of America*, 107 Suppl, pp. 1757–1764.

Feng, S. *et al.* (2010) 'Conservation and divergence of methylation patterning in plants and animals', *Proceedings of the National Academy of Sciences*, 107(19), pp. 8689–8694.

Fisher, R. A. (1930). The genetic theory of natural selection. Oxford University Press, Oxford, UK.

Fischer, H. M. *et al.* (2008) 'Evolutionary origins of a novel host plant detoxification gene in butterflies', *Molecular Biology and Evolution*, 25(5), pp. 809–820.

Foret, S. *et al.* (2012) 'DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees', *Proceedings of the National Academy of Sciences*, 109(13), pp. 4968–4973.

Forsman, A. (2015) 'Rethinking phenotypic plasticity and its consequences for individuals, populations and species', *Heredity*. Nature Publishing Group, 115(4), pp. 276–284.

Foust, C. M. *et al.* (2016) 'Genetic and epigenetic differences associated with environmental gradients in replicate populations of two salt marsh perennials', *Molecular Ecology*, 25(8), pp. 1639–1652.

Gaspar, J. M., & Hart, R. P. (2017). DMRfinder: Efficiently identifying differentially methylated regions from MethylC-seq data. *BMC Bioinformatics*, *18*(1), 1–8.

Ghalambor, C. K. *et al.* (2007) 'Adaptive versus non-adaptive phenotypic plasticity and the

potential for contemporary adaptation in new environments', *Functional Ecology*, 21(3), pp. 394–407.

Ghalambor, C. K. *et al.* (2015) 'Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature', *Nature*, 525(7569), p. 372.

Glastad, K. M., Goodisman, M. A. D., *et al.* (2016) 'Effects of DNA Methylation and Chromatin State on Rates of Molecular Evolution in Insects', *G3: Genes, Genomes, Genetics*, 6(2), pp. 357–363.

Glastad, K. M., Gokhale, K., *et al.* (2016) 'The caste- and sex-specific DNA methylome of the termite Zootermopsis nevadensis', *Scientific Reports*. Nature Publishing Group, 6(August), pp. 1–14.

Glastad, K. M. *et al.* (2017) 'Variation in DNA Methylation Is Not Consistently Reflected by Sociality in Hymenoptera', *Genome biology and evolution*, 9(6), pp. 1687–1698.

Glastad, K. M., Hunt, B. G. and Goodisman, M. A. D. (2018) 'Epigenetics in Insects : Genome Regulation and the Generation of Phenotypic Diversity', (September), pp. 1–19.

Goldberg, A. D., Allis, C. D. and Bernstein, E. (2007) 'Epigenetics: A Landscape Takes Shape', *Cell*, 128(4), pp. 635–638. doi: 10.1016/j.cell.2007.02.006.

Goll, M. G. *et al.* (2006) 'Supporting Online Material for Methylation of tRNA Asp by the DNA Methyltransferase Homolog Dnmt2', *Science*, 311(395), pp. 395–399.

Goll, M. G. and Bestor, T. H. (2005) 'Eukaryotic Cytosine Methyltransferases', *Annual Review of Biochemistry*, 74(1), pp. 481–514.

Gompert, Z. *et al.* (2013) 'Genome divergence and the genetic architecture of barriers to gene flow between lycaeides idas and l. melissa', *Evolution*, 67(9), pp. 2498–2514.

Gompert, Z., Lucas, L. K., *et al.* (2014) 'Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants', *Molecular Ecology*, 23(18), pp. 4555–4573.

Gompert, Z., Comeault, A. a, *et al.* (2014) 'Experimental evidence for ecological selection on genome variation in the wild.', *Ecology letters*, 17(3), pp. 369–79.

Gore, C. J. and Schal, C. (2004) 'Gene Expression and Tissue Distribution of the Major Human Allergen Bla g 1 in the German Cockroach, Blattella germanica L. (Dictyoptera: Blattellidae)', *Journal of Medical Entomology*, 41(5), pp. 953–960. doi: 10.1603/0022-2585-41.5.953.

van der Graaf, A. *et al.* (2015) 'Rate, spectrum, and evolutionary dynamics of spontaneous

epimutations', *Proceedings of the National Academy of Sciences*, 112(21), pp. 6676–6681.

Griffith JS, Mahler HR. DNA ticketing theory of memory. Nature 1969; 223:580 -2.

Groot, M. P. *et al.* (2018) 'Epigenetic population differentiation in field- and common garden-grown Scabiosa columbaria plants', *Ecology and Evolution*, 8(6), pp. 3505–3517.

