8-2019

Quantifying the Predictability of Evolution at the Genomic Level in Lycaeides Butterflies

Samridhi Chaturvedi
Utah State University

QUANTIFYING THE PREDICTABILITY OF EVOLUTION AT THE GENOMIC LEVEL IN *LYCAEIDES* BUTTERFLIES

by

Samridhi Chaturvedi

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Biology

Approved:

Zachariah Gompert, Ph.D. Major Professor	Karen Kapheim, Ph.D. Committee Member
Susannah French, Ph.D. Committee Member	Karen Mock, Ph.D. Committee Member
Matthew Forister, Ph.D. Committee Member	Richard S. Inouye, Ph.D. Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah
2019

ABSTRACT

QUANTIFYING THE PREDICTABILITY OF EVOLUTION AT THE GENOMIC LEVEL IN *LYCAEIDES* BUTTERFLIES

by

Samridhi Chaturvedi, Doctor of Philosophy

Utah State University,

Major Professor: Zachariah Gompert, Ph.D.
Department: Biology

Repeatable phenotypic evolution includes parallel and convergent evolution in independent populations in response to similar environmental challenges and implies natural selection. These repeated genetic changes suggest that predictable genetic changes can be indentified. However, the extent of predictability of evolution and the different circumstances in which evolution is more or less predictable remain unclear. This dissertation attempts to identify and quantify the degree to which evolution is predictable and studies different mechanisms which contribute to evolution of *Lycaeides* butterflies. I evaluate predictability in various contexts by testing for overlap in genomic loci associated with a evolving trait or associated with a specific evolutionary process. These contexts include comparing natural populations on a geographical and a temporal scale, comparing natural and laboratory populations, and comparing locations across the genome. In chapter 2, I investigated whether historical admixture can predict patterns of introgression (gene flow between species) in a contemporary hybrid zone using *Lycaeides* butterflies. Here, I first show that both ancient and contemporary hybrid zones experience consistent selection which affects patterns of introgression and genomic composition of hybrids in a similar manner. Therefore, I can predict evolutionary patterns in one hybrid zone from another. In chapter 3, I assessed the predictability of genomic changes underlying a recent host plant shift in *Lycaeides melissa* butterflies. Here, I show genomic changes accompanying this host shift are somewhat predictable depending on the contextual comparisons. Having studied genomic basis of evolution in the previous two chapters, I

address another novel mechanism underlying host plant adaptation in these butterflies. In chapter 4, I assess the sources of variation in the gut microbial community of *Lycaeides melissa* caterpillars. Here, I show that caterpillar gut microbial communities vary over time and differ between frass and whole caterpillar samples. Diet (host plant) and butterfly population have limited effects on microbial communities. Collectively, these results demonstrate that I can use different contexts to study predictability of evolution. However, the degree of predictability varies across different contextual approaches. Quantifying the extent to which evolution is predictable can be crucial in understanding the causes and consequences of evolutionary predictability.

(224 pages)

PUBLIC ABSTRACT

QUANTIFYING THE PREDICTABILITY OF EVOLUTION AT THE GENOMIC LEVEL IN *LYCAEIDES* BUTTERFLIES

Samridhi Chaturvedi

Stephen Jay Gould, a great scientist and evolutionary biologists, suggested that if we could replay the tape of life, we would not have observed similar course of events because evolution is stochastic and if affected by several events. Since then, the possibility that evolution is repeatable or predictable has been debated. Studies using large-scale evolution experiments, long-term data for individual populations, and controlled experiments in nature, have demonstrated phenotypic and genetic convergence in several taxa. These studies suggest that despite some randomness, predictable evolutionary patterns can emerge on a large temporal and spatial scale. However, a few cases also exist where evolution is unpredictable and stochastic. One way to understand evolutionary predictability better can be to have quantitative estimates of predictability at different heirarchical levels (mutations, genetic, phenotypic). This can help better understand if evolution is predictable and the extent to which it is predictable. My dissertation uses *Lycaeides* butterflies to identify and quantify evolutionary predictability in different contexts such as on a geographic scale, temporal scale and genomic scale. I accomplished this by sequencing and annotating the genomes of these butterflies across a vast geographic range and on a temporal scale and by comparing natural and experimental populations. My results show that different mechanisms can assist evolution of organisms to adapt to novel environmental challenges, and that the evolutionary changes can be somewhat predictable. Through this work I demonstrate three main findings: first, quantitative estimates of evolutionary predictability indicate that degree of predictability is variable and is highly context-dependent. Second, we can predict evolutionary patterns on a spatial as well as temporal scale, and can predict patterns in nature by controlled laboratory experiments. Additionally, genomic changes underlying repeatability vary across the genome. Lastly, the approach of quantifying predictability can help us better understand the mechanisms which drive evolution and how organisms will evolve in response to similar environmental pressures. These results suggest that evolution can be constrained and

if we actually replay the tape of life, we could see a considerably similar outcome in biodiversity compared to what Gould predicted.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Zach Gompert for his abled guidance, support, and friendship as I pursued my Ph.D. in a foreign country. Zach, thank you for choosing me to be your first student. I would like to thank my committee members: Dr. Karen Kapheim, Dr. Matt Forister, Dr. Susannah French and Dr. Karen Mock for their support and direction throughout this journey. Dr. Kapheim was there for me during the initial years of my Ph.D. providing intellectual as well as emotional guidance through some tough times. Dr. Forister has provided me with his advice, encouragement and has mentored me like his own student. I would also like to thank Lauren Lucas, who helped me take my first steps in teaching and has been a constant support throughout my Ph.D. I feel lucky to have known her and to have nurtured some really special memories with her. I would like to thank all my wonderful undergraduates who assisted me in projects which sometimes worked and sometimes failed. I would like to acknowledge funding from USU Ecology Center and USU Research and Graduate Studies during my Ph.D.

I would particularly like to thank my peers in the Department of Biology at USU. I started my Ph.D. with three exceptional friends/scientists: Matt, Alberto and Sajeena. They did not continue in the program but each of them taught me something which will stay with me forever. I also dedicate my Ph.D. experience to Matt and his exuberant and ever so beautiful memory. Matt passed away during my Ph.D. but his scientific acumen and zeal for life changed me as a person. Even from far away, these three friends have cheered me on and have helped me grow into a better scientist. Alexandre Rego was there for me in really tough times and has been a friend, confidante and family. Amy Springer held my hand through field work in the Tetons and has since been one of the most brilliant minds and patient human being I have had the privilege of knowing. Tara Saley has been the most awesome labmate and her vivacious laughter has always brightened my day. Mallory Hagadorn and Kate Hunter are my adopted lab mates and they have made this journey much easier and fun. Akila Ram has been with me every time I missed food from home and has been my voice of reason. These and many more friends and faculty members provided me with copious amounts of inspiration and support throughout my Ph.D. and are truly the reason for my success during this program.

Finally, I would like to thank my family back home in India for all their support throughout my time here. My parents, Rishi and Sudha Chaturvedi, have provided me with a lot of encouragement and have showed immense support in helping me pursue this degree. My brother Mayur Chaturvedi and sister-in-law Puja Chaturvedi have always reminded me of how proud they are of my achievements. I would like to thank my sister, Surbhi Chaturvedi, who is my lifeline and I would not have been here without her. Last, but not the least, my life partner and best friend, Rahul Vishwakarma, who stood by me since I dreamt of being a scientist and has always reminded me that I am a strong independent woman capable of pursuing anything.

Samridhi Chaturvedi

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Repeated phenotypic evolution constitutes the parallel or convergent evolution of traits in independent populations in response to similar environmental pressures [3]. Accumulating data now shows that phenotypic and genetic convergence can occur across several taxa [2, 4, 7, 18, 25, 26]. Evolutionary processes such as mutation and random drift do not cause similar evolutionary shifts repeatedly in response to environmental changes. Therefore, repeated use of same underlying genes during parallel and convergent phenotypic evolution can be indicative of constraints on genetic pathways and imply natural selection [17]. Understanding these constraints and their effects on phenotypic evolution can provide a possibility to predict genetic evolution. Therefore, instances of repeated phenotypic evolution can provide an opportunity to identify and measure the predictable genetic changes underlying adaptive evolution and speciation.

Even though several instances of parallel and convergent evolution have been recorded, the stochastic and contingent nature of evolution has been a classic topic of discussion in biology and has been presented with contrasting views [16, 21, 22]. On one end some instances suggest the unpredictable nature of evolution [1, 15], on the other end the repeated evolution of specific traits and genetic changes in response to similar environmental pressures emphasizes that evolution can be predictable [20]. However, multiple selective agents and genetic background (such as standing genetic variation, large effect sizes, higher mutation rates and linkage or epistatic relationships) can affect the probability of repeated use of same genes in natural populations [23]. Therefore, it can be said that predictability of evolution is not discrete and instead can lie along a quantitative continuum [1]. Along these lines, quantification of the extent of predictability at different levels (such as genotypes, phenotypes, genomes and mutations) and different scales (geographic, temporal and genomic scales and comparisons within these scales) can help better understand predictability of evolution. By using different contexts to quantify predictability of genetic changes, we can obtain a deeper understanding of how natural populations will cope with similar environmental challenges due to climate changes or human-mediated habitat changes. In addition, this can help us dissect the various underlying mechanisms which drive adaptive evolution and speciation.

The overarching goal of this dissertation is to better understand predictability of evolution and the mechanisms driving adaptation to novel environments by using *Lycaeides* butterflies. These butterflies provide two interesting avenues to test for predictable evolution: the first is novel host plant colonization, and the second is the existence of an ancient and contemporary hybrid zone in a restricted spatial scale wherein the ancient hybrid zone spans a wide geographic range and the contemporary hybrid zone inhabits a very small space. *Lycaeides melissa* occur throughout western North America and utilize several species of legumes as their native hosts across their geographic range in North Western United States. Since the introduction of *Medicago sativa* (alfalfa) in their host range, some populations have started to colonize this novel host [19]. Along these lines, there is evidence suggesting that populations have persisted and adapted to alfalfa even though it is a poor host compared to the native legumes [5, 6, 24]. This case of novel host colonization provides an exciting opportunity to understand the predictability of genome-wide evolutionary changes associated with a host-plant shift in these butterflies. I use this case to quantify predictability and use several contexts for studying predictable genomic changes underlying a life-history trait. In addition, I also use this background of host-plant shift to understand the role of larval gut microbial community in host use in *L. melissa*. Another aspect of *Lycaeides* biology, which makes it an ideal system to address my research goals in the occurrence of hybridization in several species of *Lycaeides*. Several species of *Lycaeides* hybridize in specific geographic regions and have even formed hybrid lineages. *Lycaeides idas* and *Lycaeides melissa* are two of the 5 nominal species of *Lycaeides* butterflies that occur in North America [10, 13]. Along these lines, there is evidence that *Lycaeides* come into secondary contact in various regions across their geographic range and have hybridized to produce three additional admixed lineages [8, 9, 11, 12]. I use this case of hybridization as another novel context to study predictability.

Using this background, I first identified and quantified the degree of evolutionary predictability in the following contexts: 1) different type of comparisons (among natural populations vs. between natural and experimental populations 2) different scales of comparisons (on a large geographic scale vs. pair of populations in close proximity; on a large temporal scale) and lastly, 3) comparison of different genomic regions (autosomes vs. sex chromosomes). Second, I was broadly interested in

understanding adaptive evolution in these butterflies by studying different mechanisms which drive adaptation and speciation. Host-plant use in herbivorous insects is an interesting case in which to understand adaptation in response to contemporary habitat changes. Therefore, I used genomic and microbiome approaches to understand how *L. melissa* butterflies adapt to novel host plant alfalfa. Hybrid zones can also be used to understand adaptive evolution since introgression can provide novel genetic variation to adapt to environmental changes and can be useful to indentify genetic barriers to gene flow [14]. Therefore, I use *Lycaeides* hybrid zones and a genomics approach to identify genetically differentiated regions in hybrid populations experiencing variable environmental pressures. Together these approaches combined with genomics analyses help me better understand the evolution of *Lycaeides* butterflies over geographic scale wherein I compare natural populations which vary in their geographic distribution (widespread vs. small distribution) and on a temporal scale (ancient vs. contemporary hybrid zone).

In Chapter 2, I approach evolutionary predictability by quantifying predictability of genomic changes in the context of temporal comparisons in natural hybrid zones. Studies of replicate hybrid zones have found evidence of similar patterns of introgression across transects, but these studies generally focus on hybrid zones of a similar age. Whether there is consistency in patterns of introgression over time (at different stages of hybrid zone formation) is less clear. In this chapter, I use relatively old admixed populations of *Lycaeides melissa* and *Lycaeides idas* butterflies (admixture occurred about 14,000 ybp) and populations from a recent, active hybrid zone (hybridization started around 200 years ago and is ongoing) to ask if evolutionary patterns in old admixed populations can predict evolutionary dynamics in the current hybrid zone. Here, I asked two questions. First, how well can I predict genomic regions which are most resistant to gene flow in recent active hybrids from patterns of ancestry in the admixed populations? Second, can I identify the processes which drive repeated patterns of introgression? I used genomic data analyses and genome annotation to first delineate candidate genomic regions which show excess local ancestry in ancient hybrids and genomic regions which show variable patterns of introgression in contemporary hybrids. By identifying these regions, I could pinpoint specific locations of the genome and their functional properties to better understand what traits underlie reproductive isolation in *Lycaiedes* hybrids. I first

found that several regions of the genome show excess ancestry in ancient hybrids and several regions restrict introgression in contemporary hybrids. These regions were spread across the autosomes and sex chromosome. I then saw that similar regions of the genome experience restricted introgression across ancient and contemporary hybrid zones. Second, the level of consistency in overlap of genomic regions is quite high between the two hybrid zones and this indicates that natural selection is the deterministic force driving these patterns of introgression across time. These results highlight that quantification of degree of predictability is possible over a large temporal scale and can be variable in different regions of the genome.

In chapter 3, I measured the predictability of genome-wide evolutionary changes associated with a recent host shift in *Lycaeides melissa*. There are various contexts which can be used to study repeatable genomic changes underlying a phenotype. In addition, quantification of degree of predictability can be crucial in drawing conclusions about the processes driving repeatable patterns of evolution. In this chapter, I used two different contextual approaches to quantify the extent of predictability of patterns of evolutionary change in nature. First, I identified genomic regions most associated with host-use in *L. melissa* and tested if these regions are enriched for specific functional properties. Second, I compared instances of repeated evolution of host shifts across different populations of *L. melissa* across a geographic scale, and, third, I used SNP $\times$ performance associations in a laboratory experiment to predict evolutionary changes underlying host use in nature. I used genomic analyses and delineated genomic regions associated with host use in several *L. melissa* populations. There were several regions which were associated with host use and these were distributed across autosomes and sex chromosome. I then found that there is evidence of parallel genomic changes underlying host use among natural populations spread across a wide geographic range. There is a significant increase in predictability when I compare populations in close proximity and associated with the same host plant. Additionally, I could partially predict genomic regions associated with host use in nature from SNP $\times$ performance associations in a laboratory experiment. However, in both these cases I could not predict the direction of allele frequency changes in nature from those in the performance experiment. These results highlight how the degree of predictability can be variable in different contexts and quantifying predictability can indicate if stochastic or

deterministic processes are driving genomic changes underlying adaptive evolution.

While addressing the main objectives of my research, I was also curious about the role of different mechanisms in adaptation to different environment. I used host-plant adaptation in *L. melissa* as a case to understand adaptive evolution through different mechanisms. In chapter 3, I address adaptation to novel host plant by understanding the genomic basis of novel host plant use. In chapter 4, I tried to understand host plant adaptation in these butterflies by dissecting the role of gut microbiome in host use by assessing the sources of variation in the gut microbial community of *Lycaeides melissa* caterpillars. Host plant use in herbivorous insects is a complex life-history trait which is affected by several aspects of the organisms biology. Insect gut microbiome can facilitate or constrain host plant use. In this chapter, I ask two questions. First, can different aspects of insect and host plant biology affect *Lycaeides melissa* caterpillar gut microbiome? Second, does gut caterpillar gut microbiome community interact with caterpillar performance? I use caterpillar rearing experiments and 16s rRNA microbiome sequencing of host plant, caterpillar frass and whole body to address these questions. I first find that caterpillar age and sample type (frass or whole body) causes variation in gut microbial communities. However, diet (host plant) and population have limited effect on gut microbiome. Second, I found that there is no association of caterpillar gut microbial communities with caterpillar performance. Our results provide general insights into the role of gut microbiome in host plant use in *Lepidotera*.

Finally, in chapter 5, I summarize the findings of the previous three chapters and present a conclusion concerning the specific questions addressed in this dissertation.

REFERENCES

[1] Daniel I Bolnick, Rowan DH Barrett, Krista B Oke, Diana J Rennison, and Yoel E Stuart. (non) parallel evolution. *Annual Review of Ecology, Evolution, and Systematics*, 49:303–330, 2018.

[2] Pascal-Antoine Christin, Daniel M Weinreich, and Guillaume Besnard. Causes and evolutionary significance of genetic convergence. *Trends in Genetics*, 26(9):400–405, 2010.

[3] Gina L Conte, Matthew E Arnegard, Catherine L Peichel, and Dolph Schluter. The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749):5039–5047, 2012.

[4] Kathryn R Elmer and Axel Meyer. Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in ecology & evolution*, 26(6):298–306, 2011.

[5] Matthew L Forister, Chris C Nice, James A Fordyce, and Zachariah Gompert. Host range evolution is not driven by the optimization of larval performance: the case of *Lycaeides melissa* (Lepidoptera: Lycaenidae) and the colonization of alfalfa. *Oecologia*, 160(3):551–561, 2009.

[6] Matthew L Forister, Zachariah Gompert, Chris C Nice, Glen W Forister, and James A Fordyce. Ant association facilitates the evolution of diet breadth in a Lycaenid butterfly. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20101959, 2010.

[7] Nicolas Gompel and Benjamin Prud'homme. The causes of repeated genetic evolution. *Developmental biology*, 332(1):36–47, 2009.

[8] Zachariah Gompert, James A Fordyce, Matthew L Forister, Arthur M Shapiro, and Chris C Nice. Homoploid hybrid speciation in an extreme habitat. *Science*, 314(5807):1923–1925, 2006.

[9] Zachariah Gompert, Chris C Nice, James A Fordyce, Matthew L Forister, and Arthur M Shapiro. Identifying units for conservation using molecular systematics: the cautionary tale of the karner blue butterfly. *Molecular ecology*, 15(7):1759–1768, 2006.

[10] Zachariah Gompert, Matthew L. Forister, James A. Fordyce, and Chris C. Nice. Widespread mitonuclear discordance with evidence for introgressive hybridization and selective sweeps in *Lycaeides*. *Molecular Ecology*, 17(24):5231–5244, 8 2008. ISSN 1365-294X. doi: 10.1111/j.1365-294x.2008.03988.x. URL http://dx.doi.org/10.1111/j.1365-294x.2008.03988.x.

[11] Zachariah Gompert, Matthew L Forister, James A Fordyce, Chris C Nice, Robert J Williamson, and C Alex Buerkle. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of lycaeides butterflies. *Molecular ecology*, 19 (12):2455–2473, 2010.

[12] Zachariah Gompert, Lauren K. Lucas, James A. Fordyce, Matthew L. Forister, and Chris C. Nice. Secondary contact between lycaeides idas and l.Âămelissa in the rocky mountains: extensive admixture and a patchy hybrid zone. *Molecular Ecology*, 19(15):3171–3192, 10 2010. ISSN 1365-294X. doi: 10.1111/j.1365-294x.2010.04727.x. URL http://dx.doi.org/10.1111/j.1365-294x.2010.04727.x.

[13] Zachariah Gompert, Lauren K. Lucas, Chris C. Nice, James A. Fordyce, Matthew L. Forister, and C. Alex Buerkle. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, 66(7):2167–2181, 12 2012. ISSN 1558-5646. doi: 10.1111/j.1558-5646.2012.01587.x. URL http://dx.doi.org/10.1111/j.1558-5646.2012.01587.x.

[14] Zachariah Gompert, Elizabeth G Mandeville, and C Alex Buerkle. Analysis of population genomic data from hybrid zones. *Annual Review of Ecology, Evolution, and Systematics*, 48: 207–229, 2017.

[15] Stephen Jay Gould. *Wonderful life: the Burgess Shale and the nature of history*. WW Norton & Company, 1990.

[16] Michael Lässig, Ville Mustonen, and Aleksandra M Walczak. Predicting evolution. *Nature Ecology & Evolution*, 1:0077, 2017.

[17] Jonathan B Losos. Convergence, adaptation, and constraint. *Evolution*, 65(7):1827–1840, 2011.

[18] Marie Manceau, Vera S Domingues, Catherine R Linnen, Erica Bree Rosenblum, and Hopi E Hoekstra. Convergence in pigmentation at multiple levels: mutations, genes and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552):2439–2450, 2010.

[19] Real Michaud, WF Lehman, and MD Rumbaugh. World distribution and historical development. *Alfalfa and alfalfa improvement*, (alfalfaandalfal):25–91, 1988.

[20] Simon Conway Morris. *Life's solution: inevitable humans in a lonely universe*. Cambridge University Press, 2003.

[21] Terry J Ord and Thomas C Summers. Repeated evolution and the impact of evolutionary history on adaptation. *BMC Evolutionary Biology*, 15(1):137, 2015.

[22] Virginie Orgogozo. Replaying the tape of life in the twenty-first century. *Interface focus*, 5(6): 20150057, 2015.

[23] H Allen Orr. The probability of parallel evolution. *Evolution*, 59(1):216–220, 2005.

[24] Cynthia F Scholl, Chris C Nice, James A Fordyce, Zachariah Gompert, and Matthew L Forister. Larval performance in the context of ecological diversification and speciation in Lycaeides butterflies. *International Journal of Ecology*, 2012, 2012.

[25] David L Stern and Virginie Orgogozo. Is genetic evolution predictable? *Science*, 323(5915): 746–751, 2009.

[26] Troy E Wood, John M Burke, and Loren H Rieseberg. Parallel genotypic adaptation: when evolution repeats itself. In *Genetics of Adaptation*, pages 157–170. Springer, 2005.

CHAPTER 2

DOES HISTORICAL ADMIXTURE PREDICT PATTERNS OF INTROGRESSION IN A CONTEMPORARY HYBRID ZONE? INSIGHTS FROM *LYCAEIDES* BUTTERFLIES

Introduction

Hybridization, or interbreeding between diverged taxa, is increasingly being recognized as an important and ubiquitous process in evolution of species [1, 2, 3, 4]. Several cases of introgression and hybridization have been now documented across various organisms such as plants, animals, microbes and humans. Hybridization is an important evolutionary phenomenon as it can lead to reinforcement of incompatibilities through prezygotic isolation and thereby help complete the speciation process [5, 6]. Hybridization can also provide genetic variation for adaptive evolution by introducing new combinations of genotypes into the species gene pool [2, 3, 5, 6]. Therefore, understanding patterns of hybridization (and introgression) can help understand the processes of speciation and diversification. Natural hybrid zones offer an ideal opportunity to study hybridization. Fine scale genetic mapping in hybrid zones can help identify genomic regions which have undergone generations of recombination and contribute to reproductive isolation or are under selection [7]. This temporal and spatial tracking of genomic changes cannot be achieved in laboratory experiments for organism with longer life spans. Although we have started to learn more about the role of hybridization in evolution and its implications in speciation by studying patterns of introgression and hybridization, we still lack knowledge about the process of hybridization in natural populations and how the reproductive isolation barriers are maintained between hybrids and their parental taxa. Two ways in which this gap in our knowledge can be filled is by understanding the processes which drive patterns of evolution in hybrid zones and understanding how these patterns are shaped over time.

Differential introgression across the genome is common in hybridizing taxa as has been highlighted by both geographic and genomic clines analyses in hybrid zones [6, 8]. Studies have also highlighted how markers involved in restricted introgression are present in excess on specific regions of the genome, such as the X (Z) chromosome [7, 8, 9, 10, 11, 12, 13, 14]. Similarly, genomic differentiation can help identify regions under divergent selection in populations. However, these

patterns of differential introgression or differentiation can be difficult to interpret mainly due to the unclear role of selection in driving variation across the genome. When considering allopatric populations, recent selection or variable recombination rates can shape diverged genomic regions which need not reflect resistance to introgression [8]. For example, variable recombination rates have led to increased divergence near chromosome centers in replicate pairs of stream and lake sticklebacks [15], and led to increased genome divergence in *Ficedula* flycatchers [16], suggesting that heterogeneous selection may be driving reproductive isolation. Taxa undergo repeated instances of allopatry and secondary contact, and regions of divergence could arise due to sorting of ancestral variation such that selection could drive allele frequency differences by acting on standing genetic variation [8]. Therefore, these regions could be shaped by heterogeneous selection or recombination rates and restricted introgression (steep clines) will provide clear signatures of genomic regions involved in speciation or reproductive isolation [6, 8]. In addition, patterns of ancestry can also be difficult to interpret as it is not always plausible to parse the role of selection versus drift in causing variation ancestry frequencies in hybrids. As population size and timing of admixture can affect recombination rates, caution should be observed while interpreting high ancestry frequencies as evidence of selection [17]. One way in which these issues can be resolved is by comparisons between independent admixed populations or between ancient and contemporary hybrids. This way hybrid zone studies can help distinguish between the role of selection versus stochastic processes in driving patterns of introgression, especially when selection causes similar/parallel patterns of change across time.

Along these lines, comparisons among multiple independent hybrid zones or transects can provide information about whether there is concordance in patterns of genomic introgression and can help connect patterns with processes. Parallel evolution, or the repeated evolution of genotypes or phenotypes in different populations, provides evidence of selection [18, 19, 20, 21, 22]. Several examples of parallel phenotypic evolution exist which demonstrate that similar ecological pressures can promote parallel phenotypic evolution (For example, armor plating in sticklebacks, [23]). Similarly, parallel genetic divergence can occur across ecotypes in patchy/mosaic habitats (For example, stick insects [24] and *Littoria* snails [25]). In hybrid zones, spatial comparisons between indepen-

dent transects have revealed variable concordance in genomic evolution. For example, patterns of introgression between two species of *Helianthus* are highly consistent [26, 27], but studies on independent hybrid zones between fish species/morphs reveal little concordance or partial parallelism [28]. Either way, consistent patterns of locus-specific introgression between two geographic regions of the hybrid zone highlight how specific genomic regions contribute to pre-zygotic barriers that isolate species irrespective of local ecological selection, population structure or ambiguity in gene flow and divergence. While replicate hybrid zones in different locations can exhibit repeatable patterns of introgression, we do not know if selection could consistently shape genomic patterns in a hybrid zone at different time points over a large temporal scale. For example, can we ask if selection can consistently act on barrier loci in ancient admixed hybrids with no ongoing hybridization (ancient hybrids) and contemporary hybrids experiencing ongoing hybridization (recent hybrids)?

In this study, we make use of a well studied hybrid zone in *Lycaeides* butterflies to ask whether consistent selection pressures shape the process of hybridization over long periods of time. Hybridization has been crucial in the evolution of *Lycaeides* with several important outcomes, with a unique scenario where ancient and contemporary hybrids coexist and show ongoing gene flow. By comparing genomic patterns of admixture in contemporary and ancient hybrids, we attempt to quantify the predictability of genome-wide patterns of admixture and inrogression. *Lycaeides idas* and *Lycaeides melissa* are two of the 5 nominal species of *Lycaeides* butterflies that occur in North America [29, 30]. These species likely diverged from their European ancestor 2.4 million years ago [31] and differ in various aspects such as male genitalic morphology [32], wing patterns [33], host plant use, phenology and behavior. Along these lines, there is evidence that *Lycaeides* come into secondary contact in various regions across their geographic range and have hybridized to produce three additional admixed lineages [34, 35, 36, 37]. One such hybrid lineage is formed between *L. melissa* and *L. idas*, in North Western Wyoming, specifically in Jackson Hole and Gross Ventre mountain region and we call this lineage Jackson Hole-*Lycaeides*[30, 37, 38]. We have shown that Jackson Hole-*Lycaeides* are similar to *L. idas* in their genomic composition, have experienced extensive introgression from *L. melissa*, and experience limited or no gene flow with parental species [30, 39]. In addition, locus specific ancestry estimates suggest that several loci in the genome have

been fixed for chromosomal ancestry blocks [30]. Low levels of linkage disequilibrium and narrow hybrid indices suggest that Jackson Hole-*Lycaeides* experience no gene flow with parental species and harbor genomic regions which are evolving without independently from their parental species. These populations formed 15,000 years ago and are in the process of genome stabilization post hybridization [40](Also see Figure 2.1 for ancient hybrids).

Jackson Hole-*Lycaeides* have a genomic composition somewhat similar to nearby *L. idas* populations but show evidence of introgression from *L. melissa*. In contrast, individuals from a single locality, Dubois, WY, show variable genetic ancestry patterns (hereafter referred to as Dubois-*Lycaeides*). Dubois-*Lycaeides* show high levels of admixture which is indicative of ongoing gene flow between parental populations and therefore, this region is a case of recent or contemporary hybridization [39]. Dubois is situated close to another Jackson Hole-*Lycaeides* locality where the individuals utilize *Astragalus miser* as their host plant and a *L. melissa* locality where the host plant used is *Medicago sativa*. The population at Dubois is restricted to a geographic space of 5 kilometers and occupies host plant patches along the road sides. Dubois-*Lycaeides* inhabit *L. melissa* like environment and feed on *M. sativa*. Therefore, hybridization in Dubois-*Lycaeides* is ongoing and individuals here could be putative hybrids between Jackson Hole-*Lycaeides* and *L. melissa* [39] (Also see Figure 2.1 for contemporary hybrids).