Guerrero-Bosagna, C. (2017) 'Evolution with no reason: A Neutral view on epigenetic changes, genomic variability, and evolutionary novelty', *BioScience*, 67(5), pp. 469–476.

Hagmann, J. *et al.* (2015) 'Century-scale Methylome Stability in a Recently Diverged Arabidopsis thaliana Lineage', *PLoS Genetics*, 11(1).

Herrera, C. M. and Bazaga, P. (2010) 'Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet Viola cazorlensis', *New Phytologist*, 187(3), pp. 867–876.

Herrera, C. M. and Bazaga, P. (2011) 'Untangling individual variation in natural populations: Ecological, genetic and epigenetic correlates of long-term inequality in herbivory', *Molecular Ecology*, 20(8), pp. 1675–1688.

Herrera, C. M., Medrano, M. and Bazaga, P. (2013) 'Epigenetic Differentiation Persists after Male Gametogenesis in Natural Populations of the Perennial Herb Helleborus foetidus (Ranunculaceae)', *PLoS ONE*, 8(7), pp. 1–8.

Herrera, C. M., Medrano, M. and Bazaga, P. (2014) 'Variation in DNA methylation transmissibility, genetic heterogeneity and fecundity-related traits in natural populations of the perennial herb Helleborus foetidus', *Molecular Ecology*, 23(5), pp. 1085–1095.

Herrera, C. M., Medrano, M. and Bazaga, P. (2016) 'Comparative spatial genetics and epigenetics of plant populations: Heuristic value and a proof of concept', *Molecular Ecology*, 25(8), pp. 1653–1664.

Herrera, C. M., Pozo, M. I. and Bazaga, P. (2012) 'Jack of all nectars, master of most: DNA methylation and the epigenetic basis of niche width in a flower-living yeast', *Molecular Ecology*, 21(11), pp. 2602–2616.

Hohenlohe, P. A. *et al.* (2010) 'Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags', *PLoS Genetics*, 6(2).

Hu, J. and Barrett, R. D. H. (2017) 'Epigenetics in natural animal populations', *Journal of Evolutionary Biology*, 30(9), pp. 1612–1632.

Huang, X. *et al.* (2017) 'Rapid response to changing environments during biological invasions: DNA methylation perspectives', *Molecular Ecology*, 26(23), pp. 6621–6633.

Hunt, B. G. *et al.* (2013) 'Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects', *Genome Biology and Evolution*, 5(3), pp. 591–598.

Jablonka, E. and Lamb, M. J. (1998) 'Epigenetic inheritance in evolution', *Journal of Evolutionary Biology*, 11(2), pp. 159–183.

Jablonka, E. and Raz, G. (2009) 'Transgenerational Epigenetic Inheritance: Prevelance, Mechanisms, and Implications for the Study of Heredity and Evolution', 84(2), pp. 131–176.

Janz, N., Nyblom, K. and Nylin, S. (2001) 'Evolutionary dynamic of host-plant specialization: A case study of the tribe Nymphalini', *Evolution*, 55(4), pp. 783–796.

Janz, N., Nylin, S. and Wahlberg, N. (2006) 'Diversity begets diversity: Host expansions and the diversification of plant-feeding insects', *BMC Evolutionary Biology*, 6, pp. 1–10.

Jenkins, D. G. . *et al.* (2010) 'A meta-analysis of isolation by distance: relic or reference standard for landscape genetics?', *Ecography*, 33(February), pp. 315–320.

Jiang, L. *et al.* (2013) 'Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos', *Cell*. Elsevier Inc., 153(4), pp. 773–784.

Johannes, F. *et al.* (2009) 'Assessing the impact of transgenerational epigenetic variation on complex traits', *PLoS Genetics*, 5(6).

Johnson, L. J. and Tricker, P. J. (2010) 'Epigenomic plasticity within populations: Its evolutionary significance and potential', *Heredity*. Nature Publishing Group, 105(1), pp. 113–121.

Jones, P. A. (2012) 'Functions of DNA methylation: Islands, start sites, gene bodies and beyond', *Nature Reviews Genetics*. Nature Publishing Group, 13(7), pp. 484–492.

Keller, T. E. and Yi, S. V. (2014) 'DNA methylation and evolution of duplicate genes', *Proceedings of the National Academy of Sciences*, 111(16), pp. 5932–5937.