For this study, we used Jackson Hole-*Lycaeides* as a case of ancient hybridization and Dubois-*Lycaeides* as a case of ongoing or contemporary hybridization to ask if we can predict contemporary patterns of introgression from ancient admixture. We used a high quality *Lycaeides melissa* genome and DNA sequence data from butterflies from 23 localities, to answer the following specific questions for our study: 1) Is Dubois, WY an active hybrid zone between *L. melissa* and Jackson Hole-*Lycaeides*? 2) To what extent does introgression vary across the genome in Dubois, WY hybrid zone? 3) To what extent does genetic ancestry vary across ancient Jackson Hole-*Lycaeides*? and 4) Can we predict regions which are most resistant to gene flow in contemporary hybrids from patterns of ancestry in the ancient hybrids and are these regions enriched for specific functional properties? We mainly hypothesize that natural selection rather than stochastic processes can drive predictable patterns of evolution in *Lycaeides* hybrids and that putative targets of selection lie on

specific genomic regions and are enriched for important biological functions.

Materials and Methods

Experimental Design

The main objective of this study was to predict patterns of introgression in a contemporary hybrid zone from historical admixture in *Lycaeides* butterflies. In addition, we wanted to quantify the degree of genomic predictability between these two cases of hybridization. The analytical framework of this study included partial genome sequencing of *Lycaeides* butterflies from key localities in the Jackson Hole region (No. of populations = 23, No. of individuals = 835), mapping these to a high-quality genome reference genome assembly of *Lycaeides melissa* and conducting statistical analyses to quantify predictability of genomic patterns of introgression between contemporary and ancient hybrids.

Samples and DNA sequencing

We used genotyping-by-sequencing (GBS) data from 835 *Lycaeides* butterflies sampled from 23 localities to include four species groups: a). *Lycaeides melissa* (N = 9), b). *Lycaeides idas* (N = 6), c). Jackson Hole-*Lycaeides* (N =7), and d). Dubois-*Lycaeides* (N =1) (also see Table 2.1 for number of individuals sampled from each locality). GBS data from 643 of these individuals was published previously in a study of admixture in the *Lycaeides* species complex [39]. Data for the remaining 192 individuals is presented here. For the data presented here, we conducted sampling at 4 localities (Table 2.1). We conducted DNA extraction, GBS library preparation and DNA sequencing to generate 100 bp single-end reads with an Illumina HiSeq 2500 following previously described approach [41].

Genome assembly and annotation

We annotated the structural and functional genetic elements in the de novo *L. melissa* genome using the MAKER pipeline (version 2.31.10) [42, 43]. This pipeline uses repeatmasking, protein and RNA alignment and ab initio gene prediction to perform evidence based gene prediction

which generates annotations which are supported by quality scores. Prior to using MAKER, we identified de novo repeats in the *L. melissa* genome using REPEATSCOUT (version 1.0.5) [44]. This program identifies repeat elements, tandem repeats and low complexity elements and removes them from the genome. We took this approach since every genome will have some repeat elements which can go unidentified. This filtered repeats library was supplied to MAKER which uses this along with Repbase in REPEATMASKER to conduct repeat masking of the genome. MAKER also requires protein sequence and transcriptome data for alignment. Since we lacked protein sequences for *L. melissa*, we downloaded 28 protein homology files of 15 butterfly species (species listed in Table 2.16) from LepBase (Version 4) [45] and concatenated the fasta files to create a master protein homology file for MAKER. We used RNA/transcriptomic data from 24 *L. melissa* samples (unpublished data, manuscript in prep.). We first used Trim_galore (version 2.6.6, https://github.com/FelixKrueger/TrimGalore) for adapter trimming and quality filtering of paired-end RNAseq reads. We then used these trimmed reads to conduct de novo transcriptome assembly using Trinity (version 2.6.6) [46, 47], which was used for genome annotation through MAKER. We ran two rounds of MAKER. We first ran MAKER without using any information from ab initio gene predictors such as AUGUSTUS, to generate de novo gene models for our genome. We then ran a second round of MAKER wherein we used the gene models from the first run to train two gene prediction softwares: AUGUSTUS and SNAP. We ran SNAP (version 2006-07-08) [48] by using models with AED scores of 0.25 or better and length of 50 or more amino acids. We ran AUGUSTUS (version 3.3) with the insecta predictions [49]. We then used both these gene predictions for our second MAKER run. We used the MAKER output to generate functional annotations for our genome. We assigned putative gene functions by using blastp command to query the MAKER output against UNIPROT/SWISSPROT database. We also used Interproscan to functionally annotate the files to add interproscan and gene ontology information to each annotation. Finally, we used a custom python script to annotate the SNP dataset for this study (N = 39,193) using the annotation information generated above. GO and IPR terms were assigned to SNPs within 1 kb of annotated genes. We used this SNP annotation to conduct randomizations for structural and functional enrichment of outlier loci in analyses described below.

Genome alignment and detecting genetic variation

For the 192 samples sequenced in the present study, we filtered the sequences for individual barcodes and then split them by individual using custom perl and python scripts. We then combined this data with the remaining individuals data to perform alignment and variant calling. We used BWA version 0.7.17 to align the GBS sequences from 835 individuals to the draft *L. melissa* genome by using MEM and SAMSE algorithms to compress, sort and index the alignments [50]. Here we used 12 threads and minimum seed length of 15 to generate this alignment. We then used SAMTOOLS (version 1.5) to compress, sort and index the alignments [51]. We conducted variant calling SAMTOOLS and BCFTOOLS (version 1.6) for variant calling and retained genetic variants for which we had sequence data for at least 80% of the sampled individuals and where posterior probability of the sequence data under a null model that the nucleotide was invariant was <0.01. Following this, we used a median sequencing depth of 2 per individual per variable site to filter variants. We finally used genotype estimates from BCFTOOLS to calculate population allele frequencies using an expectation-maximization algorithm to obtain maximum likelihood estimates while accounting for uncertainty in genotypes. In the end, we identified 39,193 SNPs which we used for all further statistical analyses. We repeated the genome alignment and variant calling with only male individuals in our dataset (N = 479).

Population genetics analyses and identifying contemporary admixture in Dubois,WY

We used a hierarchical Bayesian model, entropy, to analyse population genetic structure and admixture across 23 populations for 39,193 SNPs [39]. Entropy uses a model similar to the correlated admixture model in structure [52], but allows for variation in sequence coverage, sequencing error, and alignment error in the model itself. Similar to structure, entropy uses multilocus genotype data and a given number of k ancestral populations or clusters, to estimate admixture proportions for each individual. These admixture proportions denote the proportion of an individuals genome which is inherited from each of the k ancestral populations. In this way, entropy provides three outputs: admixture proportions, genotype probabilities of all individuals at all loci and credible intervals for all estimated parameters.

We ran entropy for k = 2 to k = 5 putative populations, with 3 chains for each k. Our approach was to speed convergence of MCMC analyses by providing starting values which do not constrain the posterior. For this, we first generated point estimates of genotypes by using genotype likelihood values from BCFTOOLS (version 1.6). We then calculated a genotype likelihood matrix of all individuals and performed a principal component analysis using the prcomp command in R [53]. We used the PCA results to perform k-means clustering (k means in R) and linear discriminant analysis (lda in R) to generate appropriate starting values for q for each individual value for each value of k to initialize MCMC and assure proper mixing and convergence of chains. We ran three chains with 15,000 MCMC iterations with a thinning interval of 5 and discarded the first 5000 values as burn-in, to generate the posterior probability distributions for admixture proportions and genotype probability. We plotted MCMC steps for a subset of parameter estimates to check for mixing of chains, stabilization and convergence of parameter estimates. We estimated posterior means, medians and credible intervals for the parameters of interest.

We used the genotype estimates from entropy to calculate a genotype covariance matrix for all individuals by taking a mean of genotype probabilities across all k values (k =2 to 5). We used the matrix generated for all SNPs (N = 39,193) to perform principal component analysis using the prcomp function in R and summarized the genotypic variation across all individuals and across all populations [53].

The PCA analysis was also an ordination-based approach to answer our first focal question for this study which is to confirm if Dubois-*Lycaeides* are experiencing recent or ongoing hybridization. We examined whether the 115 individuals from Dubois, WY form a single cluster in genotype space or show variation in their genomic composition. Results from this analyses (described in detail in the Results section) revealed that Dubois-*Lycaeides* show extreme variation in genomic composition. Therefore, we classify this population as contemporary hybrids and use them for further analyses to answer the focal questions of our study.

Analysis of introgression in contemporary hybrids at Dubois, WY

To test if individuals from Dubois, WY are undergoing recent hybridization, we used a genomic clines approach. This approach helped us quantify genome-wide variation in locus-specific

introgression among Dubois-*Lycaeides*. For this analysis, we used the Bayesian genomic cline (bgc) model [30, 54]. This model also identifies outlier loci which can be potential candidates for reproductive isolation as compared to the rest of the genome [54]. This model mainly includes two key parameters, α and β, which describe the probability of inheriting a gene copy at a locus from parent 1, given an individualâĂŹs hybrid index and that the average probability of parent 1 ancestry is equal to the individualâĂŹs hybrid index. Assuming there are two parental populations (0 and 1), genomic cline parameter α indicates an increase (positive α) or decrease (negative α) in the locus specific ancestry from parent 1 and specifies the center of the cline. The second parameter, β, indicates an increase or decrease in the rate of transition in the probability of parent 1 ancestry as a function of hybrid index. Changes in the cline parameter β can lead to a widening (positive β) or narrowing (negative β) of the cline. This parameter is also a measure of the average pairwise linkage disequilibrium between a locus and all other loci based on ancestry. Both parameters have been shown to be affected by selection on specific hybrid loci such that extreme values of α are suggestive of underdominance or Dobzhansky-Muller incompatibilities. On the other hand, extreme values of β are suggestive of population structure in the hybrid zone, strong selection against hybrids and/or gene flow from parental populations. We implemented bgc by using MCMC to estimate hybrid indices and both genomic cline parameters.

We used geographical proximity and genotypic clustering patterns from the PCA analysis to combine individuals from 2 Jackson Hole-*Lycaeides* localities and 3 *L. melissa* localities, to be used as potential parental populations 0 and 1. We used Bald Mountain, WY and Frontier Creek, WY as potential Jackson Hole-*Lycaeides* parental populations (parental population 0; n =94) and Lander, Sinclair and Cokeville as potential *L. melissa* parental populations (parental population 1; n =131) for admixed Dubois-*Lycaeides* (N = 115). We used genotype likelihoods from BCFTOOLS to calculate genotype point estimates (as described in the entropy section) to subset the three classes of ancestry informative markers (AIMS). We defined three classes of AIMS based on markers which had absolute allele frequency differences greater than 2% (N = 2126) and 3% (N = 1164). We ran bgc analyses for both these classes of AIMS. For each class, we ran five independent chains of 25,000 MCMC steps with a 5000 step burn-in and recorded samples from the posterior distribution

every 5th step. We inspected the MCMC output to assess convergence of chains to the stationary distribution and combined the output of the five chains. We repeated the analyses with only male individuals to compare patterns of introgression.

Analysis of population ancestry in Jackson Hole-*Lycaeides*

We used `popanc` to estimate local ancestry frequencies in individuals from ancient hybrid populations (No. of populations = 10, No. of individuals = 250). `popanc` provides an ideal model to estimate local ancestry frequencies in our focal populations since it aims to calculate ancestry frequencies in isolated populations in which admixture has occurred already but the ancestry blocks in the genome are still segregating. We already know that Jackson Hole-*Lycaeides* do not experience any gene flow from the parental species and have portions of the genome which have begun to stabilize with certain blocks of ancestry being fixed in the genome. Therefore, we used `popanc` to quantify ancestry frequencies across the genome in Jackson Hole-*Lycaeides*. `popanc` uses a combination of discriminant analysis and a continuous correlated beta process model to estimate local ancestry within individuals and at population-level. In addition, this method uses SNP windows within individuals to calculate autocorrelations in ancestry frequencies at the population level. Tests on simulated datasets and on human admixed datasets indicate that this method outperforms the traditional HMM linkage model in structure and is reliable to infer patterns of local ancestry in our dataset.

We ran `popanc` for three windows (3, 5 and 7) and focused on two AIMS classes. We used Soldier creek, Siyeh creek and King's hill as potential *L. idas* parents and Bonneville shoreline trail, Sinclair, Cody and Cokeville as potential *L. melissa* parents. We ran the analysis for 10,000 MCMC steps with a 5000 step burn-in and recorded samples from the posterior distribution every 5 step. We used a thinning-interval of 10. We repeated the analyses with only male individuals to compare patterns of ancestry.

Statistical Analyses: Quantifying predictability of genomic changes in contemporary hybrids from old hybrids

We measured and quantified predictable genomic changes associated with ancient and contemporary hybrids by asking how well genomic regions which are most resistant to gene flow in contemporary hybrids or Dubois-*Lycaeides* can be predicted from patterns of ancestry in Jackson Hole-*Lycaeides*. We did this by first identifying and quantifying excess overlap in SNPs which show exceptional cline parameter values (high α, high β and low α) from genomic cline analysis in Dubois-*Lycaeides* and also show extreme *L. idas* ancestry frequencies in Jackson Hole-*Lycaeides*. Therefore, we made three sets of comparisons: a). SNPs with high α values in Dubois-*Lycaeides* and extreme *L. idas* ancestry in Jackson Hole-*Lycaeides*, b) SNPs with low α values in Dubois-*Lycaeides* and extreme *L. idas* ancestry in Jackson Hole-*Lycaeides*, and c) SNPs with high β values in Dubois-*Lycaeides* and extreme *L. idas* ancestry in Jackson Hole-*Lycaeides*. We report the excess overlap values as x-fold enrichments. For example, an x-fold enrichment of 2.0 would imply that twice as many SNPs show excess restricted introgression in contemporary hybrids and exceptional ancestry patterns in old hybrids, than expected by chance. This will therefore mean that we can predict exceptional patterns of genomic change from old hybrids in contemporary hybrids twice as well as compared to a model with no information. For this analyses, we focused on the top 0.1% ancestry informative markers (AIMS) by retaining markers which differed in allele frequency between the ancestral populations (N = 1164).

Gene enrichment and ontology analyses

In addition to identifying outlier loci for each analyses described above, we conducted additional tests to ask whether outlier SNPs are present in excess on the Z chromosomes, or whether they are enriched for presence on specific structural regions. We were also interested in the overall distribution of exceptional loci throughout the genome. We performed these tests for a range of quantiles (top 0.01% to top 0.1%). We conducted these tests for three cases: (i) outlier loci associated with reproductive isolation in Dubois-*Lycaeides*, (ii) outlier loci showing exceptional ancestry patterns in Jackson Hole-*Lycaeides*, and (iii) loci associated with predictable genomic changes associated

with Dubois-*Lycaeides* and Jackson Hole-*Lycaeides*. For this, we conducted enrichment tests across a range of empirical quantiles and considered a range of cut-offs, from the top 0.01% to the top 0.1% SNPs (with increments of 1%). We quantified over-representation of loci on different regions by x-fold enrichments. We used a linkage map to classify SNPs as Z-linked or autosomes. We then used the structural annotation information we generated using MAKER (see above), to classify SNPs as on coding regions (genes, mRNA or CDS), on transposable elements or on proteins or near coding regions, near transposable elements or near proteins. Randomization tests were conducted to quantify and assess the significance of enrichments for each quantile cut-off and all three genomic clines parameters (high α, low α, and high β), population ancestry estimates and excess overlap. We conducted 10,000 randomizations for each case. In addition to structural annotation, we also identified the functional annotations for SNPs which show exceptional cline parameter values (high α, high β and low α) from genomic cline analysis in Dubois-*Lycaeides* and also show extreme *L. idas* ancestry frequencies in Jackson Hole-*Lycaeides* in the top 0.1% quantile range. We describe these results for each analyses seperately.

Results

Genome assembly, annotation and GBS sequence alignment

We identified that annotated regions were spread across 1651 scaffolds. We identified 11247 putative genes, 48765 putative coding sequences, 51464 matches to exon, 8893 UTR sequences. We had 340568 protein matches and 84004 matches to expressed sequence tags.We used genome sequence data from 835 individuals of *Lycaeides* and identified 39,193 candidate single nucleotide polymorphisms (SNPs) which we use for analyses to answer our focal questions. We annotated these SNPs using the genome annotation information generated using the MAKER pipeline. For this set of SNPs, 11569 SNPs were located on putative genes and 2468 SNPs were located near genes, 15043 were located on exons and 10070 were located near exons, 4786 were located on CDS and 9612 were located near CDS, 21505 were located on UTRs and 21134 were located near UTRs, 2468 were locate don proteins and none near proteins.

Population structuring and contemporary hybrids

We implemented the admixture model in `entropy` to estimate posterior probability distribution of admixture proportions (q), for 39,193 SNPs for 23 populations (No. of individuals = 835). We selected k = 2 model to interpret our results using estimates from q. Our results revealed, that under k = 2 model, each genetic cluster corresponds to a identified nominal species (*L. melissa* and *L. idas*) or an identified admixed lineage (Jackson Hole-*Lycaeides*), with Dubois-*Lycaeides* showing variable patterns of admixture. We then conducted principle component analysis (PCA) of average genotype likelihood estimates from entropy model (average likelihood estimates across k = 2 to 5), for 39,193 SNPs for these populations. The first two principal components accounted for most of the genetic variation in the samples (3.3%) (Figure 2.2B). The PCs revealed striking pattern of population structuring by separating entities mainly based on hybrid ancestry. Based on PC1 scores, all entities formed distinct clusters and were separated from each other. PC2 scores separated *L. idas* individuals and *L. melissa* individuals. Jackson Hole-*Lycaeides* occupied intermediate PC space between the two parental taxa (*L. idas* and *L. melissa*). Interestingly, Dubois-*Lycaeides* were dispersed across PC1 and PC2 space and had scores intermediate between Jackson Hole-*Lycaeides* and *L. melissa*. These results were consistent with our prediction that Jackson Hole-*Lycaeides* have more constrained and intermediate ancestry without recent back-crossing or ongoing hybridization with parental taxa (Figure 2.2C). On the other hand, Dubois-*Lycaeides* showed clustering patterns indicative of ongoing introgression or backcrossing between Jackson Hole-*Lycaeides* and *L. melissa* (Figure 2.2C).

Delineating candidate SNPs for restricted introgression in Dubois,WY

We used Bayesian genomic clines (bgc) method to test for variation in locus-specific introgression across the genome in Dubois-*Lycaeides* [30, 54]. We conducted this analyses on ancestry informative markers or AIMS, which are essentially SNPs which show difference in allele frequencies between the ancestral populations (or parents). Here, we present the results of AIMS which had absolute allele frequency difference greater than 3% (N = 1164). Our results revealed that introgression was variable among genetic loci in Dubois-*Lycaeides*. Genomic clines were really steep for

several loci, suggesting that these loci exhibited restricted introgression (high β, N = 39), with fewer loci with clines indicating a higher rate of introgression (low β, N = 57) (Figure 2.3B). Genomic clines for several other loci indicated an excess of *L. melissa* ancestry in admixed Dubois-*Lycaeides* (high α, N = 197), while a few other loci show excess Jackson Hole-*Lycaeides* ancestry (Low α, N = 295) (Figure 2.3B).

We were also interested in determining the distribution of loci with exceptional genomic cline parameters across the genome, to see if loci showing restricted introgression are overrepresented on certain regions of the genome, enriched to be on coding regions of the genome. We saw that loci with high cline parameter values were distributed across different linkage groups with several loci present on the Z chromosome (Figure 2.3C and 2.3D). We quantified the presence of excess number of SNPs with exceptional cline paramters on the Z chromosome. We conducted randomization tests to calculate x-fold enrichments based on null expectations for the proportion of SNPs with high β or high α values present on a specific linkage group. We found that for the top 0.1% quantile cut-off (N =117), loci with exceptional cline parameter values were spread across all linkage groups. Loci which show excess *L. melissa* ancestry, were spread across all linkage groups with a significant excess on the Z chromosome which was almost 1.5 time more than expected under random chance (high α; N = 32, x-fold = 1.42, p-value = 0.011, Figure 2.3C). Similarly, loci with high cline parameter β values were spread across all chromosomes with almost two time more SNPs on Z chromosome than expected under random chance (high β; N = 47, x-fold = 2.07, p-value = < 0.001, Figure 2.3D). We did not find a significant excess of loci with low α values on Z (N = 19, x-fold = 0.84, p-value = 0.843). We saw several loci (across a range of quantiles 0.1 - 0.01 %) with high genomic cline parameter values were present on and near coding regions, transposable elements and proteins. However, we saw limited significant enrichments across quantiles for all three genomic clines parameters (Table 2.3, 2.4 and 2.5).

Delineating candidate SNPs with excess local ancestry frequencies in old hybrids

We used `popanc` to quantify local ancestry frequencies in 10 Jackson Hole-*Lycaeides* populations [17]. We again focused on AIMs and here we present our results for AIMS which had absolute allele frequency difference greater than 3% (N = 1164). Patterns of *L. idas* ancestry were consistent

across the 10 populations. Across all populations, there were several loci which showed high or fixed *L. idas* ancestry and these loci were distributed across all linkage groups (Figure 2.4A and 2.4B). Within populations, ancestry frequencies varied with a mean low of 0.41 and a mean high of 0.82. Within linkage group, mean ancestry was variable but individuals from Bunsen Peak had highest mean ancestries for every linkage group (Figure 2.4C).

Loci showing excess *L. idas* ancestry in Jackson Hole-*Lycaeides* were distributed across the different linkage groups. We conducted randomization tests to calculate x-fold enrichments to test if observed significant excess of SNPs were more than expected by random chance. For the top 0.1% quantile, we found a significant excess of SNPs on Z chromosome with x-fold enrichments ranging from 1.28 to 3.09 (Table 2.2). For example, for Bald Mountain, we observed 66 SNPs on Z chromosome (x-fold = 2.92, p-value = <0.001). Similar results were seen for other populations when considering top 0.1% SNPs with excess *L. idas* ancestry (Table 2.2). Several loci across the range of quantiles were present near coding regions and proteins. For example, for the top 0.1% SNPs, we saw that SNPs were present in excess near gene (observed = 63, x-fold = 1.20, p-value = 0.024), near CDS (observed = 55, x-fold = 1.22, p-value = 0.024), near mRNA (observed = 63, x-fold = 1.20, p-value = 0.021) and near proteins (observed = 84, x-fold = 1.14, p-value = 0.021) (Table 2.6).

Predictability of contemporary hybrids from old hybrids

We conducted randomization tests to calculate x-fold enrichment to test for greater overlap than expected by chance between SNPs most associated with differential introgression in Dubois-*Lycaeides* and those associated with excess *L. idas* ancestry in Jackson Hole-*Lycaeides* (here again we consider results for N = 1164 SNPs). We considered SNPs which show exceptional values of three cline parameters in Dubois-*Lycaeides* (ie high α, low α, and high β). Therefore, we had three sets of comparisons. We also conducted randomizations for a range of quantiles (top 0.01% to top 0.1%) to quantify predictability at each quantile. We found almost two times more overlap between SNPs with high α values in Dubois-*Lycaeides* and those with high *L. idas* ancestry means in Jackson Hole-*Lycaeides* (obs = 22, x-fold = 1.91, p-value = 0.0007; Figure 2.5A). We also saw an increase in x-fold values as the quantile cut-off decreased with an x-fold enrichment as high as 16 for SNPs

in the top 0.01% quantile (x-fold range across quantiles; 1.90 –16.93, Figure 2.5D). Similarly, we found almost four times more overlap between SNPs with high β values in Dubois-*Lycaeides* and those with high *L. idas* folded ancestry values in Jackson Hole-*Lycaeides* (obs = 49, x-fold = 4.23, p-value = 0.001; Figure 2.5C). Here also we saw an increase in x-fold values as the quantile cut-off decreased with an x-fold enrichment as high as 35 for SNPs in the top 0.01% quantile (x-fold range across quantiles; 4.23 – 35.42, Figure 2.5D). Alternatively, we did not see a significant overlap between SNPs with low α values in Dubois-*Lycaeides* and those with high *L. idas* ancestry values in Jackson Hole-*Lycaeides* (obs = 9, x-fold = 0.78, p-value = 0.842; Figure 2.5B and 2.5D).

We then focused on the distribution of shared SNPs across various linkage groups and were interested to compare the degree of excess enrichment of shared SNPs on the Z chromosome versus autosomes. We again conducted randomization tests to calculate x-fold enrichments for shared SNPs between Dubois-*Lycaeides* and Jackson Hole-*Lycaeides* (again made three sets of comparison). Again, we conducted tests for a range of quantiles (top 0.01% to top 0.1%) as we were interested in quantifying predictability at each quantile. Our results show significant and modest excess of shared SNPs on the Z chromosome (Figure 2.5E). For the top 0.1% quantile, we see three times more enrichment of SNPs on the Z chromosome for SNPs with high α (obs = 12, x-fold = 2.82, p-value = 0.0003). Across the quantile range we saw a significant xfold enrichment in the range of 2.83 – 5.14 (Figure 2.5E). Similarly, for overlapping SNPs with high β in Dubois-*Lycaeides* and high folded population ancestry in Jackson Hole-*Lycaeides*, for the top 0.1% quantile, we see almost four times more enrichment of SNPs on the Z chromosome for SNPs with high β (obs = 37, x-fold = 3.88, p-value = 0.0001). Across the quantile range we saw a significant xfold enrichment in the range of 3.88 – 5.10 (Figure 2.5E). We again do not see signficant enrichments of SNPs with low α values in Dubois-*Lycaeides* and high *L. idas* ancestry in Jackson Hole-*Lycaeides* (2.5E). We then quantified predictability across autosomes to compare patterns across Z and autosomes. For all autosomes, our results show presence of significant excess of shared SNPs. (Figure 2.5F). For cline parameter high α, for the top 0.1% quantile, we see almost two times more enrichment of SNPs on the autosomes (obs = 21, x-fold = 2.23, p-value = 0.0001). Across the quantile range we saw a significant xfold enrichment in the range of 2.225 – 0 (Figure 2.5E).Similarly, for cline parameter high β, for the top

0.1% quantile, we see almost two times more enrichment of SNPs on the autosomes (obs = 23, x-fold = 2.44, p-value = 0.0001). Across the quantile range we saw a significant xfold enrichment in the range of 35.42 - 2.44 (Figure 2.5E). We again do not see significant enrichments of SNPs with low α values in Dubois-*Lycaeides* and high *L. idas* ancestry in Jackson Hole-*Lycaeides*. These results indicate that although we see higher repeatability on Z chromosome, the signal is not restricted to just this region of the genome also experience selective pressures. In addition, these results indicate genomic concordance between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides*. We repeated these randomization tests for the AIMS2 category of SNPs (N = 2126) and saw similar patterns of x-fold enrichments for shared SNPs and SNPs lying on autosomes with significant excess of shared SNPs on Z chromosome (Figure 2.7). We obtained similar results when only considering males, which are homozygous for the Z chromosome (Figures 2.11, 2.12).

We then tested if the shared SNPs across quantiles were present in signficant excess on any structural regions and if they were annotated for specific IPR or GO terms. Such enrichments might be expected if the top predictability associated SNPs were indeed tagging (via LD) genetic regions affecting introgression or differentiation. Several shared SNPs across the range of quantiles were present on and near coding regions, transposable elements and proteins. However, we did not see any significant enrichments across quantiles for our three comparisons for predictablity (Tables 2.7, 2.8, 2.9). We then looked at the functional annotations of the shared SNPs. In the top 0.1% quantile, the 22 overlapping SNPs with high α in Dubois-*Lycaeides* and high population ancestry means in Jackson Hole-*Lycaeides* had 12 unique interproscan (IPR) and 7 unique GO ontology categories (Tables 2.10, 2.11). The 49 overlapping SNPs with high β in Dubois-*Lycaeides* and high folded population ancestry in Jackson Hole-*Lycaeides* are enriched for and 17 unique interproscan (IPR) categories and 12 unique GO ontology categories. The 9 overlapping SNPs with low α in Dubois-*Lycaeides* and high population ancestry means in Jackson Hole-*Lycaeides* are enriched for and 4 unique interproscan (IPR) categories and 4 unique GO ontology categories (Tables 2.12, 2.13).

Discussion

We demonstrate consistent patterns of introgression in ancient and contemporary hybrids in *Lycaeides* butterflies We identify patterns of excessive ancestry and candidates for introgression

in regions involved in the process of reproductive isolation through a combination of population ancestry and genomic clines analyses and quantify predictability using randomization tests. Our results suggest that natural selection is the main process behind concordant patterns of genomic divergence and the evolution of hybrids of *Lycaeides* butterflies. We interpret these results below in the context of processes driving patterns of hybridization and how concordant patterns of evolution are crucial in our understanding of biodiversity.

Interpretations of patterns of ancestry and differential introgression

Differential genomic introgression in contemporary hybrids

Dubois-*Lycaeides* show patterns of differential introgression across the genome. This is supported by wide ranges of hybrid indices and several loci showing high values for genomic cline parameters α and β. These loci are also spread across chromosomes. These results are consistent with previous evidence from genetic ancestry analyses, which suggested that hybridization in Dubois-*Lycaeides* is recent and ongoing [39]. The wider hybrid indices indicate that this population is currently evolving with ongoing gene flow from Jackson Hole-*Lycaeides* and *L. melissa*. Genome-wide estimates of genomic cline parameter β and steep genomic clines for this parameter suggest that a subset of loci are under selection, avoid introgression and are possible candidates for reproductive isolation. Genetic drift and variable recombination rates could drive these steep clines since Dubois-*Lycaeides* has a smaller populations size, but we suspect that it is intrinsic selection rather than extrinsic selection that is driving these patterns. This can be explained by the variable environments occupied by Dubois-*Lycaeides* and one of the parental population from Jackson Hole-*Lycaeides* (Bald Mountain, WY). If extrinsic or environmental selection pressure were affecting these patterns, the difference in habitats between Dubois-*Lycaeides* and Bald Mountain population would matter. Following this, we interpet that Dobzhansky-Mueller incompatibilities act on the portion of the genome inherited from *L. melissa* in Dubois-*Lycaeides* and *L. melissa* alleles are favored. Dubois-*Lycaeides* offer an interesting contrast with Jackson Hole-*Lycaeides* for patterns of differential introgression. Jackson Hole-*Lycaeides* have low linkage disquilibrium and narrow ranges of hybrid indices and do not experience an gene flow from *L. melissa* and *L. idas* [30]. In

addition, genomic clines analyses revealed that a subset of loci have fixed for chromosomal blocks inherited from *L. melissa* and *L. idas* suggesting genomic stabilization in Jackson Hole-*Lycaeides* which could have also been driven by DM incompatablities [30], [39].

Steep genomic clines have been recorded in various organisms (For example, rabbits [55], fishes [56]) and mainly reflect some genomic regions in hybrid genomes experience strong selection for resistance to introgression. In addition, mixed admixture patterns which have parts of the genome from pure parental species and another ancient hybrid species are reflective of extensive genetic variation similar to some human populations specifically in India where there is high level of population substructure [57, 58].

Predictable genomic changes across ancient and contemporary hybrids

Randomization tests for quantifying overlap of SNPs with high ancestry estimates in Jackson Hole-*Lycaeides* and high/low genomic cline parameters in Dubois-*Lycaeides*, reveal really high degree of repeatability between ancient and contemporary hybrids. Since Dubois-*Lycaeides* have a smaller population size and Jackson Hole-*Lycaeides* are older admixed populations, stochastic processes like genetic drift could drive parallelism here [20]. These patterns could also be driven by differences in recombination rates and linkage disequilibrium in both sets of populations. In addition, extrinsic factors can differ across hybrid zones and genetic architecture of reproductive isolation itself can vary in different hybrid populations with changes in populations structure, gene flow and divergence [8]. However, the high degree of predictability supported by extensive enrichments of SNPs (x-fold for top 0.1% for beta is 41) overcomes these external chaos and lends support to our hypothesis that patterns of hybridization and introgression in *Lycaeides* butterflies are driven by natural selection and the loci underlying concordant patterns between the two hybrid zones are putative targets of selection. Here, we first discuss how this comparison of hybrid zones is novel and important and then interpret our claims in the light of predictability in evolution and structural and biological importance of loci underlying repeatability.

While studies on hybrid zones have focused on concordance in patterns of introgression across different transects or independent hybrid zones, these studies have mainly revealed concordance (or lack thereof) across space. Our study compares ancient and contemporary hybrid zones, which

offer an unique opportunity to understand the process of hybridization on a temporal scale. These studies are missing not due to lack of evidence but mainly due to the difficulty in studying ancient and contemporary hybridization within a single system. While ancient hybridization exists in several species, including humans, it is not always possible to compare contemporary hybridization with ancient hybrids mainly since these opportunities may not exist (e.g., for humans, most of the hybridizing species are now extinct). *Lycaeides* butterflies offer a novel and unique opportunity to make temporal comparisons in hybrid zones and these results can hopefully be extrapolated to other organisms. Another reason why this study is novel is in the use of genomic data to compare hybrid zones. Previous studies comparing hybrid zones on a spatial zone have mainly focused on limited sets of SNPs or targeted loci [26, 27, 59]. The use of small number of loci in testing for contingency in hybrid zones can be problematic since these are ancestry informative markers and these loci may show variation across the hybrid zone. In addition, locus specific introgression relative to the rest of the genome can be variable and cannot be interpreted correctly based on few loci [6]. Therefore, our study offers a novel approach to study contingency across hybrid zones.

While it is difficult to disentangle the role of selection versus other stochastic processes in driving predictable evolutionary processes, our results show high degree of repeatability and suggest that natural selection can drive genomic evolution on a temporal scale. Studies of parallel evolution provide evidence of parallel divergence in ecotypes inhabiting patchy habitats [60, 61] and for parallel adaptation to similar environments [18] which are focusing on a geographic scale. However, on a temporal scale, predictablity in evolution is more idiosyncratic. Nosil and colleagues used stick insect populations to study the evolution of cryptic body coloration and pattern using 25 years of field data, experiments, and genomics [62]. There results suggested that even though evolutionary outcomes are predictable on a short-term, long-term outcomes are difficult to predict. Similarly, in a 30 year long study in Darwin's finches pointed that on a long-term evolution is unpredictable since environmental fluctuations can affect selection coefficients [63]. While these patterns hold for populations with clear limits to gene flow, processes in hybrid zones are complicated since populations experience repeated instances of allopatry and secondary contact with constant mixing of the genomes. In addition, populations can show extensive substructure and genetic variation due to gene flow from

several populations of pure species and admixed species. Therefore, seeing consistency in loci involved in the process of reproductive isolation and speciation on a longer time scale in our data is quite compelling. This level of repeatability can be enhanced by gene interactions and can also occur due to the presence of standing genetic variation in hybridizing populations where ths variation can be provided by parents to hybrids which eventually aids in adaptive evolution [4, 25]. In *Lycaeides*, the repeatable patterns of ancestry and introgression in hybrid individuals and admixed populations suggests that hybridization is possibly fueling adaptation and diversification from standing genetic variation [39] and therefore few loci are maintained by selection in the genomes of hybrids across time.

Z versus autosomes

Sex chromosomes have been known to be involved in the evolution of reproductive isolation and can harbor loci which restrict introgression and play a role in speciation [9, 64, 65, 66]. These loci can underlie genomic incompatibilities, inviability and sterility in hybrids which has been associated with faster rate of adaptive evolution, reduced recombination and overrepresentation of sex related genes [67]. While previous studies have highlighted the role of sex chromosome in hybrid speciation in butterflies and birds and showed differential introgression on X in mice, studies have not explicitly quantified concordant patterns on sex chromosome across a temporal scale. Our results on variable patterns of ancestry frequencies on Z versus autosomes have been seen in humans [68]. In addition, contrasting patterns of introgression on Z versus autosomes have been demonstrated in house mouse [69], crickets [12], swordtail fishes [70]. The results of high level of overlap in genomic regions between ancient and contemporary hybrid zones is interesting suggesting that these regions consistently harbor loci which restrict introgression and reflect a stronger signal of selection relative to autosomes. Z chromosomes have reduced effective population sizes than that of autosomes due to female heterogamety, which can have significant effect on the sorting of ancestral variation and parental alleles with increased rates of fixation via drift or selection [6]. This can affect both Jackson Hole-*Lycaeides* as they are in initial stages of genome stabilization and Dubois-*Lycaedies* where the population size is not too large. However, our signals of predictability are not restricted to Z and we see considerable enrichments on across autosomes. These results, coupled with previous

evidence, further support the hypothesis that Z chromosomes are important in the adaptive evolution and hybrid genome evolution of *Lycaeides* butterflies.

Candidate barrier loci

Genome annotation analyses provide information on the structural properties and the biological functions of genes that are non-randomly associated with regions enriched with excess SNPs with high ancestry in Jackson Hole-*Lycaeides* and high/low genomic cline parameter values in Dubois-*Lycaedies*. In the top 0.1% quantile range, we find several SNPs associated with interesting gene ontologies and interproscan terms for all three predictability comparisons (high ancestry SNPs in Jackson Hole-*Lycaeides* with high β, high and low β in Dubois-*Lycaeides*). Mainly, these annotations suggest that reproductive isolation and introgression in *Lycaeides* hybrids is driven by interactions of several genes. We discuss some of these genes below.

For the 49 shared SNPs in the top 0.1% category, with high ancestry values in Jackson Hole-*Lycaeides* and high β in Dubois-*Lycaeides*, we see several interesting gene annotations. These SNPs are putative candidates for reproductive isolation and restrict introgression and therefore interesting functional annotations can suggest how specific traits are crucial for differentiation of *Lycaeides*. One SNP was annotated 5 gene ontology terms and 7 interproscan terms associated with phosphogluconate dehyrdogenase activity (Tables 2.14, 2.15, Figure 2.8). Studies focusing on role of phosphogluconate dehydrogenase in insect physiology have highlighted its possible role in insect cold hardiness, cold adaptation and diapause bhavior. Historically, fixed allele differences in phosphogluconate dehydrogenase or pgd have been described in Swallowtail butterflies, *Papilio glaucus* and *Papilio canadensis*, wherein sex-linked pgd alleles suppress melanin in glaucus X canadensis hybrids and could be possibly linked to locus (od) which is associated with obligate diapause and has recently been supported with genomic analyses [64, 71, 72, 72]. In *Bombyx mori*, 6-pgd was expressed at higher level in non-diapausing eggs [73]. In addition, it has been shown to be an important maker to track clinal variation along latitudes in *Drosophila* wherein the allele frequency of 6-pgd and pgd tends to increase with latitude [74]. Lastly, 6-pgd and NADPH have been shown to interact to aid in the process of freeze tolerance in *Ostrinia nubilalis* larvae exposed to cold temperatures [75]. Therefore, pgd could possibly play a role in distinction of obligate

versus facultative diapause in *Lycaeides*. *Lycaeides* populations inhabiting higher elevations (mostly *L.idas*) exhibit obligate diapause. However, lower elevation populations can exhibit facultative diapause (mostly *L. melissa*). Jackson Hole-*Lycaeides* inhabit similar habitats as *L. idas* and mostly exhibit obligate diapause. Dubois-*Lycaeides* inhabit similar habitat to *L. melissa* and may possibly exhibit facultative diapause. However, diapause is definitely a trait which differentiates *L. idas* and *L. melissa* [39] and these annotations can be suggestive of an important reproductive isolation trait between these species. However, since we do not have additional diapause data to support this claim, we interpret these results as possible patterns which could be dissected with additional analyses.

4 SNPs are annotated for immunoglobulin super-family (Tables 2.14, 2.15,Figure 2.8). Immunoglobulin superfamily is conserved in insects and is crucial in pathogen defense [76]. Immunology genes have also been identified as being involved in reproductive isolation in mice [77]. We also identified one SNP annotated for olfactory reception and odorant binding proteins (Tables 2.14, 2.15, Figure 2.8). Both olfactory receptors (Or) and odorant binding proteins (Odp) are chemosensory proteins (CSP) and are known to play a role in insect-plant interactions and there are more CSPs present in Lepidopteran genomes as compared to any other insect genomes [78]. Odp have been identified to influence host plant use in *Heliconius* butterflies [79] and specifically affect host plant choice for oviposition in female Swallowtail butterflies by facilitating recognition [78]. This makes sense in the context of *Lycaeides*, as host plant choice is crucial in these butterflies with populations being locally adapted to their host [80]. *Lycaeides* tend to perform better on their native host and prefer their native host for oviposition over a novel host. Another SNP was annotated for signal transduction, which is also crucial in oviposition behavior in Swallowtail butterflies [81]. One SNP is annotated for Nucleopore structural component with specific interproscan annotation for nucleopore protein Nup93(Tables 2.14, 2.15). Nucleopore proteins have been identified to play a role in hybrid sterility through DMI in Drosophila (Nup 96, Nup160) [82]. While, we do not find the same nucleoprotein in our annotation, Nup96 and Nup93 are both part of the nucleopore complex and have a significant interaction. Therefore, we interpret that our annotation is reflective of nucleopore process. Three SNPs were annotated for genes important in wing development (Wnt signalling, armadillo, Zinc finger) (Tables 2.14, 2.15). Wnt signalling and regulation of Wnt gene is crucial

in wing pattern evolution in Heliconius butterflies [83]. For the 22 shared SNPs in the top 0.1% category, with high ancestry values in Jackson Hole-*Lycaeides* and high α in Dubois-*Lycaeides*, we find two SNPs associated with immunoglobulin superfamily and many associated with protein binding.

Overall our results suggest that reproductive isolation and introgression in *Lycaeides* involves functionally diverse set of genes. However, one caveat with our results is that in most cases SNPs were present on the same scaffold and associated with the same gene with limited number of SNPs (one-four) present on or near genes with molecular functions that are variable at the genomic level. Therefore, we have limited support for the hypothesis that genes with these functions are important and signficantly associated with putative targets of selection.

Conclusion

Our study provides a novel approach to study hybrid zones on a temporal scale to quantify repeatability associated with the process of reproductive isolation and speciation. We demonstrate substantial repeatability in regions underlying restricted introgression between ancient and contemporary hybrids and provide evidence that natural selection can indeed drive contingent evolutionary patterns across a long period of time. Our study also highlights that it is important to identify distribution of outlier loci across the genome as some regions can be more involved in specific evolutionary processes than others. Lastly, specific biological functions can be crucial targets of selection and drive differences between species. Therefore, attaching functional information to loci involved in introgression or differentiation can inform ambiguities in identifying the role of selection in driving patterns of evolution. Overall, we argue that *Lycaeides* butterflies serve as a striking case where hybridization is crucial force driving species adaptation and diversification and natural selection as a process can drive predictable patterns of evolution.

REFERENCES

[1] G. M. Hewitt, Hybrid zones-natural laboratories for evolutionary studies. *Trends in Ecology Evolution* **3**, 158–167 (1988).

[2] R. G. Harrison, E. L. Larson, Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity* **105**, 795–809 (2014).

[3] B. A. Payseur, L. H. Rieseberg, A genomic perspective on hybridization and speciation. *Molecular Ecology* **25**, 2337–2360 (2016).

[4] R. J. Abbott, N. H. Barton, J. M. Good, Genomics of hybridization and its evolutionary consequences. *Molecular Ecology* **25**, 2325–2332 (2016).

[5] R. Abbott, D. Albach, S. Ansell, J. W. Arntzen, S. J. Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, *et al.*, Hybridization and speciation. *Journal of evolutionary biology* **26**, 229–246 (2013).

[6] Z. Gompert, E. G. Mandeville, C. A. Buerkle, Analysis of population genomic data from hybrid zones. *Annual Review of Ecology, Evolution, and Systematics* **48** (2017).

[7] S. H. Martin, C. D. Jiggins, Interpreting the genomic landscape of introgression. *Current opinion in genetics & development* **47**, 69–74 (2017).

[8] R. G. Harrison, E. L. Larson, Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Molecular Ecology* **25**, 2454–2466 (2016).

[9] G.-P. Sætre, T. Borge, K. Lindroos, J. Haavie, B. C. Sheldon, C. Primmer, A.-C. Syvänen, Sex chromosome evolution and speciation in ficedula flycatchers. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 53–59 (2003).

[10] B. A. Payseur, J. G. Krenz, M. W. Nachman, Differential patterns of introgression across the x chromosome in a hybrid zone between two species of house mice. *Evolution* **58**, 2064–2078 (2004).

[11] S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, D. Reich, The genomic landscape of neanderthal ancestry in present-day humans. *Nature* **507**, 354 (2014).

[12] L. S. Maroja, E. L. Larson, S. M. Bogdanowicz, R. G. Harrison, Genes with restricted introgression in a field cricket (gryllus firmus/gryllus pennsylvanicus) hybrid zone are concentrated on the x chromosome and a single autosome. *G3: Genes, Genomes, Genetics* **5**, 2219–2227 (2015).

[13] X.-S. Hu, D. A. Filatov, The large-x effect in plants: increased species divergence and reduced gene flow on the silene x-chromosome. *Molecular ecology* **25**, 2609–2619 (2016).

[14] T. O. Elgvin, C. N. Trier, O. K. TÃÿrresen, I. J. Hagen, S. Lien, A. J. Nederbragt, M. Ravinet, H. Jensen, G.-P. SÃętre, The genomic mosaicism of hybrid speciation. *Science Advances* **3**, e1602996 (2017).

[15] M. Roesti, A. P. Hendry, W. Salzburger, D. Berner, Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular ecology* **21**, 2852–2862 (2012).

[16] R. Burri, A. Nater, T. Kawakami, C. F. Mugal, P. I. Olason, L. Smeds, A. Suh, L. Dutoit, S. Bureš, L. Z. Garamszegi, *et al.*, Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of ficedula flycatchers. *Genome research* **25**, 1656–1665 (2015).

[17] Z. Gompert, A continuous correlated beta process model for genetic ancestry in admixed populations. *PloS one* **11**, e0151047 (2016).

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

[19] A. E. Lobkovsky, E. V. Koonin, Replaying the tape of life: quantification of the predictability of evolution. *Frontiers in Genetics* **3**, 246 (2012).

[20] T. J. Ord, T. C. Summers, Repeated evolution and the impact of evolutionary history on adaptation. *BMC Evolutionary Biology* **15**, 137 (2015).

[21] V. Orgogozo, Replaying the tape of life in the twenty-first century. *Interface focus* **5**, 20150057 (2015).

[22] D. I. Bolnick, R. D. Barrett, K. B. Oke, D. J. Rennison, Y. E. Stuart, (non) parallel evolution. *Annual Review of Ecology, Evolution, and Systematics* **49**, 303–330 (2018).

[23] P. F. Colosimo, K. E. Hosemann, S. Balabhadra, G. Villarreal, M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, D. Schluter, D. M. Kingsley, Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**, 1928–1933 (2005).

[24] V. Soria-Carrasco, Z. Gompert, A. A. Comeault, T. E. Farkas, T. L. Parchman, J. S. Johnston, C. A. Buerkle, J. L. Feder, J. Bast, T. Schwander, *et al.*, Stick insect genomes reveal natural selectionâĂŹs role in parallel speciation. *Science* **344**, 738–742 (2014).

[25] K. Johannesson, M. Panova, P. Kemppainen, C. André, E. Rolan-Alvarez, R. K. Butlin, Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 1735–1747 (2010).

[26] L. H. Rieseberg, B. Sinervo, C. R. Linder, M. C. Ungerer, D. M. Arias, Role of gene interactions in hybrid speciation: Evidence from ancient and experimental hybrids. *Science* **272**, 741–745 (1996).

[27] C. A. Buerkle, L. H. Rieseberg, Low intraspecific variation for genomic isolation between hybridizing sunflower species. *Evolution* **55**, 684–691 (2001).

[28] A. Nolte, Z. Gompert, C. Buerkle, Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. *Molecular Ecology* **18**, 2615–2627 (2009).

[29] Z. GOMPERT, M. L. FORISTER, J. A. FORDYCE, C. C. NICE, Widespread mitoâĂŘnuclear discordance with evidence for introgressive hybridization and selective sweeps in lycaeides. *Molecular Ecology* **17**, 5231–5244 (2008).

[30] Z. Gompert, L. K. Lucas, C. C. Nice, J. A. Fordyce, M. L. Forister, C. A. Buerkle, Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution* **66**, 2167–2181 (2012).

[31] R. Vila, C. D. Bell, R. Macniven, B. Goldman-Huertas, R. H. Ree, C. R. Marshall, Z. Bálint, K. Johnson, D. Benyamini, N. E. Pierce, Phylogeny and palaeoecology of polyommatus blue butterflies show beringia was a climate-regulated gateway to the new world. *Proceedings of the Royal Society of London B: Biological Sciences* **278**, 2737–2744 (2011).

[32] L. Lucas, J. Fordyce, C. Nice, Patterns of genitalic morphology around suture zones in north american lycaeides (lepidoptera: Lycaenidae): implications for taxonomy and historical biogeography. *Annals of the Entomological Society of America* **101**, 172–180 (2014).

[33] L. K. Lucas, C. C. Nice, Z. Gompert, 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 (2018).

[34] Z. Gompert, C. C. Nice, J. A. Fordyce, M. L. Forister, A. M. Shapiro, Identifying units for conservation using molecular systematics: the cautionary tale of the karner blue butterfly. *Molecular ecology* **15**, 1759–1768 (2006).

[35] Z. Gompert, J. A. Fordyce, M. L. Forister, A. M. Shapiro, C. C. Nice, Homoploid hybrid speciation in an extreme habitat. *Science* **314**, 1923–1925 (2006).

[36] Z. Gompert, M. L. Forister, J. A. Fordyce, C. C. Nice, R. J. Williamson, C. Alex Buerkle, Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of lycaeides butterflies. *Molecular ecology* **19**, 2455–2473 (2010).

[37] Z. Gompert, L. K. Lucas, J. A. Fordyce, M. L. Forister, C. C. Nice, Secondary contact between lycaeides idas and l.Âămelissa in the rocky mountains: extensive admixture and a patchy hybrid zone. *Molecular Ecology* **19**, 3171–3192 (2010).

[38] V. V. Nabokov, *The nearctic members of the genus Lycaeides Hübner (Lycaenidae, Lepidoptera)* (Museum of Comparative ZooÍŔÌĹlogy, 1949).

[39] Z. Gompert, L. K. Lucas, C. A. Buerkle, M. L. Forister, J. A. Fordyce, C. C. Nice, Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology* **23**, 4555–4573 (2014).

[40] Z. Gompert, L. K. Lucas, C. C. Nice, C. A. Buerkle, Genome divergence and the genetic architecture of barriers to gene flow between lycaeides idas and l. melissa. *Evolution* **67**, 2498–2514 (2013).

[41] Z. Gompert, A. A. Comeault, T. E. Farkas, J. L. Feder, T. L. Parchman, C. A. Buerkle, P. Nosil, Experimental evidence for ecological selection on genome variation in the wild. *Ecology letters* **17**, 369–379 (2014).

[42] M. Yandell, D. Ence, A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* **13**, 329 (2012).

[43] M. S. Campbell, C. Holt, B. Moore, M. Yandell, Genome annotation and curation using maker and maker-p. *Current Protocols in Bioinformatics* **48**, 4–11 (2014).

[44] A. L. Price, N. C. Jones, P. A. Pevzner, De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).

[45] R. J. Challis, S. Kumar, K. K. K. Dasmahapatra, C. D. Jiggins, M. Blaxter, Lepbase: the lepidopteran genome database. *BioRxiv* p. 056994 (2016).

[46] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, *et al.*, Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology* **29**, 644 (2011).

[47] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, *et al.*, De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494 (2013).

[48] I. Korf, Gene finding in novel genomes. *BMC bioinformatics* **5**, 59 (2004).

[49] M. Stanke, B. Morgenstern, Augustus: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* **33**, W465–W467 (2005).

[50] H. Li, R. Durbin, Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).

[51] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

[52] J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

[53] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2017).

[54] Z. Gompert, C. Buerkle, bgc: Software for bayesian estimation of genomic clines. *Molecular Ecology Resources* **12**, 1168–1176 (2012).

[55] M. Carneiro, S. J. Baird, S. Afonso, E. Ramirez, P. Tarroso, H. Teotónio, R. Villafuerte, M. W. Nachman, N. Ferrand, Steep clines within a highly permeable genome across a hybrid zone between two subspecies of the european rabbit. *Molecular ecology* **22**, 2511–2525 (2013).

[56] A. Souissi, F. Bonhomme, M. Manchado, L. Bahri-Sfar, P.-A. Gagnaire, Genomic and geographic footprints of differential introgression between two divergent fish species (solea spp.). *Heredity* **121**, 579–593 (2018).

[57] D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing indian population history. *Nature* **461**, 489 (2009).

[58] P. Moorjani, K. Thangaraj, N. Patterson, M. Lipson, P.-R. Loh, P. Govindaraj, B. Berger, D. Reich, L. Singh, Genetic evidence for recent population mixture in india. *The American Journal of Human Genetics* **93**, 422–438 (2013).

[59] E. L. Larson, T. A. White, C. L. Ross, R. G. Harrison, Gene flow and the maintenance of species boundaries. *Molecular Ecology* **23**, 1668–1678 (2014).

[60] A. P. Michel, S. Sim, T. H. Powell, M. S. Taylor, P. Nosil, J. L. Feder, Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences* **107**, 9724–9729 (2010).

[61] P. Nosil, Z. Gompert, T. E. Farkas, A. A. Comeault, J. L. Feder, C. A. Buerkle, T. L. Parchman, Genomic consequences of multiple speciation processes in a stick insect. *Proc. R. Soc. B* p. rspb20120813 (2012).

[62] P. Nosil, R. Villoutreix, C. F. de Carvalho, T. E. Farkas, V. Soria-Carrasco, J. L. Feder, B. J. Crespi, Z. Gompert, Natural selection and the predictability of evolution in timema stick insects. *Science* **359**, 765–770 (2018).

[63] P. R. Grant, B. R. Grant, Unpredictable evolution in a 30-year study of darwin's finches. *science* **296**, 707–711 (2002).

[64] R. Hagen, J. Scriber, Sex-linked diapause, color, and allozyme loci in papilio glaucus: linkage analysis and significance in a hybrid zone. *Journal of Heredity* **80**, 179–185 (1989).

[65] D. C. Presgraves, Sex chromosomes and speciation in drosophila. *Trends in Genetics* **24**, 336–343 (2008).

[66] K. Kunte, C. Shea, M. L. Aardema, J. M. Scriber, T. E. Juenger, L. E. Gilbert, M. R. Kronforst, Sex chromosome mosaicism and hybrid speciation among tiger swallowtail butterflies. *PLoS Genetics* **7**, e1002274 (2011).

[67] N. A. Johnson, J. Lachance, The genetics of sex chromosomes: evolution and implications for hybrid incompatibility. *Annals of the New York Academy of Sciences* **1256**, E1–E22 (2012).

[68] S. Sankararaman, S. Mallick, M. Dannemann, K. PrÃijfer, J. Kelso, S. PÃďÃďbo, N. Patterson, D. Reich, The genomic landscape of neanderthal ancestry in present-day humans. *Nature* **507**, 354 (2014).

[69] L. M. Turner, B. Harr, Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and dobzhansky-muller interactions. *Elife* **3**, e02504 (2014).

[70] M. Schumer, R. Cui, D. L. Powell, G. G. Rosenthal, P. Andolfatto, Ancient hybridization and genomic stabilization in a swordtail fish. *Molecular Ecology* **25**, 2661–2679 (2016).

[71] J. M. Scriber, B. L. Giebink, D. Snider, Reciprocal latitudinal clines in oviposition behavior ofpapilio glaucus andp. canadensis across the great lakes hybrid zone: possible sex-linkage of oviposition preferences. *Oecologia* **87**, 360–368 (1991).

[72] Q. Cong, D. Borek, Z. Otwinowski, N. V. Grishin, Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell reports* **10**, 910–919 (2015).

[73] L. Fan, J. Lin, Y. Zhong, J. Liu, Shotgun proteomic analysis on the diapause and non-diapause eggs of domesticated silkworm bombyx mori. *PLoS One* **8**, e60386 (2013).

[74] J. R. Adrion, M. W. Hahn, B. S. Cooper, Revisiting classic clines in drosophila melanogaster in the age of genomics. *Trends in Genetics* **31**, 434–444 (2015).

[75] B. Stanic, A. Jovanovic-Galovic, D. P. Blagojevic, G. Grubor-Lajsic, R. Worland, M. B. Spasic, Cold hardiness in ostrinia nubilalis (lepidoptera: Pyralidae): Glycerol content, hexose monophosphate shunt activity, and antioxidative defense system. *European Journal of Entomology* **101**, 459–466 (2004).

[76] M. Mandrioli, M. Monti, R. Tedeschi, Presence and conservation of the immunoglobulin superfamily in insects: current perspective and future challenges. *Invertebrate Survival Journal* **12**, 188–194 (2015).

[77] K. C. Teeter, B. A. Payseur, L. W. Harris, M. A. Bakewell, L. M. Thibodeau, J. E. OâĂŹBrien, J. G. Krenz, M. A. Sans-Fuentes, M. W. Nachman, P. K. Tucker, Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome research* **18**, 67–76 (2008).

[78] K. Ozaki, A. Utoguchi, A. Yamada, H. Yoshikawa, Identification and genomic structure of chemosensory proteins (csp) and odorant binding proteins (obp) genes expressed in foreleg tarsi of the swallowtail butterfly papilio xuthus. *Insect biochemistry and molecular biology* **38**, 969–976 (2008).

[79] A. D. Briscoe, A. Macias-Muñoz, K. M. Kozak, J. R. Walters, F. Yuan, G. A. Jamie, S. H. Martin, K. K. Dasmahapatra, L. C. Ferguson, J. Mallet, *et al.*, Female behaviour drives expression and evolution of gustatory receptors in butterflies. *PLoS genetics* **9**, e1003620 (2013).

[80] S. Chaturvedi, L. K. Lucas, C. C. Nice, J. A. Fordyce, M. L. Forister, Z. Gompert, The predictability of genomic changes underlying a recent host shift in melissa blue butterflies. *Molecular ecology* **27**, 2651–2666 (2018).

[81] K. Ozaki, M. Ryuda, A. Yamada, A. Utoguchi, H. Ishimoto, D. Calas, F. Marion-Poll, T. Tanimura, H. Yoshikawa, A gustatory receptor involved in host plant recognition for oviposition of a swallowtail butterfly. *Nature communications* **2**, 542 (2011).

[82] S. Tang, D. C. Presgraves, Lineage-specific evolution of the complex nup160 hybrid incompatibility between drosophila melanogaster and its sister species. *Genetics* **200**, 1245–1254 (2015).