Kille, P. *et al.* (2013) 'DNA sequence variation and methylation in an arsenic tolerant earthworm population', *Soil Biology and Biochemistry*. Elsevier Ltd, 57, pp. 524–532.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120.

Kong, A. *et al.* (2012) 'Rate of de novo mutations and the importance of father-s age to disease risk', *Nature*. Nature Publishing Group, 488(7412), pp. 471–475.

Krauss, V., Eisenhardt, C. and Unger, T. (2009) 'The genome of the stick insect Medauroidea extradentata is strongly methylated within genes and repetitive DNA', *PLoS ONE*, 4(9).

Krueger, F. and Andrews, S. R. (2011) 'Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications', *Bioinformatics*, 27(11), pp. 1571–1572.

Kucharski, R. *et al.* (2008) 'Nutritional Control of Reproductive Status in Honeybees via DNA methylation', *Science*, 319(5831), pp. 1827–1831.

Kucharski, R., Maleszka, J. and Maleszka, R. (2016) 'A possible role of DNA methylation in functional divergence of a fast evolving duplicate gene encoding odorant binding protein 11 in the honeybee', *Proceedings of the Royal Society B: Biological Sciences*, 283(1833).

Laland, K. *et al.* (2014) 'Does evolutionary theory need a rethink?', *Nature*, 514, pp. 161–164.

Lande, R. (2009) 'Adaptation to an extraordinary environment by evolution of phenotypic plasticity and genetic assimilation', *Journal of Evolutionary Biology*, 22(7), pp. 1435–1446.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nat Methods*, 9(4), pp. 357–359.

Larose, C., Rasmann, S. and Schwander, T. (2019) 'Evolutionary dynamics of specialisation in herbivorous stick insects', *Ecology Letters*, 22(2), pp. 354–364.

Law, J. A. and Jacobsen, S. E. (2010) 'Establishing, maintaining and modifying DNA methylation patterns in plants and animals', *Nature Reviews Genetics*. Nature Publishing Group, 11(3), pp. 204–220.

Lê, S., Josse, J. and Husson, F. (2008) 'FactoMineR: An R Package for Multivariate Analysis', *J. of Statistical Software*, 25(1), pp. 1–18.

Lea, A. J. *et al.* (2016) 'Resource base influences genome-wide DNA methylation levels in wild baboons (Papio cynocephalus)', *Molecular Ecology*, 25(8), pp. 1681–1696.

Lea, A. J. *et al.* (2017) 'Maximizing ecological and evolutionary insight in bisulfite sequencing data sets', *Nature Ecology & Evolution*, 1(8), pp. 1074–1083.

Lea, A. J., Tung, J. and Zhou, X. (2015) 'A Flexible, Efficient Binomial Mixed Model for Identifying Differential DNA Methylation in Bisulfite Sequencing Data', *PLoS Genetics*, 11(11), pp. 1–31.

Ledón-Rettig, C. C. (2013) 'Ecological epigenetics: An introduction to the symposium', *Integrative and Comparative Biology*, 53(2), pp. 307–318.

Li-Byarlay, H. *et al.* (2013) 'RNA interference knockdown of DNA methyl-transferase 3 affects gene alternative splicing in the honey bee', *Proceedings of the National Academy of Sciences*, 110(31), pp. 12750–12755.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079.

Libbrecht, R. *et al.* (2016) 'Robust DNA methylation in the clonal raider ant brain', *Current Biology*, 26(3), pp. 391–395.

Lidholm, D., Lohe, A. R. and Hart, D. L. (1993) 'The transposable element mariner mediates germline transformation in Drosophila melanogaster', *Genetics*, 134(3), pp. 859–868.

Liebl, A. L. *et al.* (2013) 'Patterns of DNA methylation throughout a range expansion of an introduced songbird', *Integrative and Comparative Biology*, 53(2), pp. 351–358.

Lindtke, D. *et al.* (2017) 'Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect', *Molecular Ecology*, 26(22), pp. 6189–6205.

Lira-Medeiros, C. F. *et al.* (2010) 'Epigenetic variation in mangrove plants occurring in contrasting natural environment', *PLoS ONE*, 5(4), pp. 1–8.

Lo, N., Simpson, S. J. and Sword, G. A. (2018) 'Epigenetics and developmental plasticity in orthopteroid insects', *Current Opinion in Insect Science*. Elsevier Inc., 25, pp. 25–34.

Lorincz, M. C. *et al.* (2004) 'Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells', *Nature Structural and Molecular Biology*, 11(11), pp. 1068–1075.

Lyko, F. *et al.* (2010) 'The honey bee epigenomes: Differential methylation of brain DNA in queens and workers', *PLoS Biology*, 8(11). doi: 10.1371/journal.pbio.1000506.

Lynch, M. (2000) 'The Evolutionary Fate and Consequences of Duplicate Genes', *Science*, 290(5494), pp. 1151–1155. doi: 10.1126/science.290.5494.1151.

Lynch, M. (2002) 'Gene Duplication and Evolution', *Science*, 297(5583), pp. 945–947. doi: 10.1126/science.293.5535.1551a.

Maderspacher, F. (2010) 'Lysenko rising', *Current Biology*. Elsevier, 20(19), pp. R835–R837. doi: 10.1016/j.cub.2010.09.009.

Manning, K. *et al.* (2006) 'A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening', *Nature Genetics*, 38(8), pp. 948–952.

Mantel, N. (1967). Cancer research. *Cancer Research*, *27*(1), 209–220.

Massicotte, R., Whitelaw, E. and Angers, B. (2011) 'DNA methylation: A source of random variation in natural populations', *Epigenetics*, 6(4), pp. 422–428.

May, R. (1990) 'How many species?', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 330, pp. 293–304.

Medrano, M., Herrera, C. M. and Bazaga, P. (2014) 'Epigenetic variation predicts regional and local intraspecific functional diversity in a perennial herb', *Molecular Ecology*, 23(20), pp. 4926–4938.

Mello, M. O. and Silva-Filho, M. C. (2002) 'Plant-insect interactions: an evolutionary arms race between two distinct defense mechaisms', *Brazilian Journal of Plant Physiology*, 14(2), pp. 71–81.

Merzendorfer, H. (2006) 'Insect chitin synthases: A review', *Journal of Comparative Physiology B: Biochemical, Systemic, and Environmental Physiology*, 176(1), pp. 1–15.

Metzger, D. C. H. and Schulte, P. M. (2018) 'Similarities in temperature-dependent gene expression plasticity across timescales in threespine stickleback (Gasterosteus aculeatus)', *Molecular Ecology*, 27(10), pp. 2381–2396.

Misof, B. *et al.* (2014) 'Phylogenomics resolves the timing and pattern of insect evolution', *Science*, 346(6210), pp. 763–767.

Mitsudome, T. *et al.* (2015) 'Biochemical characterization of maintenance DNA methyltransferase DNMT-1 from silkworm, Bombyx mori', *Insect Biochemistry and Molecular Biology*. Elsevier Ltd, 58, pp. 55–65.

Moczek, A. P. (2010) 'Phenotypic plasticity and diversity in insects', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1540), pp. 593–603.

Morgan, T. H. (1915). The mechanism of Mendelian heredity. Holt.

Morgan, H. D. *et al.* (1999) 'Epigenetic inheritance at the agouti locus in the mouse', *Nature Genetics*, 23(3), pp. 314–318.

Mukherjee, K., Twyman, R. M. and Vilcinskas, A. (2015) 'Insects as models to study the epigenetic basis of disease', *Progress in Biophysics and Molecular Biology*. Elsevier Ltd, 118(1–2), pp. 69–78.

Nicotra, A. B. *et al.* (2015) 'Adaptive plasticity and epigenetic variation in response to warming in an Alpine plant', *Ecology and Evolution*, 5(3), pp. 634–647.

Niederhuth, C. E. and Schmitz, R. J. (2017) 'Putting DNA methylation in context: from genomes to gene expression in plants', *Biochimica et Biophysica Acta - Gene Regulatory*

Mechanisms. Elsevier B.V., 1860(1), pp. 149–156.

Nolan, T. *et al.* (2011) 'Analysis of two novel midgut-specific promoters driving transgene expression in Anopheles stephensi mosquitoes', *PLoS ONE*, 6(2).

Nosil, P. (2007) 'Divergent host plant adaptation and reproductive isolation between ecotypes of Timema cristinae walking sticks.', *The American naturalist*, 169(2), pp. 151–162.

Nosil, P. *et al.* (2012) 'Genomic consequences of multiple speciation processes in a stick insect.', *Proceedings. Biological sciences / The Royal Society*, 279(1749), pp. 5058–65.

Nosil, P. *et al.* (2018) 'Natural selection and the predictability of evolution inTimemastick insects.', *Science (New York, N.Y.)*, 359(6377), pp. 765–770.