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

Tables and Figures

Fig. 2.1. Diagram shows conceptual overview and a comparative summary of genomic patterns due to hybridization in *Lycaeides* in ancient hybrids and contemporary hybrids. (A) Histogram shows hybrid index distributions for the hybrid categories. (B) Plots of ancestry blocks in the chromosome (dark gray versus light) in different hybrid individuals. For ancient hybrids, ancestry blocks have been broken up by recombination and some have stabilized with several individuals harboring many, small blocks. For contemporary hybrids, the ancestry blocks are still intact and not broken up by recmbination. (C) For ancient hybrids, plot show variation in ancestry frequency for loci across the genome. Red arrows indicate selection acting on specific regions of the genome where loci have high versus low ancestry frequency. For contemporary hybrids, genomic cline plot depicts the cline parameters α (blue) and β (red). Red arrows again indicate selection acting on specific loci across the genome which are preferred in the genomic background of either of the two parental species. (D) Diagram represents history of hybridization in *Lycaeides*.

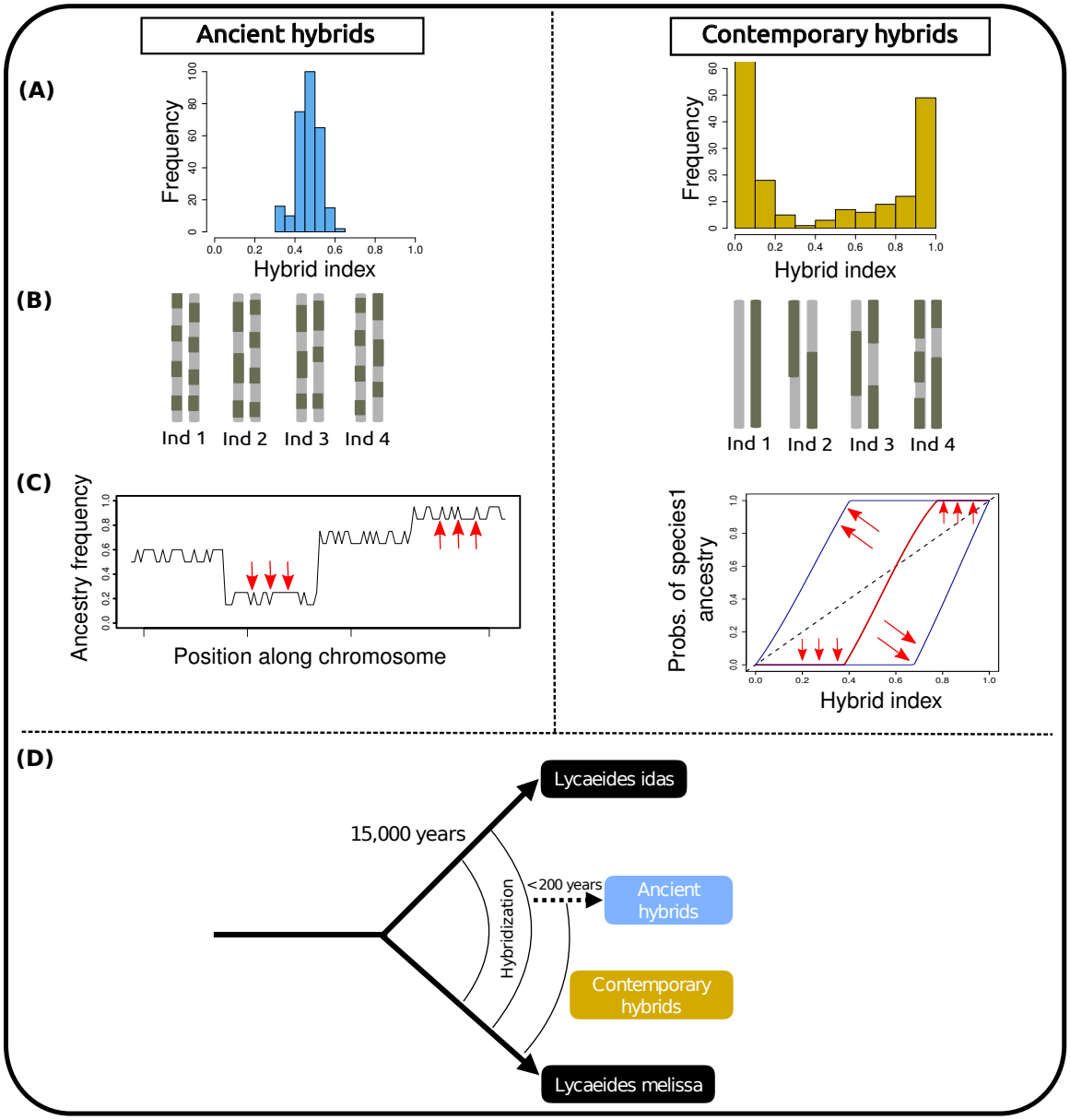


Fig. 2.2. (A) Map shows sample locations with populations colored based on species. Population colors correspond to species for geographical locations in Table 2.1. (B) Plot shows summary of population structure based on principal component analysis. The points denote individuals in each population used for the analysis. (C) Violin plot shows variation in genomic composition of individuals from 10 Jackson Hole-*Lycaeides* localities and those from Dubois-*Lycaeides*, based on PC1 scores. Abbreviations in this plot correspind to geographical locations in Table 2.1.

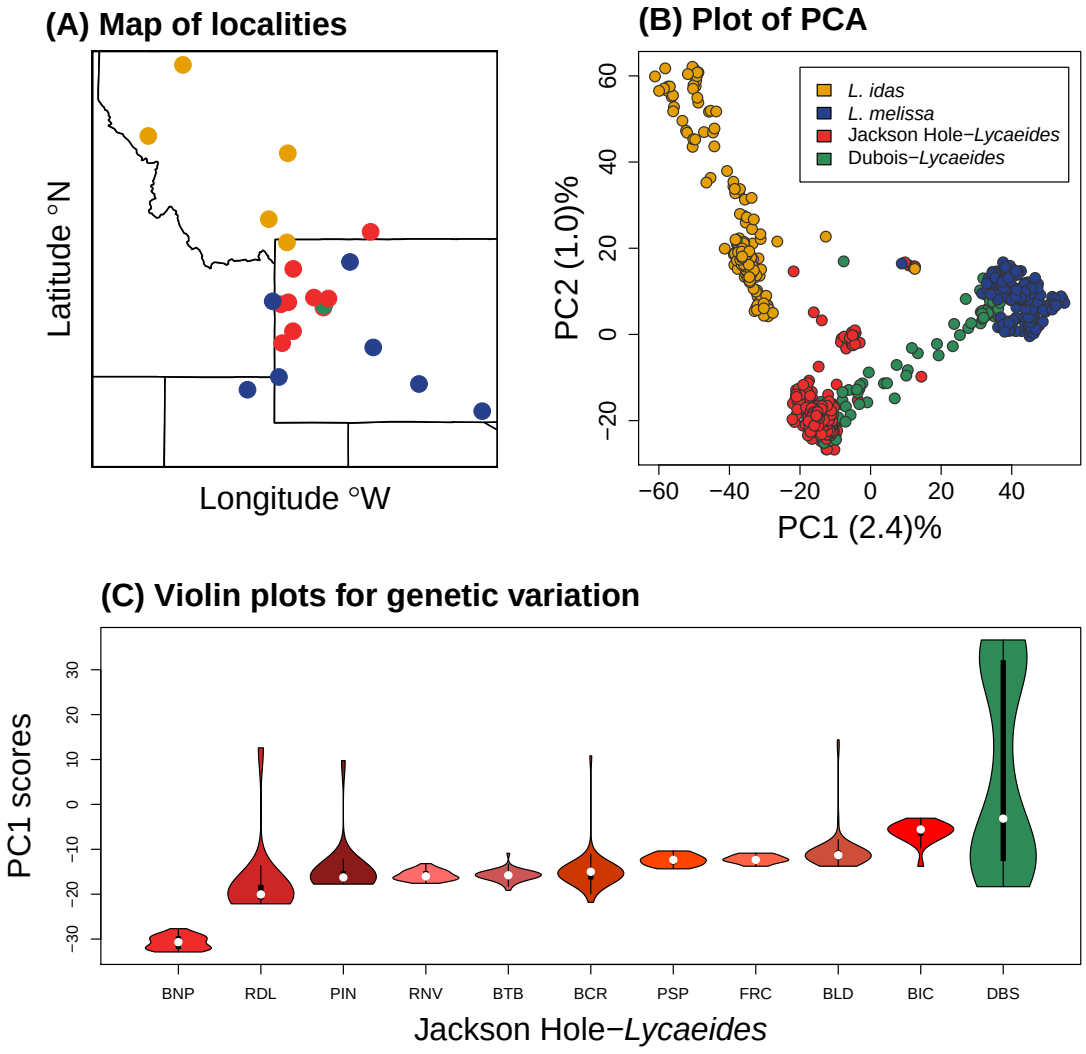


Fig. 2.3. (A) Plot shows frequency distribution of hybrid indices in Dubois-*Lycaeides*. (B) Plot shows estimated genomic clines for representative loci. Each green (locus's 95% CI for α does not include zero) or purple (locus's 95% CI for β includes zero) line represents genomic cline for a single locus. This means that each line gives the probability of Jackson Hole-*Lycaeides* ancestry at an individual locus as a function of hybrid index. The dashed black line gives the probability of ancestry is equal to the hybrid index. (C) Boxplot shows the distribution of cline parameter α values for loci across different linkage groups. (D) Boxplot shows the distribution of cline parameter β values for loci across different linkage groups.

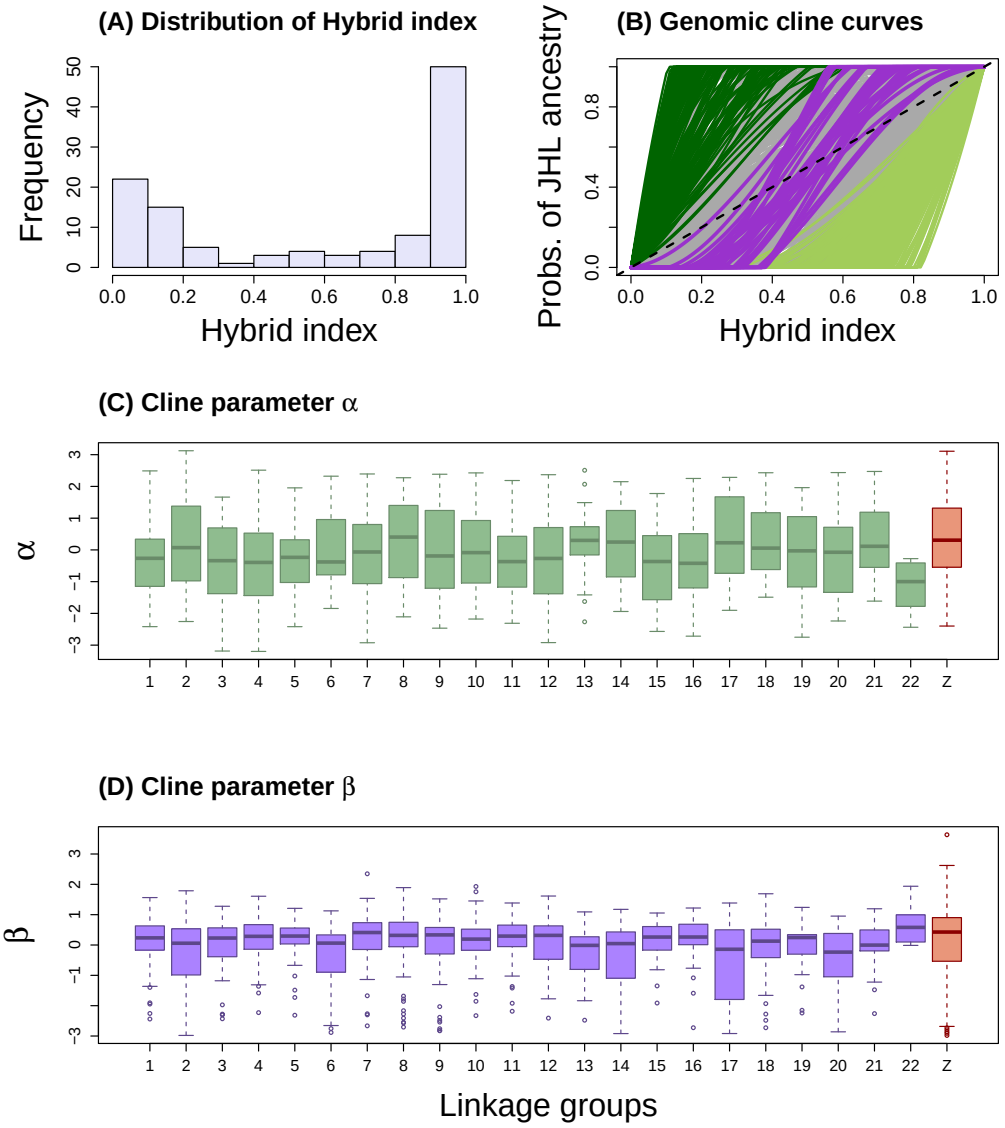


Fig. 2.4. (A) Boxplot shows the distribution *L. idas* ancestry frequencies of loci across different linkage groups based for individuals from Bald Mountain, WY. (B) Boxplot shows the distribution *L. idas* ancestry frequencies of loci across different linkage groups based for individuals from Pinnacle, WY. (Both these localities represent Jackson Hole-*Lycaeides*). (C) Line plots show mean *L. idas* ancestry for each linkage group across 10 populations representing Jackson Hole-*Lycaeides* in the study. Abbreviations in the legend correspond to geographical locations in Table 2.1.

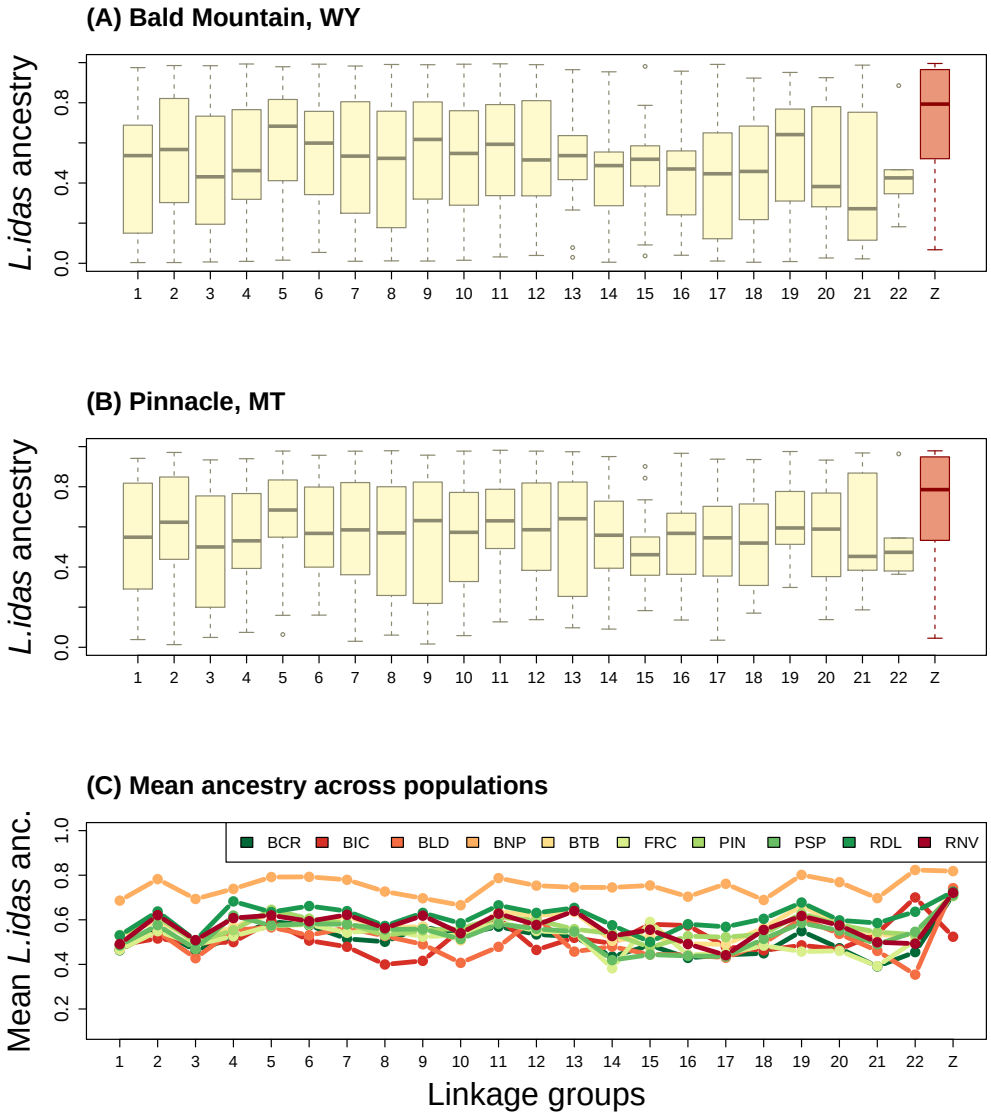
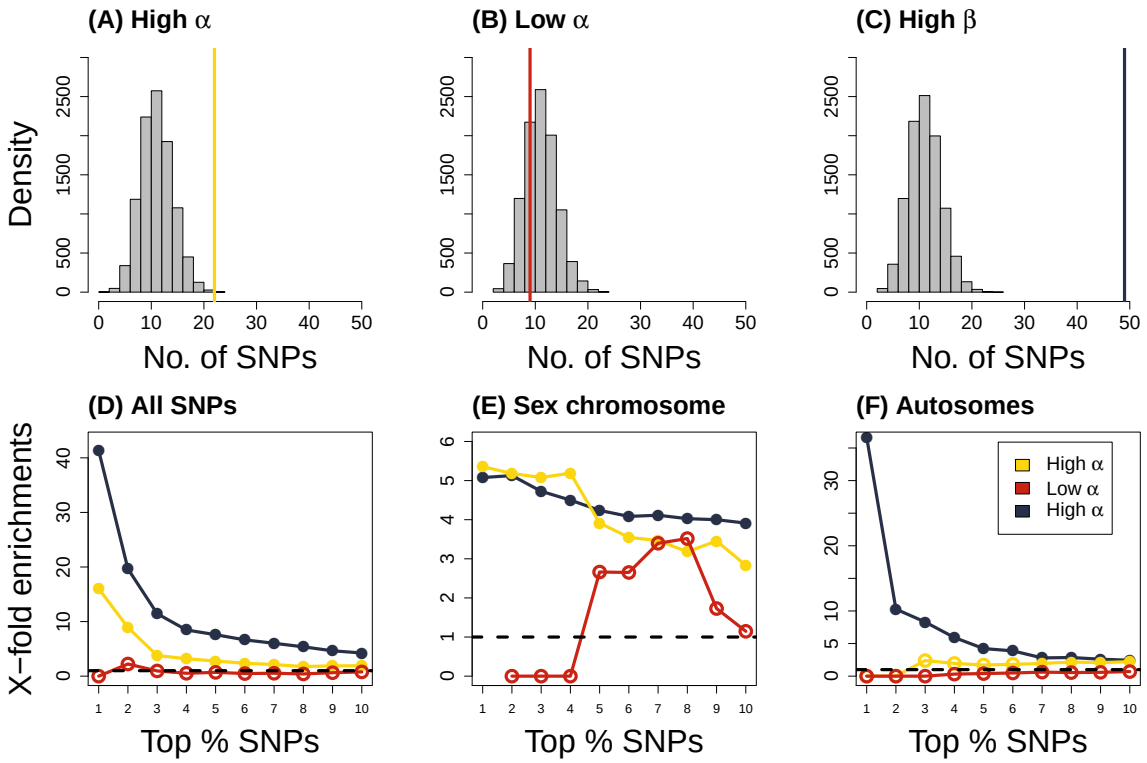


Fig. 2.5. (A) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have high cline parameter α values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Yellow line indicates the number of overlapping SNPs actually observed between the two groups. (B) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have low cline parameter α values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Red line indicates the number of overlapping SNPs actually observed between the two groups. (C) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have high cline parameter β values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Blue line indicates the number of overlapping SNPs actually observed between the two groups. (D) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides*. (E) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides* which lie in excess on Z chromosome. (F) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides* which lie in excess on autosomes. For (D), (E), and (F) open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.



Supplemental tables and figures

Fig. 2.6. S1 Density plot shows the estimated minor allele frequency distribution for all loci (N = 39,139) for all populations included in this study. The population abbreviations are defined in Table 2.1

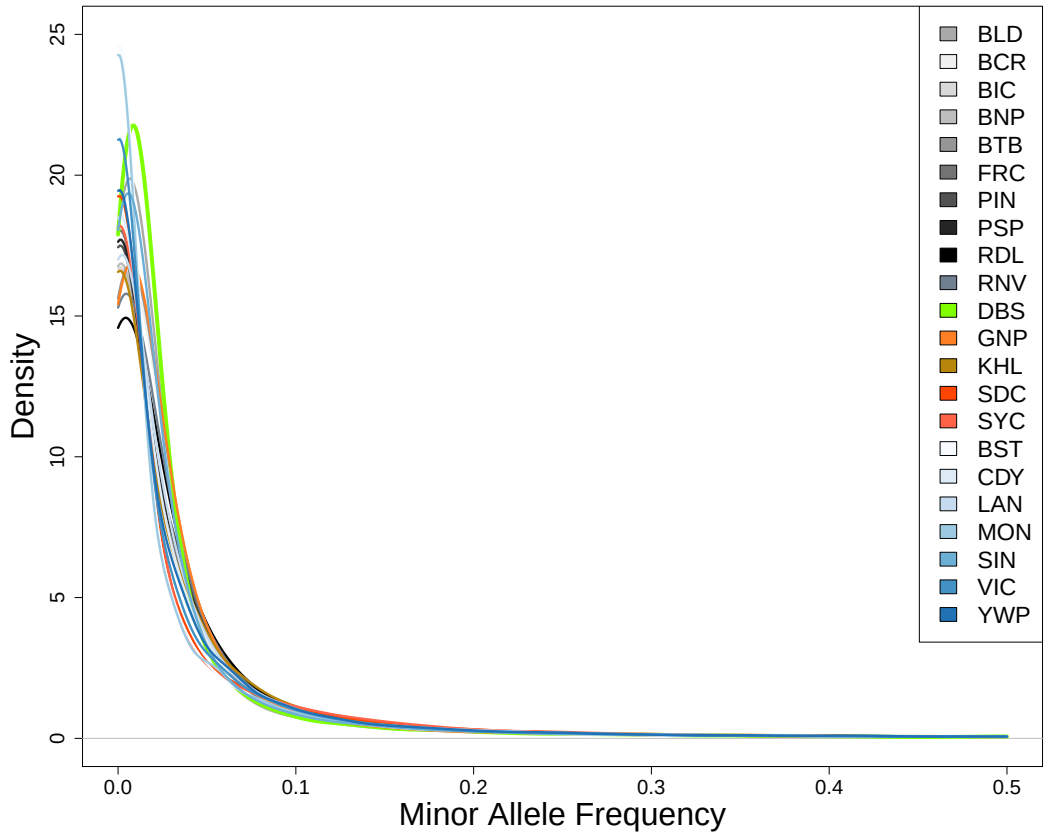


Fig. 2.7. S2 (A) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have high cline parameter α values in Dubois-*Lycaeides*, as expected under a null model for SNPs in the AIMS2 category (N = 2126). This distribution is for overlap in the top 0.1% quantile. Yellow line indicates the number of overlapping SNPs actually observed between the two groups. (B) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have low cline parameter α values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Red line indicates the number of overlapping SNPs actually observed between the two groups. (C) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have high cline parameter β values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Blue line indicates the number of overlapping SNPs actually observed between the two groups. (D) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides*. (E) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides* which lie in excess on Z chromosome. (F) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides* which lie in excess on autosomes. For (D), (E), and (F) open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

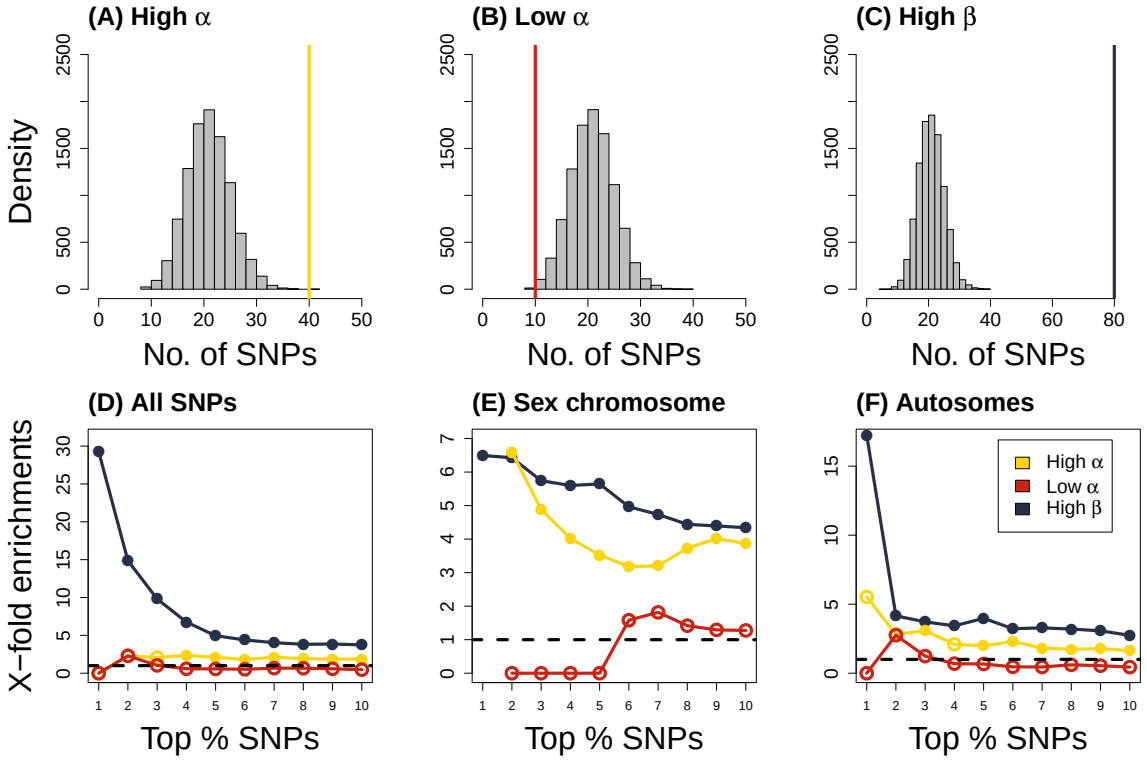


Fig. 2.8. S3 Boxplot shows distribution of hybrid index for each genotype for the six SNPs with excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have high cline parameter β values in Dubois-*Lycaeides*, as expected under a null model. Three of the SNPs (A), (C) and (E) were annotated for unique functional properties. Pg = Phospogluconate dehydrogenase activity, Or = olfactory receptor activity, Odb = Odorant binding protein, and Ig = Immunoglobulin. These are plotted against random SNPs which were not annotated for any functional properties (B), (D) and (F).

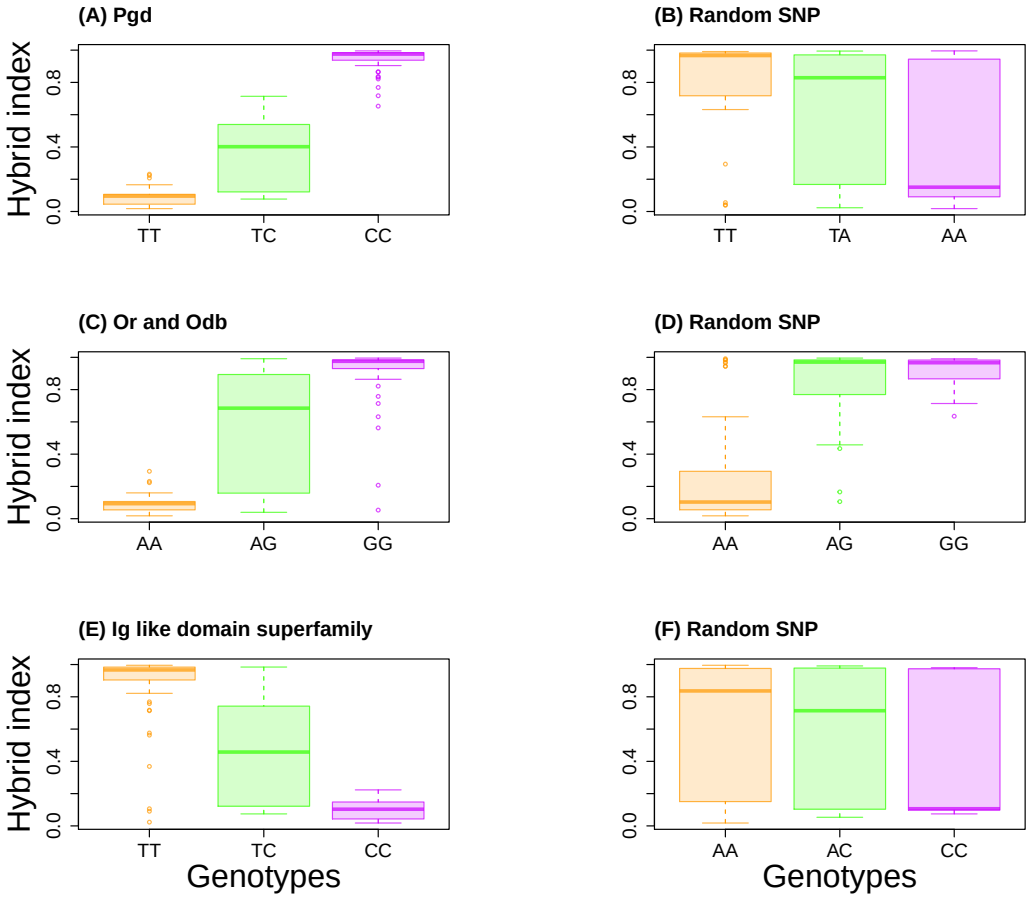


Fig. 2.9. S4 Plots show results for only male samples from Dubois, WY included in this study (N = 89). (A) Plot shows frequency distribution of hybrid index in Dubois-*Lycaeides*. (B) Plot shows estimated genomic clines for representative loci in the AIMS 3 category (N = 1223). Each green (locus's 95% CI for α does not include zero) or purple (locus's 95% CI for β includes zero) line represents genomic cline for a single locus. This means that each line gives the probability of Jackson Hole-*Lycaeides* ancestry at an individual locus as a function of hybrid index. The dashed black line gives the probability of ancestry is equal to the hybrid index. (C) Boxplot shows the distribution of cline parameter α values for loci across different linkage groups. (D) Boxplot shows the distribution of cline parameter β values for loci across different linkage groups.