Nosil, P. and Crespi, B. . J. . (2004) 'Does Gene Flow Constrain Adaptive Divergence or Vice Versa? A Test Using Ecomorphology and Sexual Isolation in Timema cristinae Walking-Sticks', *Evolution*, 58(1), pp. 102–112.

Nosil, P. and Crespi, B. J. (2006) 'Experimental evidence that competition promotes divergence in adaptive radiation', *PNAS*, 103(24), pp. 9090–9095.

Nosil, P. and Crespi, B. J. (2006) 'Experimental evidence that predation promotes divergence in adaptive radiation.', *Proceedings of the National Academy of Sciences of the United States of America*, 103(24), pp. 9090–5.

Nosil, P., Crespi, B. J. and Sandoval, C. P. (2002) 'Host-plant adaptation drives the parallel evolution of reproductive isolation', *Nature*, 417(6887), pp. 440–443.

Nosil, P., Egan, S. P. and Funk, D. J. (2008) 'Heterogeneous Genomic Differentiation between Walking-Stick Ecotypes: "Isolation by Adaptation" and Multiple Roles for Divergent Selection', *Evolution*, 62(2), pp. 316–336.

Nosil, P. and Sandoval, C. P. (2008) 'Ecological niche dimensionality and the evolutionary diversification of stick insects', *PLoS ONE*, 3(4). doi: 10.1371/journal.pone.0001907.

Nosil, P., Sandoval, C. P. and Crespi, B. J. (2006) 'The evolution of host preference in allopatric vs. parapatric populations of Timema cristinae walking-sticks', *Journal of Evolutionary Biology*, 19(3), pp. 929–942.

Nugent, B. M. *et al.* (2015) 'Brain feminization requires active repression of masculinization via DNA methylation', *Nature Neuroscience*, 18(5), pp. 690–697.

Nylin, S. and Janz, N. (2009) 'Butterfly host plant range: An example of plasticity as a promoter of speciation?', *Evolutionary Ecology*, 23(1), pp. 137–146.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., & Suggests, M. A. S. S. (2007). The vegan package. *Community ecology package*, *10*, 631-637.

Onuchic, V. *et al.* (2018) 'Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci', *Science*, 1354(September), p. eaar3146.

Pál, C. and Miklós, I. (1999) 'Epigenetic inheritance, genetic assimilation and speciation.', *Journal of theoretical biology*, 200(1), pp. 19–37.

Parchman, T. L. *et al.* (2012) 'Genome-wide association genetics of an adaptive trait in lodgepole pine', *Molecular Ecology*, 21(12), pp. 2991–3005.

Parrott, B. B. *et al.* (2013) 'Differential Incubation Temperatures Result in Dimorphic DNA Methylation Patterning of the SOX9 and Aromatase Promoters in Gonads of Alligator (Alligator mississippiensis) Embryos1', *Biology of Reproduction*, 90(1), pp. 1–11.

Paszkowski, J. and Grossniklaus, U. (2011) 'Selected aspects of transgenerational epigenetic inheritance and resetting in plants', *Current Opinion in Plant Biology*. Elsevier Ltd, 14(2), pp. 195–203.

Patalano, S. *et al.* (2015) 'Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies', *Proceedings of the National Academy of Sciences*, 112(45), pp. 13970–13975.

Pecinka, A., Abdelsamad, A. and Vu, G. T. H. (2013) 'Hidden genetic nature of epigenetic natural variation in plants', *Trends in Plant Science*. Elsevier Ltd, 18(11), pp. 624–632.

Peterson, B. K. *et al.* (2012) 'Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species', *PLoS ONE*, 7(5).

Pfennig, D. W. *et al.* (2010) 'Phenotypic plasticity's impacts on diversification and speciation', *Trends in Ecology and Evolution*. Elsevier Ltd, 25(8), pp. 459–467.

Pigliucci, M. (2005) 'Evolution of phenotypic plasticity: Where are we going now?', *Trends in Ecology and Evolution*, 20(9), pp. 481–486.

Platt, A. *et al.* (2015) 'Genome-wide signature of local adaptation linked to variable CpG methylation in oak populations', *Molecular Ecology*, 24(15), pp. 3823–3830.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *DSC Working Papers*. R

Preite, V. *et al.* (2015) 'The epigenetic footprint of poleward range-expanding plants in apomictic dandelions', *Molecular Ecology*, 24(17), pp. 4406–4418.

Preite, V. *et al.* (2018) 'Increased transgenerational epigenetic variation, but not predictable epigenetic variants, after environmental exposure in two apomictic dandelion lineages', *Ecology and Evolution*, 8(5), pp. 3047–3059.