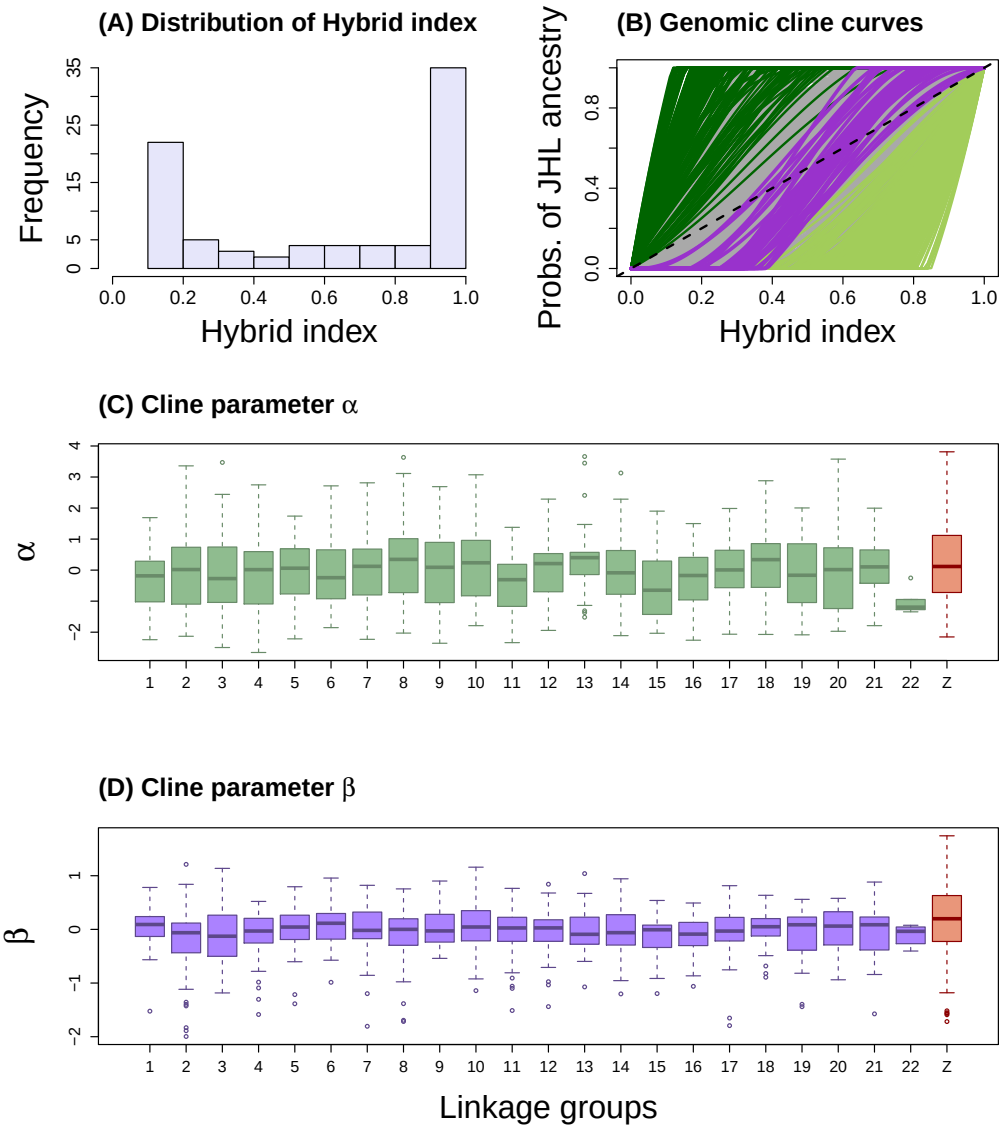


Fig. 2.10. S5 Plots show results for only male samples from Dubois, WY included in this study (N = 89). (A) Boxplot shows the distribution *L. idas* ancestry frequencies of loci across different linkage groups based for individuals from Bald Mountain, WY. (B) Boxplot shows the distribution *L. idas* ancestry frequencies of loci across different linkage groups based for individuals from Pinnacle, WY. (Both these localities represent Jackson Hole-*Lycaeides*. (Both these localities represent Jackson Hole-*Lycaeides*. (C) Line plots show mean *L. idas* ancestry for each linkage group across 10 populations representing Jackson Hole-*Lycaeides* in the study. Abbreviations in the legend correspond to geographical locations in Table 2.1.

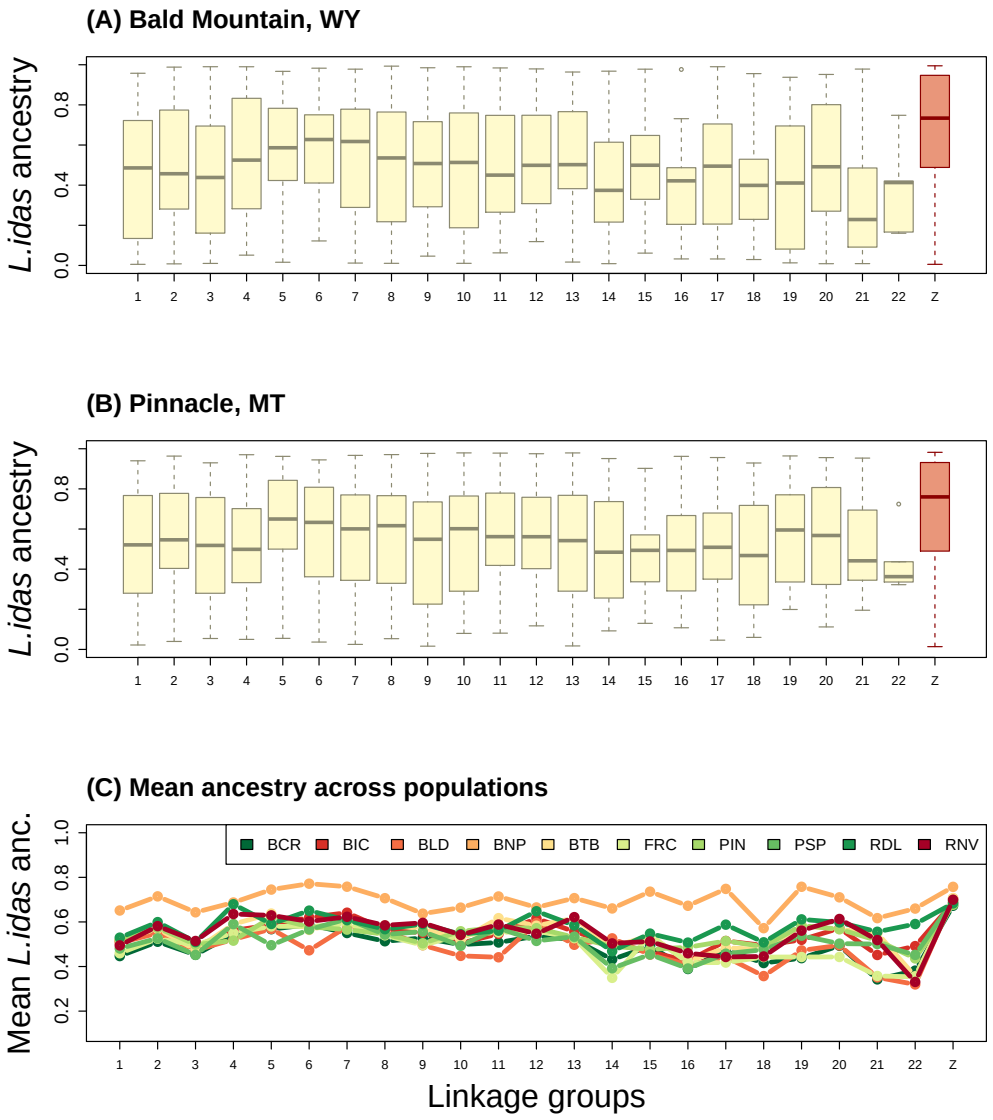


Fig. 2.11. S6 Plots show results for predictability tests for comparisons between male individuals from Jackson Hole-*Lycaeides* (N = 224) and male individuals from Dubois-*Lycaeides* (N =89). These are results for SNPs in AIMS3 category (N = 1223). (A) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have high cline parameter α values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Yellow line indicates the number of overlapping SNPs actually observed between the two groups. (B) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have low cline parameter α values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Red line indicates the number of overlapping SNPs actually observed between the two groups. (C) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have high cline parameter β values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Blue line indicates the number of overlapping SNPs actually observed between the two groups. (D) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides*. (E) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides* which lie in excess on Z chromosome. (F) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides* which lie in excess on autosomes. For (D), (E), and (F) open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

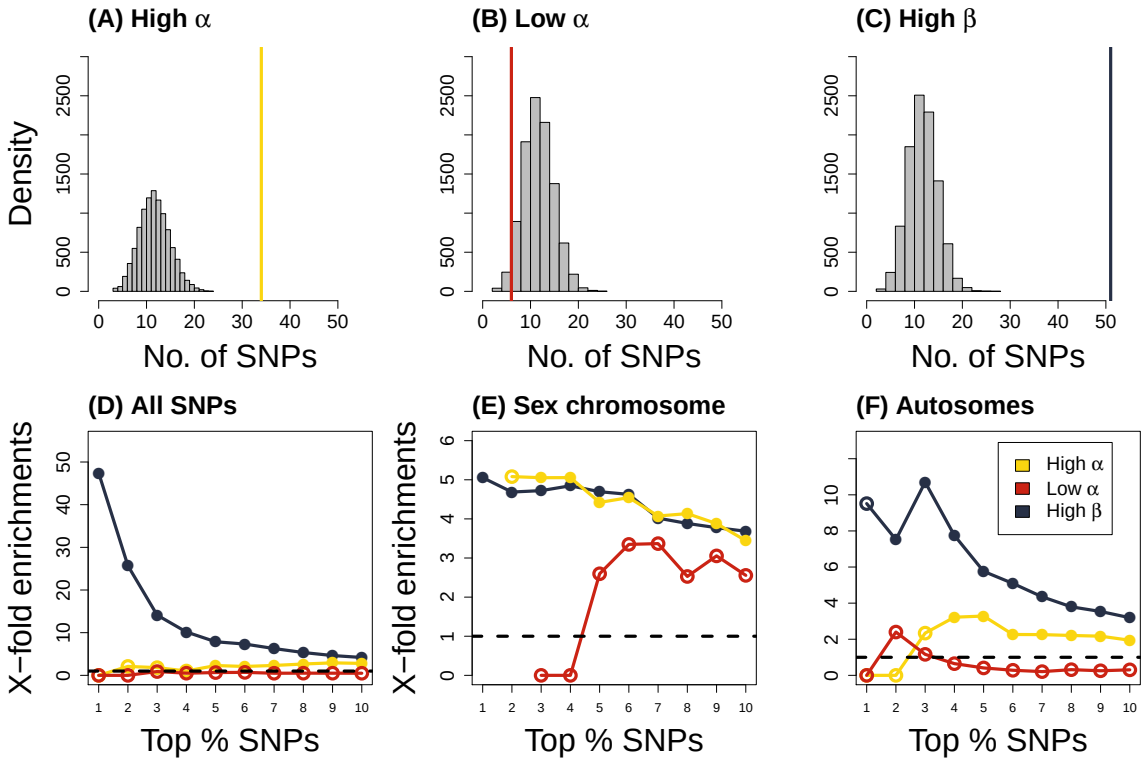


Fig. 2.12. S7 Plots show results for predictability tests for comparisons between male individuals from Jackson Hole-*Lycaeides* (N = 224) and male individuals from Dubois-*Lycaeides* (N =89). These are results for SNPs in AIMS3 category (N = 2133). (A) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have high cline parameter α values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Yellow line indicates the number of overlapping SNPs actually observed between the two groups. (B) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have low cline parameter α values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Red line indicates the number of overlapping SNPs actually observed between the two groups. (C) Histogram shows the null distribution of number of overlapping SNPs enriched for excess *L. idas* ancestry in Jackson Hole-*Lycaeides* and have high cline parameter β values in Dubois-*Lycaeides*, as expected under a null model. This distribution is for overlap in the top 0.1% quantile. Blue line indicates the number of overlapping SNPs actually observed between the two groups. (D) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides*. (E) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides* which lie in excess on Z chromosome. (F) Line plots show x-fold enrichments across quantiles for overlapping SNPs between Jackson Hole-*Lycaeides* and Dubois-*Lycaeides* which lie in excess on autosomes. For (D), (E), and (F) open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

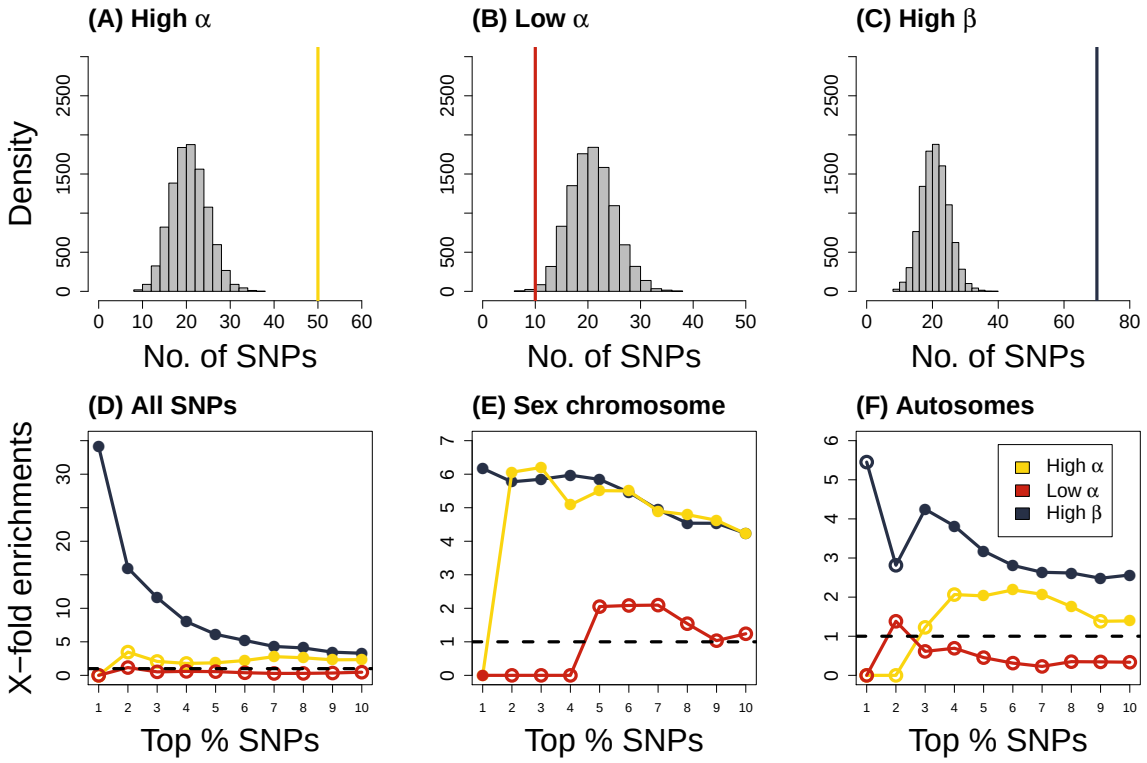


Table 2.1. S1 Locality information and sample sizes for the populations included in this study. Species denotes the species of the individuals sampled from the locality, # Ind. gives the number of individuals sequenced for this study, and Data = indicates whether the sequence data were included in previous study = “Previous" [39] , or are being presented here for the first time = “Present".

	Locality	**Abbreviation**	**Species**	**# Ind.**	**Data**
1	Bishop, CA	BIC	*melissa*	18	2014
2	Bonneville shoreline trail, UT	BST	*melissa*	24	Present
3	Cody, WY	CDY	*melissa*	23	2014
4	Cokeville, WY	CKV	*melissa*	10	2014
5	Lander, WY	LAN	*melissa*	24	2014
6	Montague, CA	MON	*melissa*	20	2014
7	Yellow Pine, WY	YWP	*melissa*	20	2014
8	Sinclair, WY	SIN	*melissa*	97	2014
9	Victor, ID	VIC	*melissa*	20	2014
10	Bald Mountain, WY	BLD	*idas*	74	Present
11	Frontier Creek, WY	FRC	*idas*	20	present
12	Garnet Peak, MT	GNP	*idas*	98	2014
13	King’s Hill, MT	KHL	*idas*	18	2014
14	Soldier Creek, MT	SDC	*idas*	20	2014
15	Siyeh Creek, MT	SYV	*idas*	20	2014
16	Bull Creek, WY	BCR	Jackson Hole-*Lycaeides*	46	2014
17	Bunsen Peak, WY	BNP	Jackson Hole-*Lycaeides*	20	2014
18	Blacktail Butte, WY	BTB	Jackson Hole-*Lycaeides*	46	2014
19	Pinnacle, MT	PIN	Jackson Hole-*Lycaeides*	20	2014
20	Periodic Springs, WY	PSP	Jackson Hole-*Lycaeides*	20	2014
21	Riddle Lake, WY	RDL	Jackson Hole-*Lycaeides*	30	2014
22	Rendevouz Mountain, WY	RNV	Jackson Hole-*Lycaeides*	32	2014
23	Dubois, WY	DBS	Dubois-*Lycaeides*	115	Present

Table 2.2. S2 Table shows summary of randomization tests for presence of top (0.01‰) SNPs showing excess *L. idas* ancestry frequency on Z chromosome for 10 Jackson Hole-*Lycaeides* localities (x-fold = Number of SNPs observed is how much more than chance; P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion). $P \leq 0.05$ are in bold.

Locality	**No. observed**	**x-fold enrichment**	***P***
BCR	29	1.28	0.078
BLD	66	2.92	< **0.01**
BNP	66	2.93	< **0.01**
BTB	58	2.57	< **0.01**
FRC	65	2.86	< **0.01**
PIN	64	2.83	< **0.01**
PSP	70	3.11	< **0.01**
RDL	52	2.30	< **0.01**
RNV	60	2.65	< **0.01**

Table 2.3. S3 Table shows summary of randomization tests for presence of top (0.01%) SNPs with high genomic cline parameter high α values in Dubois-*Lycaeides* on various regions of the genome (category = region in the genome, x-fold enrichment = Number of SNPs observed is how much more than chance; P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion). $P \leq 0.05$ are in bold.

Category	**No. observed**	**x-fold enrichment**	***P***
ongene	47	1.06	0.3292
neargene	55	1.05	0.3342
oncds	18	1.08	0.3958
nearcds	49	1.09	0.2361
onmRNA	47	1.06	0.3306
nearmRNA	55	1.05	0.3425
onte	3	0.90	0.6618
nearte	17	1.20	0.2341
onprotein	58	0.99	0.5644
nearprotein	79	1.07	0.1619

Table 2.4. S4 Table shows summary of randomization tests for presence of top (0.01%) SNPs with low genomic cline parameter low α values in Dubois-*Lycaeides* on various regions of the genome (category = region in the genome, x-fold enrichment = Number of SNPs observed is how much more than chance; P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion). $P \leq 0.05$ are in bold.

Category	**No. observed**	**x-fold enrichment**	***P***
ongene	47	1.06	0.3308
neargene	53	1.01	0.4894
oncds	24	1.45	0.029
nearcds	48	1.07	0.3037
onmRNA	47	1.06	0.3294
nearmRNA	53	1.01	0.4898
onte	3	0.89	0.6652
nearte	15	1.06	0.4432
onprotein	65	1.11	0.1171
nearprotein	77	1.05	0.2848

Table 2.5. S5 Table shows summary of randomization tests for presence of top (0.01%) SNPs with high genomic cline parameter high β values in Dubois-*Lycaeides* on various regions of the genome (category = region in the genome, x-fold enrichment = Number of SNPs observed is how much more than chance; P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion). $P \leq 0.05$ are in bold.

Category	**No. observed**	**x-fold enrichment**	***P***
ongene	47	1.06	0.3345
neargene	54	1.03	0.4108
oncds	18	1.08	0.4022
nearcds	46	1.02	0.4485
onmRNA	47	1.06	0.3236
nearmRNA	54	1.03	0.4135
onte	4	1.20	0.4303
nearte	15	1.06	0.4497
onprotein	60	1.02	0.4132
nearprotein	74	1.00	0.5201

Table 2.6. S6 Table shows summary of randomization tests for presence of top (0.01%) SNPs showing excess mean *L. idas* ancestry frequency on various genomic regions for Jackson Hole-*Lycaeides* localities (category = region in the genome, x-fold = Number of SNPs observed is how much more than chance; P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion). $P \leq 0.05$ are in bold.

Category	**No. observed**	**x-fold enrichment**	***P***
ongene	53	1.19	0.0511
neargene	63	1.20	**0.0241**
oncds	23	1.38	0.058
nearcds	55	1.22	**0.0249**
onmRNA	53	1.19	0.0518
nearmRNA	63	1.20	**0.0219**
onte	3	0.90	0.6623
nearte	13	0.91	0.675
onprotein	61	1.04	0.351
nearprotein	84	1.14	**0.0217**

Table 2.7. S7 Table shows summary of randomization tests for presence of top (0.01%) SNPs showing excess mean *L. idas* ancestry frequency for Jackson Hole-*Lycaeides* localities and high cline parameter α values in Dubois-*Lycaeides* (category = region in the genome, x-fold = Number of SNPs observed is how much more than chance; P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion). $P \leq 0.05$ are in bold.

Category	**No. observed**	**x-fold enrichment**	***P***
ongene	10	1.21	0.3026
neargene	13	1.31	0.129
oncds	5	1.59	0.1905
nearcds	12	1.42	0.0914
onmRNA	10	1.20	0.2952
nearmRNA	13	1.32	0.1238
onte	0	0	NA
nearte	2	0.75	0.77
onprotein	12	1.09	0.4149
nearprotein	16	1.16	0.2291

Table 2.8. S8 Table shows summary of randomization tests for presence of top (0.01%) SNPs showing excess mean *L. idas* ancestry frequency for Jackson Hole-*Lycaeides* localities and low cline parameter α values in Dubois-*Lycaeides* (category = region in the genome, x-fold = Number of SNPs observed is how much more than chance; P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion). $P \leq 0.05$ are in bold.

Category	**No. observed**	**x-fold enrichment**	***P***
ongene	4	1.17	0.4707
neargene	4	0.99	0.6281
oncds	3	2.31	0.1302
nearcds	4	1.16	0.4758
onmRNA	4	1.17	0.4643
nearmRNA	4	0.99	0.6315
onte	0	0	NA
nearte	0	0	NA
onprotein	6	1.33	0.2517
nearprotein	6	1.06	0.5586

Table 2.9. S9 Table shows summary of randomization tests for presence of top (0.01%) SNPs showing excess mean *L. idas* ancestry frequency for Jackson Hole-*Lycaeides* localities and high cline parameter β values in Dubois-*Lycaeides* (category = region in the genome, x-fold = Number of SNPs observed is how much more than chance; P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion). $P \leq 0.05$ are in bold.

Category	**No. observed**	**x-fold enrichment**	***P***
ongene	23	1.28	0.1211
neargene	27	1.23	0.0908
oncds	8	1.14	0.4031
nearcds	22	1.16	0.2171
onmRNA	23	1.24	0.119
nearmRNA	27	1.23	0.0912
onte	3	2.17	0.1569
nearte	6	1.02	0.5498
onprotein	24	0.98	0.6141
nearprotein	32	1.04	0.4353

Table 2.10. S10 Table shows summary of biological functions of overlapping SNPs in the top (0.1%) quantile which have high *L. idas* ancestry SNPs in Jackson Hole-*Lycaeides* and high genomic cline parameter α values in Dubois-*Lycaeides* (IPR Number = Interproscan number; IPR Term = Term associated with the function associated with IPR number).

Scaffold	**Position**	**IPR Number**	**IPR Term**
11	6234637	IPR001930	Peptidase M1, alanine aminopeptidase/leukotriene A4 hydrolase
11	6234637	IPR014782	Peptidase M1, membrane alanine aminopeptidase
11	6234637	IPR024571	ERAP1-like C-terminal domain
11	15712950	IPR009851	Modifier of rudimentary, Modr
11	15712950	IPR037859	Vacuolar protein sorting-associated protein 37
503	6014039	0	0
833	6668579	0	0
1631	1200324	IPR002289	Gamma-aminobutyric-acid A receptor, beta subunit
1631	1200324	IPR006028	Gamma-aminobutyric acid A receptor/Glycine receptor alpha
1631	1200324	IPR006029	Neurotransmitter-gated ion-channel transmembrane domain
1631	1200324	IPR006201	Neurotransmitter-gated ion-channel
1631	1200324	IPR006202	Neurotransmitter-gated ion-channel ligand-binding domain
1631	1200324	IPR036719	Neurotransmitter-gated ion-channel transmembrane domain superfamily
1631	1200324	IPR036734	Neurotransmitter-gated ion-channel ligand-binding domain superfamily
1631	2570345	0	0
1631	4238526	IPR011011	Zinc finger, FYVE/PHD-type
1631	9958857	IPR006076	FAD dependent oxidoreductase
1631	9958857	IPR006222	Aminomethyltransferase, folate-binding domain
1631	9958857	IPR027266	GTP-binding protein TrmE/Glycine cleavage system T protein, domain 1
1631	9958857	IPR036188	FAD/NAD(P)-binding domain superfamily
1631	10080421	IPR003598	Immunoglobulin subtype 2
1631	10080421	IPR003599	Immunoglobulin subtype
1631	10080421	IPR007110	Immunoglobulin-like domain
1631	10080421	IPR013098	Immunoglobulin I-set
1631	10080421	IPR013783	Immunoglobulin-like fold
1631	10080421	IPR036179	Immunoglobulin-like domain superfamily
1631	10080466	IPR003598	Immunoglobulin subtype 2
1631	10080466	IPR003599	Immunoglobulin subtype
1631	10080466	IPR007110	Immunoglobulin-like domain

Continued on next page

Table 2.10 – continued from previous page

Scaffold	Position	IPR Number	IPR Term
1631	10080466	IPR013098	Immunoglobulin I-set
1631	10080466	IPR013783	Immunoglobulin-like fold
1631	10080466	IPR036179	Immunoglobulin-like domain superfamily
1631	12954191	0	0
1631	14218357	0	0
1631	14760795	IPR000034	Laminin IV
1631	14760795	IPR002049	Laminin EGF domain
1631	14760795	IPR002172	Low-density lipoprotein DL receptor class A repeat
1631	14760795	IPR003598	Immunoglobulin subtype 2
1631	14760795	IPR003599	Immunoglobulin subtype
1631	14760795	IPR007110	Immunoglobulin-like domain
1631	14760795	IPR013783	Immunoglobulin-like fold
1631	14760795	IPR023415	Low-density lipoprotein DL receptor class A, conserved site
1631	14760795	IPR036055	DL receptor-like superfamily
1631	14760795	IPR036179	Immunoglobulin-like domain superfamily
1631	15408704	0	0
1631	15428179	0	0
1631	20910914	IPR000644	CBS domain
1631	20910914	IPR039170	5'-AMP-activated protein kinase subunit gamma-2
1632	6146628	IPR011047	Quinoprotein alcohol dehydrogenase-like superfamily
1639	13668785	0	0
1641	9384153	IPR004878	Otopetrin
1642	9023853	IPR001781	Zinc finger, LIM-type
1648	13426790	0	0
1648	13619133	IPR000076	K/Cl co-transporter
1648	13619133	IPR004841	Amino acid permease/ SLC12A domain
1648	13619133	IPR018491	SLC12A transporter, C-terminal

Table 2.11. S11 Table shows summary of biological functions of overlapping SNPs in the top (0.1%) quantile which have high *L. idas* ancestry SNPs in Jackson Hole-*Lycaeides* and high genomic cline parameter α values in Dubois-*Lycaeides* (GO ID = Gene ontology reference ID; GO name = Name of the function associated with GO ID).