Prentis, P. J. *et al.* (2008) 'Adaptive evolution in invasive species', *Trends in Plant Science*, 13(6), pp. 288–294.

Provataris, P. *et al.* (2018) 'Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola', *Genome Biology and Evolution*, 10(March), pp. 1185–1197.

QGIS Development Team. (2016). QGIS Geographic Information System. Open Source Geospatial Foundation

Qiu, C. *et al.* (2002) 'The PWWP domain of mammalian DNA methyltransferase Dnmt3b defines a new family of DNA-binding folds', *Nature Structural Biology*, 9(3), pp. 217–224.

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

Ramsahoye, B. H. *et al.* (2000) 'Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a', *Proceedings of the National Academy of Sciences*, 97(10), pp. 5237–5242.

Randall, T. A. *et al.* (2013) 'Genomic, RNAseq, and molecular modeling evidence suggests that the major allergen domain in insects evolved from a homodimeric origin', *Genome Biology and Evolution*, 5(12), pp. 2344–2358.

Reik, W. (2007) 'Stability and flexibility of epigenetic gene regulation in mammalian development.', *Nature*, 447(7143), pp. 425–432.

Reznick, D. N. and Ghalambor, C. K. (2001) 'The population ecology of contemporary adaptations: What empirical studies reveal about the conditions that promote adaptive evolution', *Genetica*, 112–113(1956), pp. 183–198. doi: 10.1023/A:1013352109042.

Richards, C. L. *et al.* (2017) 'Ecological plant epigenetics: Evidence from model and non-model species, and the way forward', *Ecology Letters*, 20(12), pp. 1576–1590.

Richards, C. L., Bossdorf, O. and Pigliucci, M. (2010) 'What Role Does Heritable Epigenetic Variation Play in Phenotypic Evolution?', *BioScience*, 60(3), pp. 232–237.

Richards, C. L., Schrey, A. W. and Pigliucci, M. (2012) 'Invasion of diverse habitats by few Japanese knotweed genotypes is correlated with epigenetic differentiation', *Ecology Letters*, 15(9), pp. 1016–1025.

Richards, E. J. (2006) 'Inherited epigenetic variation - Revisiting soft inheritance', *Nature Reviews Genetics*, 7(5), pp. 395–401. d

Richards, E. J. (2008) 'Population epigenetics', *Current opinion in genetics & development*, 18(2), pp. 221–226.

Riesch, R. *et al.* (2017) 'Transitions between phases of genomic differentiation during stick-insect speciation', *Nature Ecology and Evolution*. Macmillan Publishers Limited, part of Springer Nature., 1(4), pp. 1–13.

Rodin, S. N. and Riggs, A. D. (2003) 'Epigenetic silencing may aid evolution by gene duplication', *Journal of Molecular Evolution*, 56(6), pp. 718–729.

Rousset, F. (1997). Genetic Differentiation and estimation of Gene Flow from F-Statistics Under Isolation by Distance. *Genetics*, *145*, 1219–1228.

Sandoval, C. P. (1994) 'Differential visual predation on morphs of Timema cristinae (Phasmatodeae:Timemidae) and its consequences for host range', *Biological Journal of the Linnean Society*, 52, pp. 341–356.

Sandoval, C. P. (1994) 'The effects of the relative geographic scales of gene flow and selection on morph frequencies in the walking-stick Timema cristinae', *Evolution*, 48(6), pp. 1866–1879.

Sandoval, C. (2000). 'Persistence of a walking-stick population (Phasmatoptera: Timematodea) after a wildfire.' *The Southwestern Naturalist*, 123-127.

Sandoval, C. P. and Nosil, P. (2005) 'Counteracting selective regimes and host preference evolution in ecotypes of two species of walking-sticks.', *Evolution; international journal of organic evolution*, 59(11), pp. 2405–13.

Sati, S. *et al.* (2012) 'High resolution methylome map of rat indicates role of intragenic DNA methylation in identification of coding region', *PLoS ONE*, 7(2), pp. 1–12.

Savković, U. *et al.* (2016) 'Experimentally induced host-shift changes life-history strategy in a seed beetle', *Journal of Evolutionary Biology*, 29(4), pp. 837–847.

Schlichting, C. D. and Smith, H. (2002) 'Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes', *Evolutionary Ecology*, 16(1), pp. 189–211.

Schmid, M. W. *et al.* (2018) 'Contribution of epigenetic variation to adaptation in Arabidopsis', *Nature Communications*. Springer US, 9(1).