Scaffold	**Position**	**IPR Number**	**IPR Term**
11	6234637	GO:0006508	Proteolysis
11	6234637	GO:0008237	Metallopeptidase activity
11	6234637	GO:0008270	Zinc ion binding
11	15712950	GO:0000813	ESCRT I complex
11	15712950	GO:0032509	Endosome transport via multivesicular body sorting pathway
503	6014039	NA	NA
833	6668579	NA	NA
1631	1200324	GO:0004888	Transmembrane signaling receptor activity
1631	1200324	GO:0004890	GABA-A receptor activity
1631	1200324	GO:0005216	Ion channel activity
1631	1200324	GO:0005230	Extracellular ligand-gated ion channel activity
1631	1200324	GO:0006811	Ion transport
1631	1200324	GO:0016020	Membrane
1631	1200324	GO:0016021	Integral component of membrane
1631	1200324	GO:0034220	Ion transmembrane transport
1631	2570345	NA	NA
1631	4238526	NA	NA
1631	9958857	GO:0005515	Protein binding
1631	9958857	GO:0016491	Oxidoreductase activity
1631	9958857	GO:0055114	Oxidation-reduction process
1631	10080421	NA	NA
1631	10080466	NA	NA
1631	12954191	NA	NA
1631	14218357	NA	NA
1631	14760795	GO:0005515	Protein binding
1631	15408704	NA	NA
1631	15428179	NA	NA
1631	20910914	GO:0032559	Adenyl ribonucleotide binding
1631	20910914	GO:0071900	Regulation of protein serine/threonine kinase activity

Continued on next page

Table 2.11 – continued from previous page

Scaffold	Position	IPR Number	IPR Term
1632	6146628	NA	NA
1639	13668785	NA	NA
1641	9384153	NA	NA
1642	9023853	NA	NA
1648	13426790	NA	NA
1648	13619133	GO:0005215	Transporter activity
1648	13619133	GO:0005887	Integral component of plasma membrane
1648	13619133	GO:0006811	Ion transport
1648	13619133	GO:0015379	Potassium:chloride symporter activity
1648	13619133	GO:0016020	Membrane
1648	13619133	GO:0016021	Integral component of membrane
1648	13619133	GO:0055085	Transmembrane transport
1648	13619133	GO:0071477	Cellular hypotonic salinity response
1648	13619133	GO:1902476	Chloride transmembrane transport

Table 2.12. S12 Table shows summary of biological functions of overlapping SNPs in the top (0.1%) quantile which have high *L. idas* ancestry SNPs in Jackson Hole-*Lycaeides* and low genomic cline parameter α values in Dubois-*Lycaeides* (IPR Number = Interproscan number; IPR Term = Term associated with the function associated with IPR number).

Scaffold	**Position**	**IPR Number**	**IPR Term**
1628	10120991	IPR001680	WD40 repeat
1628	10120991	IPR015943	WD40/YVTN repeat-like-containing domain superfamily
1628	10120991	IPR017986	WD40-repeat-containing domain
1628	10120991	IPR036322	WD40-repeat-containing domain superfamily
1631	520759	NA	NA
1631	1171652	IPR002059	Cold-shock protein, DNA-binding
1631	1171652	IPR011129	Cold shock domain
1631	1171652	IPR012340	Nucleic acid-binding, OB-fold
1631	1171652	IPR019844	Cold-shock conserved site
1632	8419694	NA	NA
1632	8419712	NA	NA
1641	6446919	IPR000418	Ets domain
1641	6446919	IPR003118	Pointed domain
1641	6446919	IPR013761	Sterile alpha motif/pointed domain superfamily
1641	6446919	IPR033077	Ets DNA-binding protein pokkuri
1641	6446919	IPR036388	Winged helix-like DNA-binding domain superfamily
1641	6446919	IPR036390	Winged helix DNA-binding domain superfamily
1641	6453863	NA	NA
1642	12252816	NA	NA
1646	10245305	IPR001254;	Serine proteases, trypsin domain
1646	10245305	IPR009003	Peptidase S1, PA clan

Table 2.13. S13 Table shows summary of biological functions of overlapping SNPs in the top (0.1%) quantile which have low *L. idas* ancestry SNPs in Jackson Hole-*Lycaeides* and high genomic cline parameter α values in Dubois-*Lycaeides* (GO ID = Gene ontology reference ID; GO name = Name of the function associated with GO ID).

Scaffold	**Position**	**GO Number**	**GO Term**
1628	10120991	GO:0005515	Protein binding
1631	520759	NA	NA
1631	1171652	GO:0003676	Nucleic acid binding
1632	8419694	NA	NA
1632	8419712	NA	NA
1641	6446919	GO:0000122	Negative regulation of transcription by RNA polymerase II
1641	6446919	GO:0001709	Cell fate determination
1641	6446919	GO:0003700	DNA-binding transcription factor activity
1641	6446919	GO:0005634	Nucleus
1641	6446919	GO:0006355	Regulation of transcription, DNA-templated
1641	6446919	GO:0043565	Sequence-specific DNA binding
1641	6446919	GO:0045596	Negative regulation of cell differentiation
1641	6446919	GO:0045892	Negative regulation of transcription, DNA-templated
1641	6453863	NA	NA
1642	12252816	NA	NA
1646	10245305	GO:0004252;	Serine-type endopeptidase activity;
1646	10245305	GO:0006508	Proteolysis

Table 2.14. S14 Table shows summary of biological functions of overlapping SNPs in the top (0.1%) quantile which have high *L. idas* ancestry SNPs in Jackson Hole-*Lycaeides* and high genomic cline parameter β values in Dubois-*Lycaeides* (IPR Number = Interproscan number; IPR Term = Term associated with the function associated with IPR number).

Scaffold	**Position**	**IPR Number**	**IPR Term**
4	12406546	IPR003316	E2F/DP family, winged-helix DNA-binding domain
4	12406546	IPR015633	E2F Family
4	12406546	IPR032198	E2F transcription factor, CC-MB domain
4	12406546	IPR036388	Winged helix-like DNA-binding domain superfamily
4	12406546	IPR036390	Winged helix DNA-binding domain superfamily
4	12406546	IPR037241	E2F-DP heterodimerization region
4	12406549	IPR003316	E2F/DP family, winged-helix DNA-binding domain
4	12406549	IPR015633	E2F Family
4	12406549	IPR032198	E2F transcription factor, CC-MB domain
4	12406549	IPR036388	Winged helix-like DNA-binding domain superfamily
4	12406549	IPR036390	Winged helix DNA-binding domain superfamily
4	12406549	IPR037241	E2F-DP heterodimerization region
11	8117866	IPR004117	Olfactory receptor, insect
833	6668579	NA	NA
1628	11727258	IPR007231	Nucleoporin interacting component Nup93/Nic96
1628	12893859	NA	NA
1631	1090637	NA	NA
1631	4238467	IPR011011	Zinc finger, FYVE/PHD-type
1631	5365561	NA	NA
1631	5516528	IPR013602	Dynein heavy chain, domain-2
1631	5516528	IPR024317	Dynein heavy chain, AAA module D4
1631	5516528	IPR024743	Dynein heavy chain, coiled coil stalk
1631	5516528	IPR026980	Dynein heavy chain 6, axonemal
1631	5516528	IPR026983	Dynein heavy chain
1631	5516528	IPR027417	P-loop containing nucleoside triphosphate hydrolase
1631	5516528	IPR035699	Dynein heavy chain, hydrolytic ATP-binding dynein motor region
1631	6146721	NA	NA
1631	7534776	NA	NA
1631	7688497	IPR011989	Armadillo-like helical

Continued on next page

Table 2.14 – continued from previous page

Scaffold	Position	IPR Number	IPR Term
1631	7688497	IPR014768	Rho GTPase-binding/formin homology 3 (GBD/FH3) domain
1631	7688497	IPR015425	Formin, FH2 domain
1631	7688497	IPR016024	Armadillo-type fold
1631	7688497	IPR027651	FH1/FH2 domain-containing protein 3
1631	8161313	IPR001251	CRAL-TRIO lipid binding domain
1631	8161313	IPR008936	Rho GTPase activation protein
1631	8161313	IPR036865	CRAL-TRIO lipid binding domain superfamily
1631	9616373	IPR007110	Immunoglobulin-like domain
1631	9616373	IPR013162	CD80-like, immunoglobulin C2-set
1631	9616373	IPR013783	Immunoglobulin-like fold
1631	9616373	IPR036179	Immunoglobulin-like domain superfamily
1631	10080466	IPR003598	Immunoglobulin subtype 2
1631	10080466	IPR003599	Immunoglobulin subtype
1631	10080466	IPR007110	Immunoglobulin-like domain
1631	10080466	IPR013098	Immunoglobulin I-set
1631	10080466	IPR013783	Immunoglobulin-like fold
1631	10080466	IPR036179	Immunoglobulin-like domain superfamily
1631	10096614	IPR003598	Immunoglobulin subtype 2
1631	10096614	IPR003599	Immunoglobulin subtype
1631	10096614	IPR007110	Immunoglobulin-like domain
1631	10096614	IPR013098	Immunoglobulin I-set
1631	10096614	IPR013783	Immunoglobulin-like fold
1631	10096614	IPR036179	Immunoglobulin-like domain superfamily
1631	10096721	IPR003598	Immunoglobulin subtype 2
1631	10096721	IPR003599	Immunoglobulin subtype
1631	10096721	IPR007110	Immunoglobulin-like domain
1631	10096721	IPR013098	Immunoglobulin I-set
1631	10096721	IPR013783	Immunoglobulin-like fold
1631	10096721	IPR036179	Immunoglobulin-like domain superfamily
1631	12342151	NA	NA
1631	12682671	NA	NA

Continued on next page

Table 2.14 – continued from previous page

Scaffold	Position	IPR Number	IPR Term
1631	13728571	NA	NA
1631	13908210	NA	NA
1631	13944328	NA	NA
1631	14175992	NA	NA
1631	14799474	NA	NA
1631	15003188	NA	NA
1631	15044549	NA	NA
1631	15347619	IPR000219	Dbl homology (DH) domain
1631	15347619	IPR001849	Pleckstrin homology domain
1631	15347619	IPR002017	Spectrin repeat
1631	15347619	IPR011993	PH-like domain superfamily
1631	15347619	IPR018159	Spectrin/alpha-actinin
1631	15347619	IPR035899	Dbl homology (DH) domain superfamily
1631	15428089	NA	NA
1631	15428102	NA	NA
1631	15630833	IPR006114	6-phosphogluconate dehydrogenase, C-terminal
1631	15630833	IPR006115	6-phosphogluconate dehydrogenase, NADP-binding
1631	15630833	IPR006183	6-phosphogluconate dehydrogenase
1631	15630833	IPR006184	6-phosphogluconate-binding site
1631	15630833	IPR008927	6-phosphogluconate dehydrogenase-like, C-terminal domain superfamily
1631	15630833	IPR013328	6-phosphogluconate dehydrogenase, domain 2
1631	15630833	IPR036291	NAD(P)-binding domain superfamily
1631	16563238	NA	NA
1631	16563245	NA	NA
1631	16742878	NA	NA
1631	17005201	NA	NA
1631	17266869	NA	NA
1631	18504663	NA	NA
1631	20910914	IPR000644	CBS domain
1631	20910914	IPR039170	5'-AMP-activated protein kinase subunit gamma-2
1631	21035262	NA	NA
1631	21035340	NA	NA

Continued on next page

Table 2.14 – continued from previous page

Scaffold	Position	IPR Number	IPR Term
1631	21250737	NA	NA
1631	21250758	NA	NA
1631	21303464	IPR000719	Protein kinase domain
1631	21303464	IPR001245	Serine-threonine/tyrosine-protein kinase, catalytic domain
1631	21303464	IPR011009	Protein kinase-like domain superfamily
1639	8438185	IPR001031	Thioesterase
1639	8438185	IPR023102	Fatty acid synthase, domain 2
1639	8438185	IPR029058	Alpha/Beta hydrolase fold
1639	10026163	IPR000648	Oxysterol-binding protein
1639	10026163	IPR001849	Pleckstrin homology domain
1639	10026163	IPR011993	PH-like domain superfamily
1641	9383455	IPR004878	Otopetrin
1642	9017721	IPR001781	Zinc finger, LIM-type
1646	12073166	IPR000225	Armadillo
1646	12073166	IPR009223	Adenomatous polyposis coli protein repeat
1646	12073166	IPR009240	Adenomatous polyposis coli protein, 15 residue repeat
1646	12073166	IPR011989	Armadillo-like helical
1646	12073166	IPR016024	Armadillo-type fold
1646	12073166	IPR026818	Adenomatous polyposis coli (APC) family
1646	12073190	IPR000225	Armadillo
1646	12073190	IPR009223	Adenomatous polyposis coli protein repeat
1646	12073190	IPR009240	Adenomatous polyposis coli protein, 15 residue repeat
1646	12073190	IPR011989	Armadillo-like helical
1646	12073190	IPR016024	Armadillo-type fold
1646	12073190	IPR026818	Adenomatous polyposis coli (APC) family

Table 2.15. S15 Table shows summary of biological functions of overlapping SNPs in the top (0.1%) quantile which have high *L. idas* ancestry SNPs in Jackson Hole-*Lycaeides* and high genomic cline parameter β values in Dubois-*Lycaeides* (GO ID = Gene ontology reference ID; GO name = Name of the function associated with GO ID).

Scaffold	**Position**	**IPR Number**	**IPR Term**
4	12406546	GO:0003700	DNA-binding transcription factor activity
4	12406546	GO:0005667	Transcription factor complex
4	12406546	GO:0006355	Regulation of transcription, DNA-templated
4	12406546	GO:0046983	Protein dimerization activity
4	12406549	GO:0003700;;;	DNA-binding transcription factor activity
4	12406549	GO:0005667	Transcription factor complex
4	12406549	GO:0006355	Regulation of transcription, DNA-templated
4	12406549	GO:0046983	Protein dimerization activity
11	8117866	GO:0004984	Olfactory receptor activity
11	8117866	GO:0005549	Odorant binding
11	8117866	GO:0007608	Sensory perception of smell
11	8117866	GO:0016020	Membrane
833	6668579	NA	NA
1628	11727258	GO:0005643	Nuclear pore
1628	11727258	GO:0017056	Structural constituent of nuclear pore
1628	12893859	NA	NA
1631	1090637	NA	NA
1631	4238467	NA	NA
1631	5365561	NA	NA
1631	5516528	GO:0003777	Microtubule motor activity
1631	5516528	GO:0005524	ATP binding
1631	5516528	GO:0005858	Axonemal dynein complex
1631	5516528	GO:0007018	Microtubule-based movement
1631	5516528	GO:0016887	ATPase activity
1631	5516528	GO:0060285	Cilium-dependent cell motility
1631	6146721	NA	NA
1631	7534776	NA	NA
1631	7688497	GO:0007015	Actin filament organization
1631	8161313	GO:0007165	Signal transduction

Continued on next page

Table 2.15 – continued from previous page

Scaffold	Position	GO Number	GO Term
1631	9616373	NA	NA
1631	10080466	NA	NA
1631	10096614	NA	NA
1631	10096721	NA	NA
1631	12342151	NA	NA
1631	12682671	NA	NA
1631	13728571	NA	NA
1631	13908210	NA	NA
1631	13944328	NA	NA
1631	14175992	NA	NA
1631	14799474	NA	NA
1631	15003188	NA	NA
1631	15044549	NA	NA
1631	15347619	GO:0005089	Rho guanyl-nucleotide exchange factor activity
1631	15347619	GO:0005515	Protein binding
1631	15347619	GO:0035023	Regulation of Rho protein signal transduction
1631	15428089	NA	NA
1631	15428102	NA	NA
1631	15630833	GO:0004616	Phosphogluconate dehydrogenase (decarboxylating) activity
1631	15630833	GO:0006098	Pentose-phosphate shunt
1631	15630833	GO:0016491	Oxidoreductase activity
1631	15630833	GO:0050661	NADP binding
1631	15630833	GO:0055114	Oxidation-reduction process
1631	16563238	NA	NA
1631	16563245	NA	NA
1631	16742878	NA	NA
1631	17005201	NA	NA
1631	17266869	NA	NA
1631	18504663	NA	NA
1631	20910914	GO:0032559	Adenyl ribonucleotide binding
1631	20910914	GO:0071900	Regulation of protein serine/threonine kinase activity
1631	21035262	NA	NA

Continued on next page

Table 2.15 – continued from previous page

Scaffold	Position	GO Number	GO Term
1631	21035340	NA	NA
1631	21250737	NA	NA
1631	21250758	NA	NA
1631	21303464	GO:0004672	Protein kinase activity
1631	21303464	GO:0005524	ATP binding
1631	21303464	GO:0006468	Protein phosphorylation
1639	8438185	GO:0004312	Fatty acid synthase activity
1639	8438185	GO:0009058	Biosynthetic process
1639	8438185	GO:0016788	Hydrolase activity, acting on ester bonds
1639	10026163	NA	NA
1641	9383455	NA	NA
1642	9017721	NA	NA
1646	12073166	GO:0005515	Protein binding
1646	12073166	GO:0008013	Beta-catenin binding
1646	12073166	GO:0016055	Wnt signaling pathway
1646	12073166	GO:0030178	Negative regulation of Wnt signaling pathway
1646	12073190	GO:0005515	Protein binding
1646	12073190	GO:0008013	Beta-catenin binding
1646	12073190	GO:0016055	Wnt signaling pathway
1646	12073190	GO:0030178	Negative regulation of Wnt signaling pathway

Table 2.16. S16 Table gives a list of sequences used from LepBase version 4 to create the protein homology file for Genome Annotation using MAKER pipeline.

	Sequence name
1	Amyelois_transitella_v1_-_proteins.fa
2	Bicyclus_anynana_nBa.0.1_-_proteins.fa
3	Bicyclus_anynana_v1.2_-_proteins.fa
4	Bombyx_mori_ASM15162v1_-_proteins.fa
5	Calycopis_cecrops_v1.1_-_proteins.fa
6	Chilo_suppressalis_CsuOGS1.0_-_proteins.fa
7	Danaus_plexippus_v3_-_proteins.fa
8	Heliconius_erato_demophoon_v1_-_proteins.fa
9	Heliconius_erato_lativitta_v1_-_proteins.fa
10	Heliconius_melpomene_melpomene_Hmel1_-_proteins.fa
11	Heliconius_melpomene_-_proteins.fa
12	Junonia_coenia_JC_v1.0_-_proteins.fa
13	Lerema_accius_v1.1_-_proteins.fa
14	Limnephilus_lunatus_v1_-_proteins.fa
15	Manduca_sexta_Msex_1.0_-_proteins.fa
16	Operophtera_brumata_v1_-_proteins.fa
17	Papilio_glaucus_v1.1_-_proteins.fa
18	Papilio_machaon_Pap_ma_1.0_-_proteins.fa
19	Papilio_polytes_Ppol_1.0_-_proteins.fa
20	Papilio_polytes_Ppol_1.0_Refseq_-_proteins.fa
21	Papilio_xuthus_Pap_xu_1.0_-_proteins.fa
22	Papilio_xuthus_Pxut_1.0_-_proteins.fa
23	Papilio_xuthus_Pxut_1.0_Refseq_-_proteins.fa
24	Phoebis_sennae_v1.1_-_proteins.fa
25	Plodia_interpunctella_v1_-_proteins.fa
26	Plutella_xylostella_DBM_FJ_v1.1_-_proteins.fa
27	Plutella_xylostella_pacbiov1_-_proteins.fa

CHAPTER 3

THE PREDICTABILITY OF GENOMIC CHANGES UNDERLYING A RECENT HOST SHIFT IN *MELISSA* BLUE BUTTERFLIES [1]

Abstract

Despite accumulating evidence that evolution can be predictable, studies quantifying the predictability of evolution remain rare. Here, we measured the predictability of genome-wide evolutionary changes associated with a recent host shift in the Melissa blue butterfly (*Lycaeides melissa*). We asked whether and to what extent genome-wide patterns of evolutionary change in nature could be predicted (1) by comparisons among instances of repeated evolution, and (2) from SNP × performance associations in a lab experiment. We delineated the genetic loci (SNPs) most strongly associated with host use in two *L. melissa* lineages that colonized alfalfa. Whereas most SNPs were strongly associated with host use in none or one of these lineages, we detected a ~two-fold excess of SNPs associated with host use in both lineages. Similarly, we found that host-associated SNPs in nature could also be partially predicted from SNP × performance (survival and weight) associations in a lab rearing experiment. But the extent of overlap, and thus degree of predictability, was somewhat reduced. Although we were able to predict (to a modest extent) the SNPs most strongly associated with host use in nature (in terms of parallelism and from the experiment), we had little to no ability to predict the direction of evolutionary change during the colonization of alfalfa. Our results show that different aspects of evolution associated with recent adaptation can be more or less predictable, and highlight how stochastic and deterministic processes interact to drive patterns of genome-wide evolutionary change.

Introduction

Repeated evolution of similar traits in populations or species under similar ecological conditions has been widely documented, and often involves the same genes or alleles [4, 10, 39, 45, 46]. Such

[1] This manuscript has been published in Molecular Ecology and was coauthored by Lauren K Lucas, Chris C Nice, James A Fordyce, Matthew L Forister and Zachariah Gompert. Permission has been granted by the required coauthors for this research to be included in my dissertation (Appendix A) and copyright permission is included (Appendix B).

repeatability suggests patterns of evolutionary change may be predictable, either by comparison with other instances of evolution (i.e., in the context of parallel or convergent evolution) or from a mechanistic understanding of the sources and targets of selection [34]. The degree to which evolution is repeatable and predictable is of general interest, as high repeatability would suggest a more central role for deterministic evolutionary processes (e.g., natural selection), and perhaps increased constraint or bias in terms of the trait combinations, developmental pathways or mutations that can result in adaptation to a given environment. Such constraints could impose general limits on patterns of biological diversity [37, 42, 53, 55]. But further progress requires moving beyond documenting instances of repeated evolution, and instead quantifying the degree to which, and context in which, evolution is repeatable or predictable, as well as identifying the factors mediating this [e.g., 13, 14, 25, 54, 55].

Instances of repeated evolution provide just one of several ways to assess the predictability of evolution; the predictability of evolution can also be considered in terms of comparisons between (i) experiments linking genotype to phenotype or fitness and (ii) evolutionary patterns in natural populations [3, 34]. For example, field transplant experiments can be used to identify genes or traits under divergent selection between two environments, and one can then ask whether or to what extent patterns of genetic differentiation between natural populations occupying those different environments could be predicted from the experimental results [e.g., 5, 13, 54]. This approach has received relatively little attention compared to direct tests for parallel or convergent evolution in nature [3], but it may have a greater ability to identify the mechanisms underlying predictability by better isolating components of the many evolutionary and ecological processes affecting natural populations [22, 37, 55]. With that said, a lack of consistency between experimental and natural populations can be difficult to interpret, as experiments can miss key features of the natural environment. Whereas both of these approaches (i.e. prediction from experiments and studies of parallelism) have been used in isolation to assess the predictability of evolution, they have rarely been used in a single system and in a comparative manner [but see, e.g., 13, 54].

Here we consider the predictability of genome-wide evolutionary changes (hereafter genomic change) associated with a host-plant shift in the Melissa blue butterfly, *Lycaeides melissa* (Ly-

caenidae). We focus on a quantitative comparison of the two aspects of predictability discussed above. *Lycaeides melissa* occurs throughout western North America, where it feeds on legumes, particularly species of *Astragalus*, *Lupinus*, and *Glycyrrhiza*. *Medicago sativa* (alfalfa, a common forage crop and also a legume) was introduced to western North America in the mid 1800s, and has since been colonized by *L. melissa* [41]. This is a poor host in terms of caterpillar survival, weight, and adult fecundity [16, 17, 52]. Nonetheless, many *L. melissa* populations persist on and have partially adapted to this plant species [e.g., on average, populations on *M. sativa* exhibit increased larval performance and adult oviposition preference relative to populations that do not feed on *M. sativa*; 15, 23]. At present, we do not know whether *L. melissa* colonized *M. sativa* once or multiple times, nor do we know whether the alfalfa-feeding populations are connected by appreciable levels of gene flow. But such information is critical for assessing the degree to which different populations or groups of populations represent independent instances of adaptation, and thus whether they can be used to quantify the repeatability of evolution.

We have additional reasons to be interested in gene flow among *L. melissa* populations. In a previous lab experiment, *L. melissa* caterpillars from populations feeding on *M. sativa* (Goose Lake Ag. = GLA; 41.9860° N, 120.2925° W) and from a population feeding on *Astragalus canadensis*, which is a native host (Silver Lake = SLA; 39.64967° N, 119.92629° W), were reared in a crossed design on either *M. sativa* or *A. canadensis* [23]. We then used a multi-locus genome-wide association mapping approach to identify SNPs associated with variation in larval performance (survival and weight) for each population × host combination. This experiment showed that genetic variants associated with performance on each host plant were mostly independent, and thus, we failed to find evidence for genetic trade-offs in performance across hosts. Such trade-offs are often hypothesized to drive host plant specialization in phytophagous insects [19, 20]. Despite the popularity of this hypothesis, very few studies have found evidence of resource-based trade-offs between hosts [but see 25, 59]. Based on these results, we raised an alternative hypothesis that host plant specialization, and particularly the loss of adaptation to an ancestral host (in this case *A. canadensis*), results from genetic drift in isolated populations that are not well connected by gene flow [similar to 26]. In other words, reduced performance on an ancestral host in an alfalfa feeding population could result solely

from genetic drift if alleles increasing fitness on the ancestral host do not affect fitness on alfalfa, and if alfalfa feeding populations experience little to no gene flow with these ancestral populations. This hypothesis has implications for the repeatability of evolution, as it would predict a greater role for stochastic processes (i.e., genetic drift) in patterns of genomic change (i.e., evolutionary change across the genome) during repeated host shifts than would be expected if trade-offs were prevalent. Evaluating this hypothesis requires additional data on gene flow among *L. melissa* populations.

Herein, we first test whether *L. melissa* has colonized *M. sativa* one or multiple times and quantify levels of contemporary gene flow, and second quantify the predictability of genomic change associated with the colonization of *M. sativa* by *L. melissa*. Specifically, we analyze genotyping-by-sequencing (GBS) data from 26 *L. melissa* populations to ask the following questions: (i). Have *L. melissa* populations colonized the novel host *M. sativa* repeatedly in independent colonization events? (ii) To what extent do parallel genetic changes underlie repeated instances of the colonization of *M. sativa* by *L. melissa*? (iii) To what extent do SNP $\times$ larval performance associations in the aforementioned rearing experiment predict patterns of genetic differentiation between natural populations feeding on alfalfa versus native legume hosts? (iv). Is the degree of predictability higher in the context of (ii) or (iii)? See Fig. 3.1 for a summary of research questions and primary analyses.

Methods

Samples and DNA sequencing

In this study we considered GBS data from 526 *L. melissa* butterflies collected from 26 populations distributed across the western USA (Table 3.1). This includes 15 populations that use *M. sativa* (alfalfa) as a host, and 11 populations that use one of several native legume species (i.e., species of *Astragalus*, *Lupinus* or *Glycrrhiza*). GBS data from 414 of these individuals (20 populations) were previously published in a study of admixture in the *Lycaeides* species complex [24], whereas the GBS data from the other 112 individuals (6 populations) are presented here. DNA extraction, GBS library preparation, and DNA sequencing (100 bp single-end reads with an Illumina HiSeq 2500) occurred concurrently for all 526 samples [for details refer to 22].

Genome alignment and genetic variation in populations

We used the `aln` and `samse` algorithms from `bwa` 0.7.5a-r405 to align 100-bp single-end reads (525 million reads) to our draft *L. melissa* genome [the draft genome is described in 23]. This included re-aligning the data from Gompert *et al.* [24] as those results preceded the current genome assembly and annotation. We allowed a maximum of four differences between each sequence and the reference (no more than two differences were allowed in the first 20 bp of the sequence). We trimmed all bases with a phred-scaled quality score lower than 10 and only placed sequences with a unique best match in our data. We then used `samtools` (version 0.1.19) to compress, sort and index the alignments [36]. We identified (verified) single nucleotide variants and calculated genotype likelihoods, but considering only the set of SNPs identified previously by Gompert *et al.* [23]. The original variant set was called using many of the *L. melissa* samples included here, as well as butterflies from the rearing experiment described above. By focusing on this variant set, we ensured that clear comparisons could be made between the data from the experimental and natural populations in terms of tests of predictability. Variants were called using `samtools` and `bcftools` (version 0.1.19) and were only output if the posterior probability that the nucleotide was invariant was less than 0.01 (with a full prior with $\theta = 0.001$), and if data were present for at least 80% of the individuals. All SNPs from Gompert *et al.* [23] were also identified as SNPs in the current data set based on these criteria, resulting in 206,028 high-quality SNPs. These SNPs had an average sequencing depth of 17.29 (SD = 11.0) per individual. We used an expectation-maximization algorithm to obtain maximum-likelihood estimates of population allele frequencies while accounting for uncertainty in genotypes [based on the calculated genotype likelihoods from `bcftools`; 35, 54].

Colonization history and tests for gene flow

We used a series of analyses to assess the degree of independence in evolutionary change across the *M. sativa* feeding populations. We were interested in independence both in terms of historical colonization and admixture/gene flow, and in terms of contemporary gene flow. We first used principal components analysis (PCA) as an ordination-based approach to examine whether the *M. sativa* populations formed a single coherent cluster in genotype space, as would be predicted if alfalfa

was colonized a single time. We ran the PCA in `R` (version 3.4.1) on the among individual similarity (i.e., genetic covariance) matrix, which was calculated from genotype point estimates for 14,051 common (global minor allele frequency > 5%) SNPs [genotypes were inferred using a mixture model for allele frequencies with k = 2 to 5 source populations, as in 24]. We then used `TREEMIX` (version 1.12) to construct a population graph depicting the relationships among the focal populations [48]. This method first fits a bifurcating tree based on the population allele frequency covariance matrix (based on the maximum likelihood allele frequency estimates), and then adds migration/admixture edges to the tree to improve the fit. Thus, it allowed us to test for both the monophyly of *M. sativa*-feeding populations (i.e., to test whether there were one or multiple successful colonization events), and to ask whether, if alfalfa was colonized multiple times, the populations have since experienced appreciable historical gene flow/admixture which would reduce their evolutionary independence. We rooted the population tree with two *Lycaeides anna* populations, which were set as the outgroup [data from 24] and fit graphs allowing 0-10 admixture events. We calculated the proportion of variance in allele frequency covariances explained by the population graph with varying numbers of admixture events to quantify model fit, and to determine whether individual admixture events substantially improved model fit [48].

We then used stochastic character mapping to estimate the number of host shifts to *M. sativa* based on the tree from `TREEMIX` [7]. We treated host use (native host vs. *M. sativa*) as a trait for ancestral character state reconstruction [as in, e.g., 12, 30]. We fixed the root of the tree as native feeding because of the known recent introduction of *M. sativa* to North America. We used the `make.simmap` function in the `R` package `phytools` (version 0.6-44) for this analysis [50], and based our inference on two Markov chain Monte Carlo (MCMC) runs each with a 10,000 iteration burn-in, 100,000 sampling steps, and a thinning interval of 50. The probabilistic character state simulations used to estimate the number of shifts to *M. sativa* incorporated uncertainty in the character transition matrix.

We then used an assignment-based approach, namely discriminant analysis, to identify individuals that were likely migrants from another population. Our goal here was to assess evidence of contemporary gene flow in terms of actual migrants [we were not attempting to detect later

generation hybrids or estimate admixture proportions; the latter can be found in 24]. We used the `lda` function from the `MASS` package in `R` to assign individuals to populations based on the first four PCs of the genotypic data (see above; these accounted 95% of the genetic variation), and this was done in a pair-wise manner for all populations, although we were most interested in pairs of adjacent populations. We used k-fold cross-validation to estimate assignment probabilities. Results from this set of analyses (described in detail in the Results below) indicated that *L. melissa* have colonized *M. sativa* at least twice (and probably more times than that), once in the western Great Basin and once in the central/eastern Great Basin and Rocky Mountains, and that there has been little gene flow between these groups of populations (see e.g., Fig. 3.2). We thus use these two groups of populations, hereafter referred to as *melissa*-west and *melissa*-east, respectively, to quantify the extent of parallel genomic change associated with alfalfa-colonization and adaptation [experimental evidence of adaptation to alfalfa in general comes from, e.g., 15, 23].

Quantifying the predictability of genomic change

We measured and compared the predictability of genomic change associated with colonization of alfalfa by *L. melissa* in two ways: (i) the degree of parallelism in genomic change during two independent host shifts onto *M. sativa* in nature, and (ii) how well patterns of genomic change in nature could be predicted from performance × SNP associations in a rearing experiment. We did this by testing for and quantifying an excess overlap in SNPs associated with host use, that is, the SNPs with the greatest allele frequency differences between native and alfalfa-feeding populations in nature and the SNPs most strongly associated with performance in the rearing experiment. We report these values as x-fold enrichments. As an example, an x-fold enrichment of 2.0 would imply that twice as many SNPs are associated with host use in, e.g., repeated instances of colonization of *L. melissa*, as expected by chance (see details of null models below) and thus would mean that exceptional patterns of genomic change can be predicted from one colonization event to the other twice as well as would be the case with no information. We considered x-fold enrichments as measures of predictability both in terms of the SNPs showing host association and in terms of the direction of these effects. In other words, we distinguished between being able to predict host-associated SNPs, and being able to predict the direction of the association. As populations will necessarily vary in the details of

linkage disequilibrium (LD) between causal variants and genetic markers (such as in our SNP set), the former might be more predictable than the later (see the Discussion for details).

Delineating SNP × host use associations in nature

We first delineated the SNPs most strongly associated with feeding on *M. sativa* in nature. We used the software package `BAYPASS version 2.1` [21] to do this by identifying SNPs with the greatest allele frequency differences between populations feeding on *M. sativa* and those feeding on native hosts; this method controls for background population genetic structure. We were interested in these SNPs as they presumably exhibited the greatest change in allele frequencies following the colonization of alfalfa, and some subset of them might be in LD with causal variants affecting host (alfalfa) adaptation (given the sparsity of GBS data, we doubt that any of these SNPs directly confer host adaptation, but this is not critical for our questions and approach).

The `BAYPASS` software used here is based on the `BAYENV` method introduced by Günther & Coop [27]. `BAYPASS` uses a hierarchical Bayesian model with a binary auxiliary variable to classify each locus (i.e., SNP) as associated or unassociated with some environmental covariate. The model attempts to control for population genetic structure by approximating the history of the populations with an allele frequency variance-covariance matrix. We ran `BAYPASS` [21] with three sets of populations: (i) all populations, (ii) 8 *melissa*-west populations, and (iii) 17 *melissa*-east populations. We treated host use, coded as a binary variable indicating whether or not a population was on alfalfa, as the environmental covariate, and ran the standard covariate model. For each data set, we ran four MCMC simulations, each with a 20,000 iteration burnin and 50,000 sampling iterations with a thinning interval of 100. The regression coefficient (β_i) describing the association of each SNP (i) with host use (after controlling for population structure) was calculated using the default option of importance sampling, which also allows for computation of Bayes factors. Bayes factors were used to compare the marginal likelihoods of models with non-zero versus zero values of β_i.

To further characterize the top host-associated SNPs, we conducted additional tests wherein we asked whether the SNPs most associated with host use were overrepresented on the Z chromosome [in butterflies males are ZZ and females are ZW, and the Z chromosome tends to harbor an excess

of QTL for adaptive traits; 32, 56], or whether they were enriched for specific gene ontology (GO) classifications. Such enrichments might be expected if the top host-associated SNPs were indeed tagging (via LD) genetic regions affecting host use. We defined “host-associated SNPs" for these and subsequent tests as those with the largest Bayes factors from the `BAYPASS` analysis. We did this using empirical quantiles, and considered a range of cut-offs, from the top 0.1% to the top 0.01% of SNPs (with increments of 0.01%). Considering multiple quantile cut-offs (here and in additional analyses described below) let us evaluate the sensitivity of our results to particular empirical quantiles. We used a new linkage map (Gompert *et al.*, manuscript in prep.) to classify SNPs as autosomal, Z-linked, or unknown. SNPs were classified as in coding regions (exons only), genic (in gene exons or introns), or intergenic based on the structural annotation described in Gompert *et al.* [23]. GO annotation were based on 14713 `PFAM-A` matches from `INTERPROSCAN`; GO terms were assigned to SNPs within 1 kb of annotated genes. Randomization tests were used to quantify and assess the significance of enrichments for each quantile cut-off and all three data sets, all 25 populations, *melissa*-east, and *melissa*-west, and in each case 1000 randomizations were conducted.

Tests of parallel genomic change in nature

Our first framework for quantifying predictability was to test for parallel evolution of host use between two groups of *L. melissa* populations. Following the `TREEMIX` results, we used the *melissa*-east (N=17) and *melissa*-west (N=8) populations to ask if parallel genetic changes underlie host plant use in these butterfly populations. We used randomization tests (10000 randomizations per test) to generate null expectations for the proportion of top host-associated SNPs shared between *melissa*-east and *melissa*-west populations, and tested if this was more than expected by chance (x-fold enrichments). Herein, we refer to this procedure as *ran1* (see Fig. 3.1). We performed this randomization procedure twice, first with raw Bayes factors and again using residuals from regressing Bayes factors on mean allele frequencies (averaged over the relevant populations) (we focus on the latter in the Results). We repeated *ran1* considering the top 0.01% to 0.1% (with 0.01% increments) host-associated SNPs to determine whether the degree of parallelism (i.e., predictability in the context of repeated evolution) was robust to different cut-offs for defining host-associated SNPs.

Next, we asked whether the top host-associated SNPs that were shared between *melissa*-east and *melissa*-west populations showed a consistent direction in terms of the allele frequency difference between populations feeding on alfalfa versus those on native hosts. We would expect differences in a consistent direction if the same allele was favored in both colonization events and if patterns of LD (including the sign of D, the coefficient of linkage disequilibrium) between causal variants and our SNP markers were consistent between these groups of populations. We tested for consistency in the sign of allele frequency differences using both raw allele frequencies and standardized allele frequencies from `BAYPASS`; the latter are residuals after controlling for background population structure (we focus on the latter in the Results). For each SNP and group (*melissa*-east or *melissa*-west) we calculated $\delta_p = \bar{p_a} - \bar{p_n}$, where $\bar{p_a}$ and $\bar{p_n}$ are the mean (raw or standardized) allele frequencies for the alfalfa and native-feeding populations, respectively. For the top shared host-associated SNPs, we enumerated the cases where the sign of the allele frequency difference (δ_p) was the same in *melissa*-east and *melissa*-west. We conducted two sets of randomization tests to compare this to null expectations. First, we asked whether the number of shared top host-associated SNPs with the same sign for δ_p was greater than expected if the δ_p vectors in *melissa*-east and *melissa*-west were independent. We did this by permuting one of the sign vectors; we refer to this procedure as *ran2A* (see Fig. 3.1). In an additional randomization test (hereafter *ran2B*), we asked whether a greater proportion of the shared top host-associated SNPs had the same sign for δ_p than expected based on sign overlap for the rest of the SNPs. This was done by permuting the classification of SNPs as shared top host-associated or not.

Predictability of patterns in natural populations from experimental outcomes

We next asked how well SNP × host association in nature can be predicted from SNP × performance association from a published lab experiment [23]. In [23], we quantified the association between each SNP and host-specific survival or adult weight as a model-averaged locus effect (MAE) by fitting Bayesian Sparse Linear Mixed Models (BSLMMs). This method includes a genetic relatedness matrix and the genotype of each individual at each SNP as predictors of each individual's phenotype [62]. MAEs are given by the formula $\hat{b_j} = \beta_j \gamma_j + \alpha_j$, where β_j is SNP j's main effect if it is included in the model (i.e., if it has a main effect), γ_j is the posterior inclusion

probability for SNP j (i.e. the probability that SNP j has a main effect, that is, that it tags a causal variant), and α_j is SNP j's contribution to the phenotypic variation via the genetic relatedness matrix. In this experiment, we estimated MAE for the following treatments: (i) larvae from GLA (which feeds on *M. sativa*) reared on *M. sativa* (GLA-Ms) 2) larvae from GLA reared on *A. canadensis* (GLA-Ac) 3) larvae from SLA (which feeds on *A. canadensis*) reared on *M. sativa* (SLA-Ms) and 4) larvae from SLA reared on *A. canadensis* (SLA-Ac). We used these MAEs to delimit SNPs with the greatest association with host-specific performance in the experiment. We then asked whether these SNPs also showed substantial allele frequency differences between alfalfa and non-alfalfa feeding populations in nature. We enumerated the SNPs that were top host-associated SNPs in the `BAYPASS` analysis and that were top performance-associated SNPs (based on the MAEs). We considered classifications based on the top 0.1% to 0.01% of SNPs and based on each experimental treatment and performance metric (weight or survival) [different SNPs were associated with performance on each host; 23]. Comparisons were made between the experimental results and `BAYPASS` results for: *melissa*-east, *melissa*-west and all populations. We used the *ran1* approach (Fig. 3.1) to test for and quantify an excess of overlap between the top host-associated (in nature) and top performance-associated (in the experiment) SNPs.

Next, we asked whether, for the shared top host-associated and top performance-associated SNPs, the direction of allele frequency differences in nature (δ_p, see the previous section) was consistent with the direction of the SNP $\times$ performance association from the experiment (see Fig. 3.7). For example, an allele associated with increased survival on alfalfa in the experiment would be predicted to be at higher frequency in the alfalfa-feeding populations. We considered all combinations of definitions for top host-associated SNPs (*melissa*-east, *melissa*-west, all populations) and top performance-associated SNPs (all experimental treatments and both weight and survival), and in each case enumerated the instances where the sign for δ_p and the sign of the MAE were consistent. Then, as we did for the tests of parallelism in nature (see preceding section), we used randomization tests to ask whether and to what extent there was more consistency than expected (i) assuming the direction of SNP $\times$ host and SNP $\times$ performance association were independent (*ran2A*), and (ii) based on the consistency of δ_p and MAEs for all other SNPs (*ran2B*).

Even if SNP × performance associations were not generally predictive of SNP × host association in nature, they could be locally predictive of exceptional genetic differentiation between the specific populations used in the rearing experiment. To test this, we first quantified genetic differentiation between GLA and SLA at each SNP locus using Hudson's estimator of F_{ST} [6]. Then, similar to the analyses described above, we identified the most differentiated loci between GLA and SLA, that is the 0.1% to 0.01% most differentiated SNPs, and tested for significant overlap between these SNPs and the top performance-associated SNPs (*ran1*), and for whether the direction of allele frequency difference between these populations was consistent with the direction of the SNP × performance association from the experiment (*ran2A* and *ran2B*). We then repeated these analyses with the most differentiated SNPs between GLA and a relatively close native-feeding population (ABC, host = *Lupinus*, distance from GLA = 220.8 km), and with SLA and a nearby alfalfa-feeding population (VCP, distance from SLA = 17.5 km) (GLA and SLA are themselves 184.9 km apart). Here, we considered only the performance-associated SNPs from GLA (for the GLA × ABC comparison) or the performance-associated SNPs from SLA (for the SLA × VCP comparison).

Results

Colonization history and tests for gene flow

Ordination with PCA indicated that most (86.7%) of the genetic variation in the samples was accounted for by the first two principal components. These PCs largely separated individuals and populations based on geography rather than host plant (Fig. 3.2). The best bifurcating tree from `TREEMIX` explained 94.3% of the variance in population covariances. Consistent with the PCA results, *L. melissa* populations formed two major clades that grouped populations by geography; each major clade included a mixture of populations feeding on alfalfa and native hosts (Fig. 3.2C). Adding migration edges to the tree increased the variance explained (Fig. 3.8), with the biggest gain from a single migration edge. This tree (graph) explained 97% of the variation in the data and allowed for gene flow from the outgroup *Lycaeides anna* to a single high-elevation *L. melissa* population at Albion meadows, UT. As such gene flow is unlikely in terms of geography, and because this population is phenotypically distinct from other *L. melissa*, it was excluded from further analyses.

Stochastic character mapping of host use on the TREEMIX tree suggested four shifts from native hosts to *M. sativa* (95% credible intervals [CIs] = 2–10, posterior probability of two or more shifts = 0.95), with seven likely reversals back to native feeding (95% CIs = 4–11) (Fig. 3.9). Results from these analyses indicate that *L. melissa* have colonized *M. sativa* at least twice (and probably more times than that), once in the western Great Basin and once in the central/eastern Great Basin and Rocky Mountains, and that there has been little historical gene flow between these groups of populations.

With discriminant analysis, most individuals were confidently assigned to the population from which they were sampled (average assignment probability to the population of origin = 0.9918; Figs. 3.10, 3.11). Mean assignment probabilities to the population of origin were similar for same (0.964, sd = 0.0524) and different (0.984, sd = 0.953) host comparisons. Very few individuals were confidently assigned to the alternative population, that is, the one they were not sampled from (0 in 221 population pairs, 1 in 66 pairs, and 2 in 13 pairs, assignment prob. > 0.9), and we never had more than two individuals assigned to the population they were not collected from (Fig. 3.11). Based on all of these results we used the two clades which included populations located in the eastern and western geographical ranges of the species (hereafter *melissa*-east [N=17 populations] and *melissa*-west [N=8 populations]) to test for predictable genetic changes underlying host plant use (Fig. 3.2A, Table 3.1).

Delineating SNP × host use associations in nature

Before formally quantifying the predictability of genomic change, we identified SNPs associated with host plant use in *L. melissa* populations, which we refer to as “host-associated SNPs". Most SNPs across *melissa*-east and *melissa*-west populations had low Bayes factors (Fig. 3.12). However, Bayes factors were large (i.e., > 5, meaning the likelihood of the host-association model was at least five times greater than the null model) for some SNPs (1068 in *melissa*-west and 1611 in *melissa*-east) (Fig. 3.12).

Here we report the results for top 0.01% host-associated SNPs (N=2061). For all populations, an excess of host-associated SNPs were present on the Z-chromosome (obs. = 195, x-fold enrichment = 2.26, $P < 0.01$; randomization test). Similar results were seen for *melissa*-east (obs. = 193; x-fold enrichment = 2.23, $P < 0.01$; randomization test) and *melissa*-west populations (obs. = 134; x-fold

enrichment = 1.55, $P < 0.01$; randomization test) (also see Table 3.2). A significant excess of the host-associated SNPs were also present in gene exons (for all populations; x-fold = 1.45; $P < 0.01$; randomization test, Tables 3.3, 3.4). These results held for *melissa*-east (x-fold enrichment = 1.46; $P < 0.01$; randomization test) and *melissa*-west (x-fold enrichment = 1.46; $P < 0.01$; randomization test) populations. Gene ontology (GO) analysis revealed that the host-associated SNPs are present in regions of the genome containing genes involved in a range of biological and cellular processes, and exhibit various molecular functions (Tables 3.5, 3.6, 3.7). Some GO terms are over-represented among the top host-associated SNPs, but not enough so to warrant particular attention at this time.

Parallel genomic change in nature

For the top 0.01% SNPs (N=2061) with the largest Bayes factors in *melissa*-east or *melissa*-west from `BAYPASS` analysis, there was a significant excess of overlap, such that more SNPs were highly associated with host use in *melissa*-east and *melissa*-west than expected by chance (*ran1*, obs. = 58 shared SNPs, expected = 21, x-fold enrichment = 2.82, $P < 0.01$; Fig. 3.3). Six of the 58 shared top host-associated SNPs were on the Z-chromosome, which is also an excess relative to null expectations (randomization test, x-fold = 2.43; $P = 0.03$; Table 3.2). Nonetheless, the majority of top host-associated SNPs differed between *melissa*-east and *melissa*-west, resulting in a low overall correlation in Bayes factors (Pearson $r = 0.06$; $P < 0.01$). We found that the x-fold enrichments for shared top host-associated SNPs held across a range of empirical quantiles, with the greatest excess seen in the most extreme quantiles (Figure 3.4).

We found minimal evidence of concordance in the direction of allele frequency differences between alfalfa and native-feeding population when comparing *melissa*-east and *melissa*-west and considering the shared top host-associated SNPs. Specifically, for the top 0.01% SNPs (N=2061) with the largest Bayes factors in both population groups, we found no evidence of greater than expected concordance in the sign of allele frequency differences between alfalfa and native-feeding populations (*ran2A*) based on standardized or raw allele frequencies (standardized: x-fold = 1.04, $P = 0.222$; raw: x-fold = 1.05, $P = 0.257$). For the same empirical quantile, we found limited and weak evidence of greater sign coincidence for the shared top host-associated SNPs than random SNPs (*ran2B*) based on the standardized allele frequencies (x-fold = 1.05, $P = 0.05$), and a slight excess

of sign coincidence based on the raw allele frequencies (x-fold = 1.21, $P = 0.046$).

Predictability of natural processes from experimental outcomes

We found some cases where there was greater overlap than expected by chance between SNPs most associated with performance in the rearing experiment (top performance-associated SNPs) and those most associated with host use in nature (top host-associated SNPs), but this depended on the specific comparison being considered (here we again focus on results for top 0.01% SNPs, N=2061, but also provide results for other quantiles graphically; Figs. 3.5, 3.6, 3.13). For example, we found an excess overlap in top survival-associated SNPs for GLA reared on *M. sativa* and top host-associated SNPs for all *L. melissa* (*ran1*, obs. = 14, x-fold enrichment = 1.74X, $P = 0.03$), but not for *melissa*-east or *melissa*-west (these were marginally significant with $P = 0.06$ and 0.07, respectively). We found excess overlap in top survival-associated SNPs for GLA reared on *A. canadensis* and top host-associated SNPs for (i) *melissa*-east (*ran1*, obs. = 16, x-fold enrichment = 1.95X, $P = 0.01$), and (ii) *melissa*-west (*ran1*, obs. = 14, x-fold enrichment = 1.71X, $P = 0.04$) (Table 3.8). Survival-associated SNPs in SLA were not predictive of host-associated SNPs in nature, but we did detect an excess of shared weight-associated SNPs in SLA and host-associated SNPs. For example, top weight-associated SNPs for SLA when reared on *A. canadensis* overlapped more than expected by chance with *melissa*-east host-associated SNPs (*ran1*, obs. = 17, x-fold enrichment = 2.23; $P = < 0.01$; Table 3.9). In addition, for top weight-associated SNPs for SLA reared on *M. sativa* overlapped more than expected by chance with *melissa*-west host-associated SNPs (*ran1*, obs. = 19, x-fold enrichment = 2.5; $P < 0.01$; Table 3.9). We found a single case of excess overlap in top weight-associated SNPs for GLA reared on *M. sativa*, which was with the top host-associated SNPs for *melissa*-west (*ran1*, obs. = 16, x-fold enrichment = 2.06X, $P = 0.01$).

For the top 0.01% SNPs (N=2061) host-associated SNPs and performance-associated SNPs, we found weaker evidence for and a lesser degree of concordance in the direction of allele frequency differences between alfalfa and native-feeding populations (δ_p) and signs of model average effects (MAE) of performance-associated SNPs (Figs. 3.14, 3.15, 3.16, 3.17; Tables 3.10, 3.11). Most notably, there was a modest excess of sign coincidence for performance (survival)-associated SNPs for SLA reared on *M. sativa* and host-associated SNPs in *L. melissa* east (*ran2A* and *ran2B*, x-fold

enrichment = 1.39–1.59X, $P \leq 0.01$) and west (*ran2A* and *ran2B*, x-fold enrichment = 1.42–1.53X, $P \leq 0.03$). We also found excess sign concordance for survival-associated SNPs for GLA reared on *M. sativa* and host associated SNPs in *L. melissa* east (*ran2A*, x-fold enrichment = 1.21, $P = 0.05$; *ran2B*, x-fold enrichment = 1.38, $P = 0.04$; Table 3.10).

We generally found greater overlap and sign-consistency between performance-associated SNPs and the most differentiated SNPs between GLA and SLA in nature than between performance-associated SNPs and more broadly host-associated SNPs in nature (described in the previous two paragraphs). We found substantial overlap between the top 0.01% of performance-associated and the top 0.01% most differentiated SNPs ($F_{ST} \geq 0.23$), with x-fold enrichments ranging from 4.53 (weight-associated SNPs for SLA reared on *A. canadensis*) to 1.42 (survival-associated SNPs for SLA reared on *A. canadensis*), with $P < 0.05$ for all but one of these comparisons (Table 3.12). Somewhat weaker overlap was detected when considering the most differentiated SNPs between GLA and ABC or SLA and VCP (x-fold enrichment = 0.77–2.06), but still in most cases the overlap was greater than expected by chance. And in general, the results were consistent across different top-SNP quantiles (Fig. 3.18). Results for tests of sign coincidence were more idiosyncratic, but with most cases of significant excess overlap between the sign of genetic differentiation between populations and the sign of the SNP×performance effect estimate involving experimental populations reared on *M. sativa*, though there was also some evidence of significant excess involving survival-associated SNPs for SLA reared on *A. canadensis* (Tables 3.13). With that said, x-fold enrichment across all comparisons never exceeded 1.55X (for *ran2B*, F_{ST} for GLA vs. SLA × weight-associated SNPs for GLA reared on *M. sativa*). Similar results were obtained for other quantile cut-offs (Figs. 3.19, 3.20).

Discussion

Several studies have shown that evolution can be predicted, at least in part, but fewer studies have quantified the degree of predictability, and in general less attention has been paid to the predictability of genomic changes underlying complex life history traits [55]. Here we first showed that *L. melissa* butterflies have colonized a novel host plant (alfalfa, *M. sativa*) two or more times, with little to no gene flow connecting the two clades of alfalfa-feeding butterflies. We used these two independent

instances of colonization and results from a rearing experiment to quantify the degree to which genomic changes following alfalfa colonization were predictable. We found a modest overlap in SNPs showing the greatest allele frequency differences between alfalfa and native-feeding *L. melissa* in the western and eastern populations (~1.5–2.8 times more than expected by chance, depending on the quantile considered), and a significant but weaker and more idiosyncratic excess in overlap of SNPs associated with host-specific larval performance in an experiment and those with the greatest allele frequency differences between host-associated populations in nature (~0.53–2.5 times more than expected by chance, depending on the quantile considered and performance measure). Although we were able to predict (to a modest extent) the SNPs with the greatest genomic change in nature (in terms of parallelism in nature and from the experimental results), we generally had little to no ability to predict the direction of change (i.e., even if the same SNP showed exceptional change in *melissa*-east and *melissa*-west, the direction of change was not necessarily the same). SNP × performance associations were, however, more predictive of patterns of genetic differentiation in nature between the specific populations used in the rearing and mapping experiments. We discuss and interpret these results in more detail below.

Predictability of genomic changes associated with a host shift

We identified a significant excess of shared SNPs between *melissa*-east and *melissa*-west populations, specifically ~ 1.5–2.8 times more than expected by chance. This means that knowing which SNPs exhibited the greatest genomic change in one geographic group (i.e., in one colonization event) improves our ability to predict those with the greatest genomic change in the other group about two-fold. In some ways, such predictability is not surprising as the western and eastern populations likely had access to much of the same standing genetic variation [10]. But there are also reasons to think that parallelism at the genetic level (and thus predictability of genomic change) might be more limited. For example, alfalfa is not a homogeneous resource, and our previous work has documented variation in caterpillar performance based on the source of alfalfa [28], which suggests that the way in which a population adapts to alfalfa might depend on the specific host plant population. Interestingly, a nearly identical excess of parallel genomic change/genetic differentiation was detected in comparisons of host-associated stick insects (*Timema cristinae*) where different

cryptic color patterns are favored on different hosts [54]. For simpler morphological traits, such as armor plating in sticklebacks, the degree of parallelism is considerably higher (91% of the high genetic differentiation regions shared between marine and freshwater populations across population pairs are also shared between additional populations) [33]. Another study in sticklebacks shows high parallel allele frequency changes between lake and stream ecotype populations in genomic regions associated with incipient ecological speciation (51% of the genomic islands of differentiation show parallel changes) [38].

SNP × performance associations in lab experiments predicted genomic change in nature, but in a more limited and more idiosyncratic way; x-fold enrichments for survival range from 0.63–1.95 and x-fold enrichments for weight range from 0.53–2.5, with only ~25% of combinations showing significant excess. Predictability was notably higher in terms of predicting the most differentiated SNPs between the populations used in the rearing experiment, that is, GLA (host = *M. sativa*) and SLA (host = *A. canadensis*); the x-fold enrichment ranged from 1.42 to 4.53 (across comparisons), with all but one case significant. Given the simplified lab rearing environment (e.g., no predators, controlled growth conditions, only some fitness components considered, etc.), it is intriguing that the experiment provided even these level of predictive power about genomic change in nature. With that said, these results are consistent with two other recent studies that predicted genomic change in nature from short-term experiments. In *Timema cristinae* stick insects, Soria-Carrasco *et al.* [54] found a modest but significant overlap between genetic regions associated with survival in a field experiment and those most differentiated between hosts in nature (obs. = 32 shared loci; expected = 23; x-fold enrichment = 1.4). In a similar study with *Rhagoletis pomonella* fruit flies, genomic change during a lab selection experiment was even more predictive of patterns of differentiation in nature (obs. = 154 shared loci; expected = 53.6; x-fold = 2.87) [13]. Substantial genomic change and increased predictability in *Rhagoletis* might be due, at least in part, to the high levels of LD in that system.

Beyond host-use in herbivorous insects, a limited number of studies have tried to predict genome-wide patterns of genetic differentiation in nature from lab or field experiments, and mostly these have involved predicting genetic differentiation from QTL studies. The outlier loci underlying

highest genetic differentiation between *Drosophila yakuba* mainland (Cameroon and Kenya) and Mayotte populations, show concordance with *Drosophila sechellia* noni-tolerance (performance) QTLs (four of nine tolerance QTL, expected = 0.125, $P = 0.013$) but there is no overlap for *D. sechellia* preference QTLs (expected = 0.35, $P = 0.37$), suggesting that noni-performance is more predictable than noni preference [58, 61]. Similarly, QTL for ecologically relevant traits co-localized with possible genetic regions affected by selection (as identified in genome scans) more than expected by chance in comparisons of lake whitefish ecotypes [51]. The predictability of evolution from lab selection and mapping experiments will depend on whether the same genetic variants affect key traits in similar ways in the lab and nature. This has been investigated in several taxa. For example, in *Arabidopsis thaliana*, only one QTL associated with flowering time in the greenhouse also affected flowering time in a field experiment meant to better approximate nature [8, 60].

Despite results suggesting the SNPs with the greatest genomic change during host adaptation could be predicted, both via parallel change in nature and from a short-term rearing/mapping experiment, we found much less evidence that the direction of genomic change was predictable (there were a few, limited exceptions). Even if the same alleles are repeatedly favored on alfalfa (including in the lab), the direction of change at genetic markers could vary if patterns of LD between sequenced SNPs and causal variants differ. For example, if the favored allele at a causal locus is positively associated with one SNP marker allele in one population and negatively associated with the same SNP marker allele in another population, selection on the causal locus could drive substantial evolutionary change at the SNP locus in both populations, but in opposite directions. Such shifts in patterns of LD could occur when new populations are founded (possibly by one or a few gravid females), and thus, this phenomenon could explain our results. Consistent with this possibility, we found slightly more cases (18.8% vs. 14.6% of tests with $P \leq 0.05$) of excess sign coincidence when comparing the experimental results to genetic differentiation between GLA and SLA than when comparing them to overall patterns of SNP $\times$ host-use association. In a related sense, if recombination rates vary across the genome, regions of exceptional genomic change might be predictable if they are simply the regions with the lowest recombination rate and thus the lowest local effective population size. Of course, the rate of evolutionary change by drift or selection is

proportional to the allele frequencies (specifically to $p(1-p)$), which could make the regions of greatest change (but not their direction) predictable, but this is something we have controlled for by working with the residuals after regressing change (or effect sizes from the rearing experiment) on allele frequencies.