Schmitz, R. J. *et al.* (2011) 'Transgenerational epigenetic instability is a source of novel methylation variants', *Science*, 334(6054), pp. 369–373.

Schmitz, R. J. *et al.* (2013) 'Patterns of population epigenomic diversity', *Nature*, 495(7440), pp. 193–198.

Schoonhoven, L. M., Van Loon, B., van Loon, J. J., & Dicke, M. (2005). Insect-plant biology. Oxford University Press on Demand.

Schrey, A. W. *et al.* (2013) 'Ecological epigenetics: Beyond MS-AFLP', *Integrative and Comparative Biology*, 53(2), pp. 340–350.

Schübeler, D. (2015) 'Function and information content of DNA methylation', *Nature*, 517(7534), pp. 321–326.

Scoville, A. G. and Pfrender, M. E. (2010) 'Phenotypic plasticity facilitates recurrent rapid adaptation to introduced predators', *Proceedings of the National Academy of Sciences*, 107(9), pp. 4260–4263.

Sexton, J. P., Hangartner, S. B. and Hoffmann, A. A. (2014) 'Genetic isolation by environment or distance: Which pattern of gene flow is most common?', *Evolution*, 68(1), pp. 1–15.

Shafer, A. B. A. and Wolf, J. B. W. (2013) 'Widespread evidence for incipient ecological speciation: A meta-analysis of isolation-by-ecology', *Ecology Letters*, 16(7), pp. 940–950.

Silva, C. P. *et al.* (2001) 'Induction of digestive α-amylases in larvae of Zabrotes subfasciatus (Coleoptera: Bruchidae) in response to ingestion of common bean α-amylase inhibitor 1', *Journal of Insect Physiology*, 47(11), pp. 1283–1290.

Simmen, M. W. (2008) 'Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals', *Genomics*, 92(1), pp. 33–40.

Simonsen, M., Mailund, T. and Pedersen, C. N. S. (2008) 'Rapid Neighbour-Joining', *Proceedings of the 8th Workshop in Algorithms in Bioinformatics*, pp. 113–122.

Simonsen, M., & Pedersen, C. N. S. (2011). Rapid computation of distance estimators from nucleotide and amino acid alignments. *Proceedings of the ACM Symposium on Applied Computing*, (1), 89–93.

Simpson, S. J., Sword, G. A. and Lo, N. (2011) 'Polyphenism in insects', *Current Biology*. Elsevier Ltd, 21(18), pp. R738–R749.

Skinner, M. K. *et al.* (2014) 'Epigenetics and the evolution of darwin's finches', *Genome Biology and Evolution*, 6(8), pp. 1972–1989.

Smith, G. and Ritchie, M. G. (2013) 'How might epigenetics contribute to ecological speciation?', *Current Zoology*, 59(5), pp. 686–696.

Smith, T. A. *et al.* (2016) 'Epigenetic divergence as a potential first step in darter speciation', *Molecular Ecology*, 25(8), pp. 1883–1894.

Soria-Carrasco, V., Gompert, Z., Comeault, A. a, *et al.* (2014) 'Stick insect genomes reveal natural selection's role in parallel speciation.', *Science (New York, N.Y.)*, 344(6185), pp. 738–42.

Soria-Carrasco, V., Gompert, Z., Comeault, A. a., *et al.* (2014) *Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation*, *Science*.

Standage, D. S. *et al.* (2016) 'Genome, transcriptome and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect', *Molecular Ecology*, 25(8), pp. 1769–1784.

Stevenson, T. J. (2017) 'Environmental and hormonal regulation of epigenetic enzymes in the hypothalamus', *Journal of Neuroendocrinology*, 29(5), pp. 1–9.

Suazo, A., Gore, C. and Schal, C. (2009) 'RNA interference-mediated knock-down of Bla g 1 in the German cockroach, Blattella germanica L., implicates this allergen-encoding gene in digestion and nutrient absorption', *Insect Molecular Biology*, 18(6), pp. 727–736.

Suzuki, M. M. and Bird, A. (2008) 'DNA methylation landscapes: Provocative insights from epigenomics', *Nature Reviews Genetics*, 9(6), pp. 465–476.

Taudt, A., Colomé-Tatché, M. and Johannes, F. (2016) 'Genetic sources of population epigenomic variation', *Nature Reviews Genetics*. Nature Publishing Group, 17(6), pp. 319–332.

Trapnell, C. *et al.* (2013) 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nature Protocols*, 7(3), pp. 562–578.

Trucchi, E. *et al.* (2016) 'BsRADseq: Screening DNA methylation in natural populations of non-model species', *Molecular Ecology*, 25(8), pp. 1697–1713.

Verhoeven, K. J. F. *et al.* (2010) 'Stress-induced DNA methylation changes and their heritability in asexual dandelions', *New Phytologist*, 185(4), pp. 1108–1118.

Verhoeven, K. J. F., VonHoldt, B. M. and Sork, V. L. (2016) 'Epigenetics in ecology and evolution: What we know and what we need to know', *Molecular Ecology*, 25(8), pp. 1631–1638.

Vickery, Vernon R. (1993). Revision of Timema Scudder (Phasmatoptera: Timematodea) including three new species. The Canadian Entomologist 125.4 p657-692.

Villoutreix, R., de Carvalho, C.F., Soria-Carrasco, V., Lindtke, D., De-la-Mora, M., Muschick,

M., Feder, J.L., Gompert, Z. & Nosil, P. (2019) ECryptic coloration in stick insects exhibits both supergene architecture and recombination between color genes. Manuscript in preparation

Waddington, C. H. (1939). An introduction to modern genetics. George Allen And Unwin Ltd Museum Street; London.

Waddington, C. H. (1942). The epigenotype. Endeavour, 1, 18-20.

Waddington, C. H. (1953) 'Genetic Assimilation of an Acquired Character', *Evolution*, 7(2), pp. 118–126.

Walsh, T. K. *et al.* (2010) 'A functional DNA methylation system in the pea aphid, Acyrthosiphon pisum', *Insect Molecular Biology*, 19(SUPPL. 2), pp. 215–228.

Wang, J., Marowsky, N. C. and Fan, C. (2014) 'Divergence of gene body DNA methylation and evolution of plant duplicate genes', *PLoS ONE*, 9(10).

Wang, X. *et al.* (2013) 'Function and Evolution of DNA Methylation in Nasonia vitripennis', *PLoS Genetics*, 9(10).

Wang, X. *et al.* (2014) 'The locust genome provides insight into swarm formation and long-distance flight', *Nature communications*, 5, p. 2957.

Wang, X., Werren, J. H. and Clark, A. G. (2016) 'Allele-Specific Transcriptome and Methylome Analysis Reveals Stable Inheritance and Cis-Regulation of DNA Methylation in Nasonia', *PLoS Biology*, 14(7), pp. 1–21.

Waterland, R. and Jirtle, R. (2003) 'Transposable Elements: Targets for early nutritional effects on epigenetic gene regulation', *Molecular and Cellular Biology*, 23(15), pp. 5293–5300.

West-Eberhard, Mary Jane. Developmental plasticity and evolution. Oxford University Press, 2003.

Wheat, C. W. *et al.* (2007) 'The genetic basis of a plant insect coevolutionary key innovation', *Proceedings of the National Academy of Sciences*, 104(51), pp. 20427–20431.

Xiang, H. *et al.* (2010) 'Single base-resolution methylome of the silkworm reveals a sparse epigenomic map', *Nature Biotechnology*, 28(5), pp. 516–520.

Xie, H. J. *et al.* (2015) 'ICE1 demethylation drives the range expansion of a plant invader through cold tolerance divergence ', *Molecular Ecology*, 24(4), pp. 835–850.

Yoder, J. A., Walsh, C. P. and Bestor, T. H. (1997) 'Cytosine methylation and the ecology of intragenomic parasites', *Trends*, 13, pp. 335–340. doi: 10.1016/S0168-9525(97)01181-5.

Yu, Y. *et al.* (2013) 'Cytosine Methylation Alteration in Natural Populations of Leymus chinensis Induced by Multiple Abiotic Stresses', *PLoS ONE*, 8(2), pp. 1–10.

Zemach, A. *et al.* (2010) 'Genome-wide evolutionary analysis of eukaryotic DNA methylation.', *Science (New York, N.Y.)*, 328(5980), pp. 916–9.

Zhang, X. *et al.* (2006) 'Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis', *Cell*, 126(6), pp. 1189–1201.

Zhang, Y. Y. *et al.* (2013) 'Epigenetic variation creates potential for evolution of plant phenotypic plasticity', *New Phytologist*, 197(1), pp. 314–322.

Zhou, X., Carbonetto, P. and Stephens, M. (2013) 'Polygenic modeling with bayesian sparse linear mixed models.', *PLoS genetics*, 9(2), p. e1003264.