In contrast to our results, both the magnitude and direction of genomic change between host races of *Rhagoletis pomonella* fruit flies was predictable from a controlled experiment [13]. Two factors likely contribute to this difference. First, the same populations were used in the experiment and for the natural comparisons, whereas our experiment focused on a small subset of the natural populations we analyzed (but see the discussion of the comparison with genetic differentiation between GLA and SLA above). Second, inversions and large blocks of LD are common in *Rhagoletis* and could increase the consistency of evolutionary patterns and associations between SNP markers and causal variants. Although not concerned with host adaptation, another study which has tested for direction of allele frequency changes underlying rapid adaptation focuses on adaptation to fragmented landscapes in Glanville Fritillary butterflies, *Melitaea cinxia* [18]. This study reports predictable allele frequency changes in most divergent outlier loci between newly colonized versus old local populations, and these allele frequency shifts are in the same direction indicating that selection can drive particular candidate genetic regions in the direction of adaptation to fragmented landscapes (Extinct populations: linear model, $r^2 = 0.36$, $P = 0.02$; Introduced populations: linear model, $r^2 = 0.14$, $P = 0.13$).

Finally, we found higher predictability of genomic change in terms of the greater overlap of top host-associated SNPs between two host shifts onto alfalfa than overlap between performance-associated SNPs in an experiment and host-associated SNPs in nature (predictability in terms of overlap was even higher for patterns of genetic differentiation between the population pair used in the experiment). Perhaps this is not surprising, as we might expect greater similarity in conditions across the natural populations than between the lab experiment and natural populations. Indeed, perhaps it is more surprising that the difference in predictability (1.5–2.7 x-fold enrichment via parallelism in nature vs. 0.53–2.5 x-fold enrichment via predictions from the lab to nature) wasn't greater. This means that genetic and phenotypic determinants of caterpillar performance in the lab have some

bearing on fitness, and thus evolutionary change during host shifts in nature. Posing the exciting possibility that, by integrating outcomes from multiple experiments probing different populations and different components of a host shift (e.g., larval performance, adult preference, etc.), it might be possible to build a more mechanistic model to predict genomic change than would ever be possible by only examining patterns of parallel change in nature.

Interpretation of demographic patterns

We found little evidence for contemporary gene flow among *L. melissa* populations, even those separated by only a few kilometers. This suggests that gene flow from populations feeding on native hosts to populations feeding on alfalfa is not a major factor constraining host adaptation. Moreover, coupled with our previous results indicating a lack of genetic trade-offs for larval performance across hosts [22], this suggests that host plant specialization in *L. melissa* could occur via the loss of adaptation to an ancestral host by genetic drift. Similarly, a recent study in moths (*Thyrinteina leucoceraea*) found that the loss of adaptation to a native host was due to mutation accumulation rather than trade-offs [26]. Thus, whereas resource-based genetic trade-offs do drive host specialization in some systems [25], this and other recent work indicates that other processes that need not include selection can lead to host-plant specialization as well.

Our results strongly suggest that *L. melissa* colonized alfalfa multiple times since the introduction of this plant to North America; at least twice and probably ~four times. Shifts from *M. sativa* to native hosts appear to be even more common, which is consistent with our own observations that populations associated with *M. sativa* are more ephemeral (i.e., less likely to persist over multiple decades) than those feeding on native hosts. Still, considerable uncertainty exists in our estimates of the exact number and nature of these host shifts. Along these lines, more than one colonization event likely occurred within our eastern and western *L. melissa* groups. Thus, while treating these groups as our level of replication is conservative, it also means that the metrics of parallelism discussed above probably represent averages over these putative additional host shifts, and that the true history of colonization is more complex than captured in these analyses.

Genomic context of the host-associated SNPs

An excess of the SNPs most associated with host use in nature were on the Z chromosome in *L. melissa*. This is consistent with findings from other studies suggesting a disproportionate role for sex chromosomes in adaptation and speciation, such as beak morphology in Darwin's finches [1, 2], coat color in mice (*Chaetodipus intermedius*) [29, 43], reduction in armor plating in threespine sticklebacks (*Gasterosteus aculeatus*) [9], and wing pattern variation in *Heliconious* butterflies [47]. In butterflies, other studies of host plant specialization have found evidence of a disproportionate role of Z-linked genes in host specificity and oviposition preference [40]. For example, oviposition preference in the comma butterfly (*Polygonia c-album*) is Z-linked [31, 44]. Similarly, oviposition differences in some swallowtail butterflies (*Papilio zelicaon* and *Papilio oregonius*) are known to be sex-linked [57], and in general, species differences in butterflies often map to the Z chromosome [49, 56]. With that said, it is important to note that the Z chromosome likely has a lower effective population size than the autosomes (this depends some on patterns of mating), and thus the signal of an excess of host-associated SNPs on the Z chromosome could partially reflect genetic drift.

We also found an excess of top host-associated SNPs on the coding regions of genes (1.45 times more than expected by chance). This does not necessarily imply a greater role for structural (vs. regulatory) changes in host adaptation, as these SNPs could also be in LD with nearby regulatory elements. But it does bolster the evidence that these SNPs are tagging (via LD) some causal variants (i.e., that their status as top host-associated SNPs does not solely reflect a greater role for genetic drift). Similar results have been seen in several other genome scans for selection or adaptation [11, 33, 54]. Finally, we found that the genes nearest to the top host-associated SNPs have annotations suggesting a diversity of molecular functions and biological/cellular processes (Table 3.5). None of these stands out in a clear way, but this does suggest that host adaptation is likely a multifaceted process with selection shaping many different genes and molecular or developmental pathways.

REFERENCES

[1] Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR, Tabin CJ (2006) The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature*, **442**, 563–567.

[2] Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ (2004) Bmp4 and morphological variation of beaks in Darwin's finches. *Science*, **305**, 1462–1465.

[3] Agrawal AA (2017) Toward a predictive framework for convergent evolution: Integrating natural history, genetic mechanisms, and consequences for the diversity of life. *The American Naturalist*, **190**, S000–S000.

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

[5] Barrett RD, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science*, **322**, 255–257.

[6] Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting fst: the impact of rare variants. *Genome research*, **23**, 1514–1521.

[7] Bollback JP (2006) Simmap: stochastic character mapping of discrete traits on phylogenies. *BMC bioinformatics*, **7**, 88.

[8] Brachi B, Faure N, Horton M, *et al.* (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genetics*, **6**, e1000940.

[9] Colosimo PF, Hosemann KE, Balabhadra S, *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.

[10] Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. In: *Proceedings of Royal Society B*, vol. 279, pp. 5039–5047, The Royal Society.

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

[12] Dobler S, Mardulyn P, Pasteels JM, Rowell-Rahier M (1996) Host-plant switches and the evolution of chemical defense and life history in the leaf beetle genus *Oreina*. *Evolution*, **50**, 2373–2386.

[13] Egan SP, Ragland GJ, Assour L, *et al.* (2015) Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. *Ecology Letters*, **18**, 817–825.

[14] Erickson PA, Glazer AM, Killingbeck EE, *et al.* (2016) Partially repeatable genetic basis of benthic adaptation in threespine sticklebacks. *Evolution*, **70**, 887–902.

[15] Forister M, Scholl C, Jahner J, *et al.* (2013) Specificity, rank preference, and the colonization of a non-native host plant by the Melissa blue butterfly. *Oecologia*, **172**, 177–188.

[16] Forister ML, Gompert Z, Nice CC, Forister GW, Fordyce JA (2010) Ant association facilitates the evolution of diet breadth in a Lycaenid butterfly. *Proceedings of the Royal Society of London B: Biological Sciences*, p. rspb20101959.

[17] Forister ML, Nice CC, Fordyce JA, Gompert Z (2009) Host range evolution is not driven by the optimization of larval performance: the case of *Lycaeides melissa* (Lepidoptera: Lycaenidae) and the colonization of alfalfa. *Oecologia*, **160**, 551–561.

[18] Fountain T, Nieminen M, Sirén J, Wong SC, Lehtonen R, Hanski I (2016) Predictable allele frequency changes due to habitat fragmentation in the Glanville Fritillary butterfly. *Proceedings of the National Academy of Sciences*, **113**, 2678–2683.

[19] Fry JD (1996) The evolution of host specialization: are trade-offs overrated? *The American Naturalist*, **148**, S84–S107.

[20] Futuyma DJ, Moreno G (1988) The evolution of ecological specialization. *Annual Review of Ecology and Systematics*, **19**, 207–233.

[21] Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555–1579.

[22] Gompert Z, Comeault AA, Farkas TE, *et al.* (2014) Experimental evidence for ecological selection on genome variation in the wild. *Ecology letters*, **17**, 369–379.

[23] Gompert Z, Jahner JP, Scholl CF, *et al.* (2015) The evolution of novel host use is unlikely to be constrained by trade-offs or a lack of genetic variation. *Molecular Ecology*, **24**, 2777–2793.

[24] Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, Nice CC (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.

[25] Gompert Z, Messina FJ (2016) Genomic evidence that resource-based trade-offs limit host-range expansion in a seed beetle. *Evolution*, **70**, 1249–1264.

[26] Grosman AH, Molina-Rugama AJ, Mendes-Dias R, *et al.* (2015) No adaptation of a herbivore to a novel host but loss of adaptation to its native host. *Scientific Reports*, **5**.

[27] Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

[28] Harrison JG, Gompert Z, Fordyce JA, *et al.* (2016) The many dimensions of diet breadth: Phytochemical, genetic, behavioral, and physiological perspectives on the interaction between a native herbivore and an exotic host. *PloS One*, **11**, e0147971.

[29] Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

[30] Hwang WS, Weirauch C (2012) Evolutionary history of assassin bugs (Insecta: Hemiptera: Reduviidae): insights from divergence dating and ancestral state reconstruction. *PLoS ONE*, **7**, e45523.

[31] Janz N (1998) Sex–linked inheritance of host–plant specialization in a polyphagous butterfly. *Proceedings of the Royal Society of London B: Biological Sciences*, **265**, 1675–1678.

[32] Janz N (2003) Sex-linkage of host plant use in butterflies. *Butterflies: Ecology and evolution taking flight*, pp. 229–239.

[33] Jones FC, Grabherr MG, Chan YF, *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

[34] Lässig M, Mustonen V, Walczak AM (2017) Predicting evolution. *Nature Ecology & Evolution*, **1**, 0077.

[35] Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

[36] Li H, Handsaker B, Wysoker A, *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

[37] Losos JB (2011) Convergence, adaptation, and constraint. *Evolution*, **65**, 1827–1840.

[38] Marques DA, Lucek K, Meier JI, *et al.* (2016) Genomics of rapid incipient speciation in sympatric threespine stickleback. *PLoS Genetics*, **12**, e1005887.

[39] Martin A, Orgogozo V (2013) The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution*, **67**, 1235–1250.

[40] Matsubayashi KW, Ohshima I, Nosil P (2010) Ecological speciation in phytophagous insects. *Entomologia Experimentalis et Applicata*, **134**, 1–27.

[41] Michaud R, Lehman W, Rumbaugh M (1988) World distribution and historical development. *Alfalfa and alfalfa improvement*, pp. 25–91.

[42] Morris SC (2008) *The deep structure of biology: is convergence sufficiently ubiquitous to give a directional signal*. 45, Templeton Foundation Press.

[43] Nachman MW, Hoekstra HE, D'Agostino SL (2003) The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences*, **100**, 5268–5273.

[44] Nygren G, Nylin S, Stefanescu C (2006) Genetics of host plant use and life history in the comma butterfly across europe: varying modes of inheritance as a potential reproductive barrier. *Journal of Evolutionary Biology*, **19**, 1882–1893.

[45] Ord TJ, Summers TC (2015) Repeated evolution and the impact of evolutionary history on adaptation. *BMC Evolutionary Biology*, **15**, 137.

[46] Orgogozo V (2015) Replaying the tape of life in the twenty-first century. *Interface focus*, **5**, 20150057.

[47] Papa R, Morrison CM, Walters JR, *et al.* (2008) Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in *Heliconius* butterflies. *BMC Genomics*, **9**, 345.

[48] Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.

[49] Prowell DP (1998) Sex linkage and speciation in Lepidoptera. *Endless forms: species and speciation*, pp. 309–319.

[50] Revell LJ (2012) phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.

[51] Rogers S, Bernatchez L (2007) The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonus sp*. Salmonidae) species pairs. *Molecular Biology and Evolution*, **24**, 1423–1438.

[52] Scholl CF, Nice CC, Fordyce JA, Gompert Z, Forister ML (2012) Larval performance in the context of ecological diversification and speciation in Lycaeides butterflies. *International Journal of Ecology*, **2012**.

[53] Scotland RW (2011) What is parallelism? *Evolution & Development*, **13**, 214–227.

[54] Soria-Carrasco V, Gompert Z, Comeault AA, *et al.* (2014) Stick insect genomes reveal natural selectionâĂŹs role in parallel speciation. *Science*, **344**, 738–742.

[55] Speed MP, Arbuckle K (2017) Quantification provides a conceptual basis for convergent evolution. *Biological Reviews*, **92**, 815–829.

[56] Sperling FA (1994) Sex-linked genes and species differences in Lepidoptera. *The Canadian Entomologist*, **126**, 807–818.

[57] Thompson JN (1988) Evolutionary genetics of oviposition preference in Swallowtail butterflies. *Evolution*, pp. 1223–1234.

[58] Vertacnik KL, Linnen CR (2017) Evolutionary genetics of host shifts in herbivorous insects: insights from the age of genomics. *Annals of the New York Academy of Sciences*, **1389**, 186–212.

[59] Via S, Hawthorne DJ (2002) The genetic architecture of ecological specialization: correlated gene effects on host use and habitat choice in pea aphids. *The American Naturalist*, **159**, S76–S88.

[60] Weinig C, Ungerer MC, Dorn LA, *et al.* (2002) Novel loci control variation in reproductive timing in *Arabidopsis thaliana* in natural environments. *Genetics*, **162**, 1875–1884.

[61] Yassin A, Debat V, Bastide H, Gidaszewski N, David JR, Pool JE (2016) Recurrent specialization on a toxic fruit in an island *Drosophila* population. *Proceedings of the National Academy of Sciences*, **113**, 4771–4776.

[62] Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, **9**, e1003264.

Tables and Figures

Table 3.1. Locality information and sample sizes for the populations included in this study. Group denotes the lineage based on TREEMIX results, # Ind. gives the number of individuals sequenced for this study, and Data = indicates whether the sequence data were included in Gompert *et al.* [24] = "2014", or are being presented here for the first time = "Present".

	Locality	**Host-plant**	**Long. (W)**	**Lat. (N)**	**Group**	**# Ind.**	**Data**
1	Bishop, CA (BHP)	*Glycyrrhiza sp.*	118.28	37.17	*melissa*-west	20	2014
2	Crystal Creek Park, NV (VCP)	*Medicago sativa*	119.99	39.51	*melissa*-west	20	Present
3	Gardnerville, NV (GVL)	*Medicago sativa*	119.78	38.81	*melissa*-west	18	2014
4	Red Earth Way, NV (REW)	*Medicago sativa*	118.84	38.98	*melissa*-west	20	2014
5	Silver Lake, NV (SLA)	*Astragalus canadensis*	119.93	39.66	*melissa*-west	18	2014
6	Sierra Valley, CA (SVY)	*Medicago sativa*	121.14	39.09	*melissa*-west	20	2014
7	Trout Pond Trailhead, CA (TPT)	*Lupinus sp.*	116.58	32.97	*melissa*-west	13	Present
8	Washoe Lake, NV (WLA)	*Astragalus candensis*	118.82	38.65	*melissa*-west	20	2014
9	Abel Creek, NV (ABC)	*Lupinus sp.*	117.65	41.44	*melissa*-east	19	Present
10	Brandon, SD (BSD)	*Medicago sativa*	96.54	43.63	*melissa*-east	20	Present
11	Cody, WY (CDY)	*Medicago sativa*	108.98	44.51	*melissa*-east	23	2014
12	Cokeville, WY (CKV)	*Medicago sativa*	110.90	42.01	*melissa*-east	10	2014
13	De Beque, CO (DBQ)	*Medicago sativa*	108.21	39.32	*melissa*-east	20	2014
14	Deeth-Charleston, NV (DCR)	*Lupinus sp.*	115.38	41.30	*melissa*-east	20	2014
15	Goose Lake, CA (GLA)	*Medicago sativa*	120.29	41.30	*melissa*-east	20	2014
16	Lander, WY (LAN)	*Medicago sativa*	108.36	42.65	*melissa*-east	24	2014
17	Lamoille Canyon, NV (LCN)	*Lupinus sp.*	115.47	40.68	*melissa*-east	20	2014
18	Montrose, CO (MON)	*Medicago sativa*	107.82	38.37	*melissa*-east	20	2014
19	Montague, CA (MTU)	*Medicago sativa*	122.53	41.73	*melissa*-east	19	2014
20	Ophir City, NV (OPC)	*Lupinus sp.*	117.24	38.94	*melissa*-east	19	2014
21	Star Creek Canyon, NV (SCC)	*Lupinus sp.*	118.12	40.55	*melissa*-east	16	2014
22	Surprise Valley, CA (SUV)	*Medicago sativa*	120.10	41.28	*melissa*-east	20	2014
23	Upper Alkali Lake, CA (UAL)	*Medicago sativa*	120.15	41.74	*melissa*-east	20	Present
24	Victor, ID (VIC)	*Medicago sativa*	111.11	43.66	*melissa*-east	20	2014
25	Yellow Pine, WY (YWP)	Unknown native legume	105.40	41.25	*melissa*-east	20	2014
26	Albion Meadows, UT (ABM)	*Lupinus sp.*	111.92	40.48	*melissa*-east	46	2014

Fig. 3.1. Diagram shows a schematic representation of the primary analyses conducted in this study for main objectives. Each box presents a question asked in this study and the analyses conducted to answer these questions.

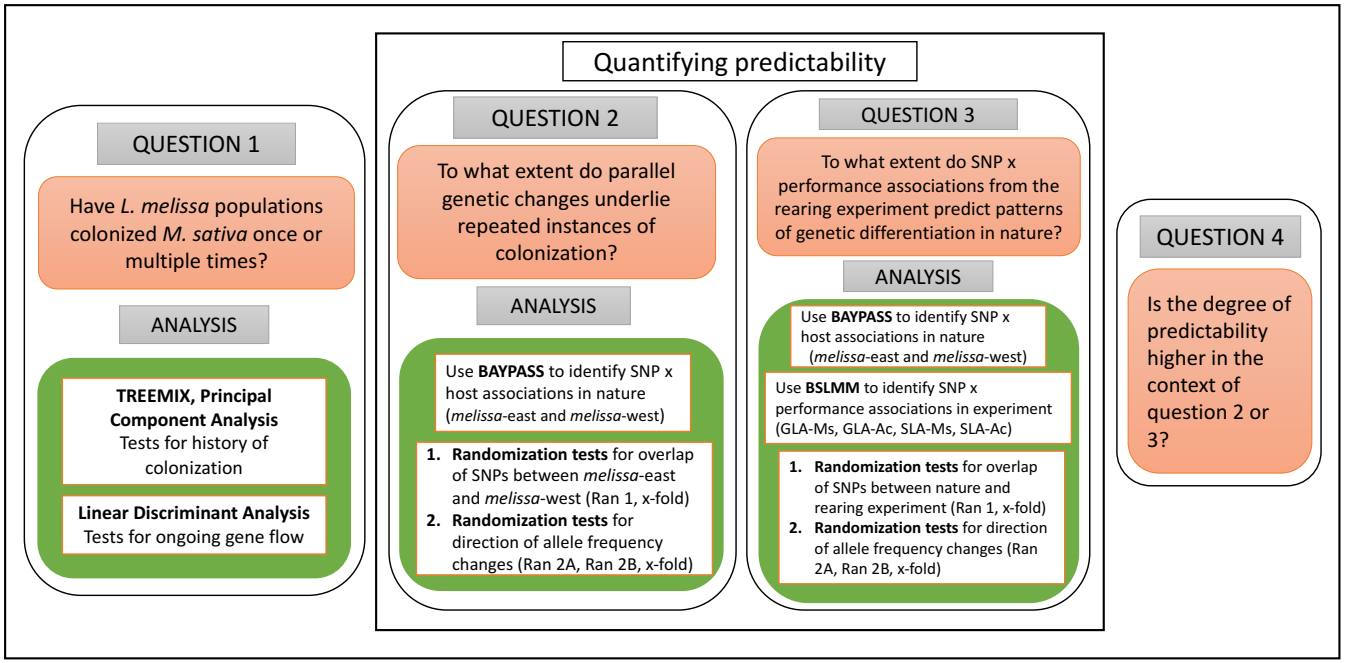


Fig. 3.2. (A) Map shows sample locations with populations colored based on host association. Population labels correspond to abbreviations for geographical locations in Table 3.1, and the line separates populations belonging to the eastern and western clades. (B) Plot shows summary of population structure based on principal component analysis. Abbreviations indicate populations corresponding to the map (A). The points denote individuals in each population used for the analysis. (C) Population graph from `TREEMIX` for *L. melissa* populations used in this study (N=26), allowing one migration or admixture event (the actual migration edge from the outgroup to ABM is not shown). Terminal nodes are labeled by abbreviations for geographical locations from where samples were collected and colored according to host-plant association.

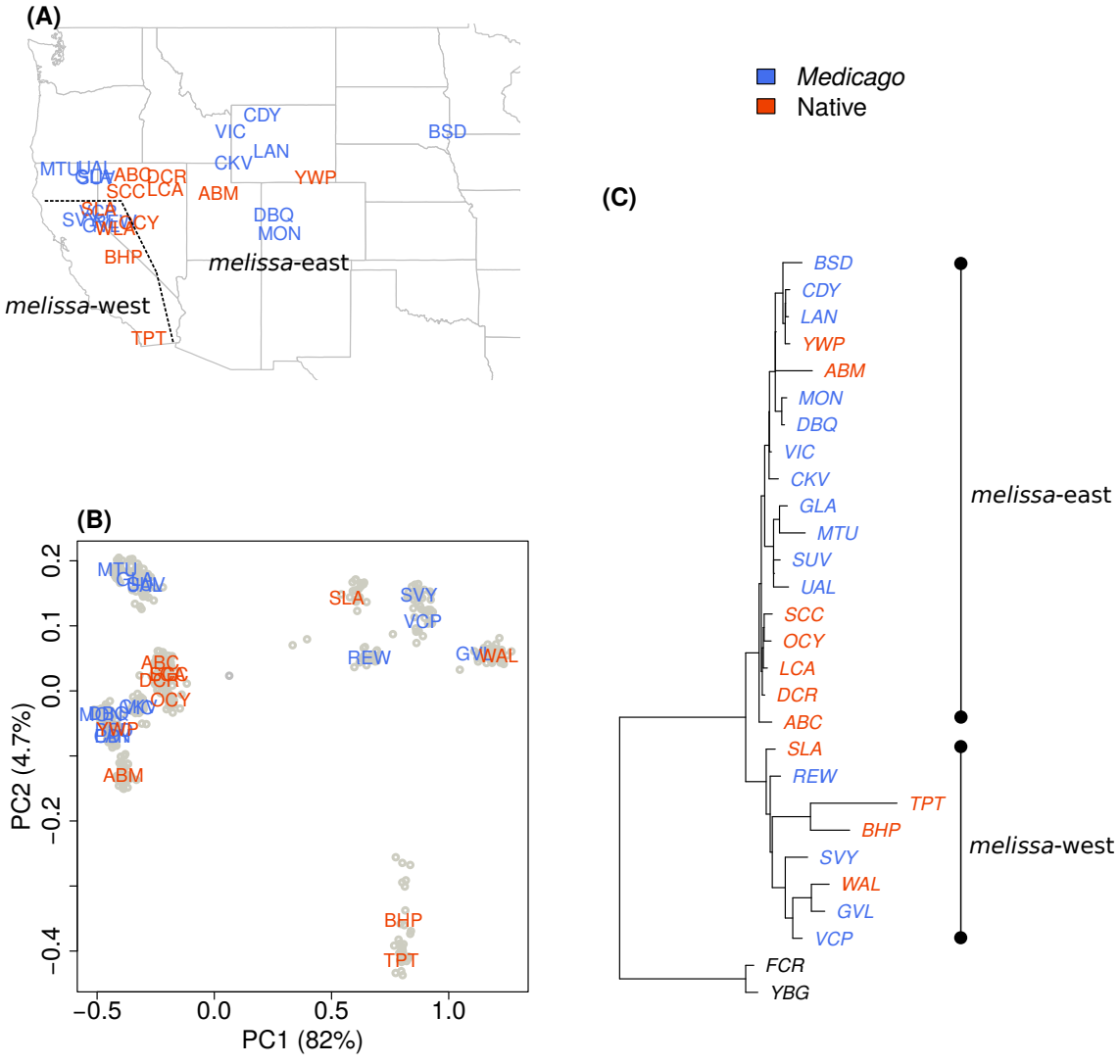


Fig. 3.3. Manhattan plot shows SNPs from (a) *melissa*-east (N=17) and (b) *melissa*-west (N=8) population groups, along linkage groups. The horizontal dashed line delineates the top 0.01% SNPs with the highest Bayes factors. Red points denote the 58 SNPs shared by the two groups. NA indicates SNPs which did not map on any linkage group.

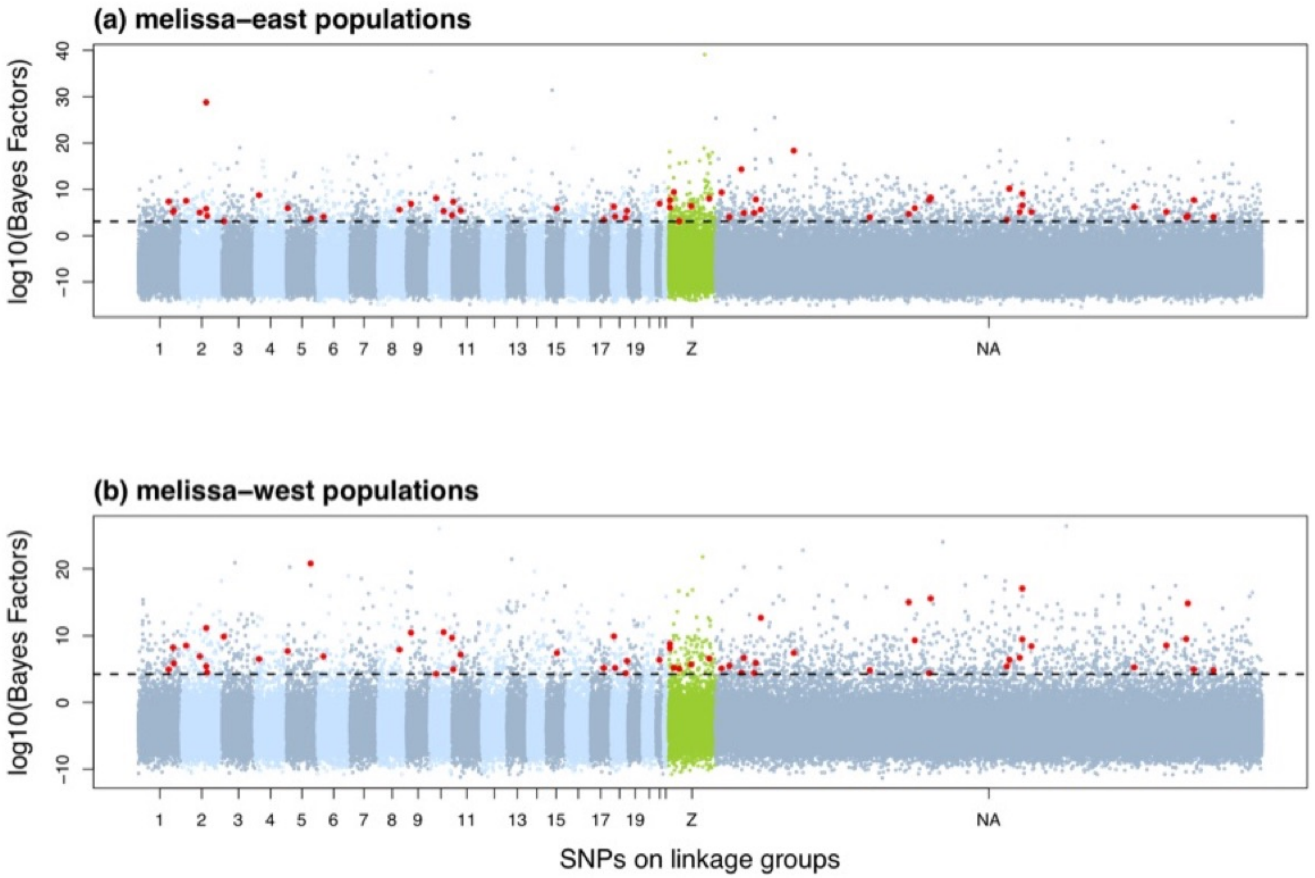


Fig. 3.4. Barplot shows x-fold enrichments for shared SNPs between *melissa*-east and *melissa*-west populations. Results are shown for different quantile cut-offs for defining the top host-associated SNPs. The null expectation is shown with a solid horizontal line.

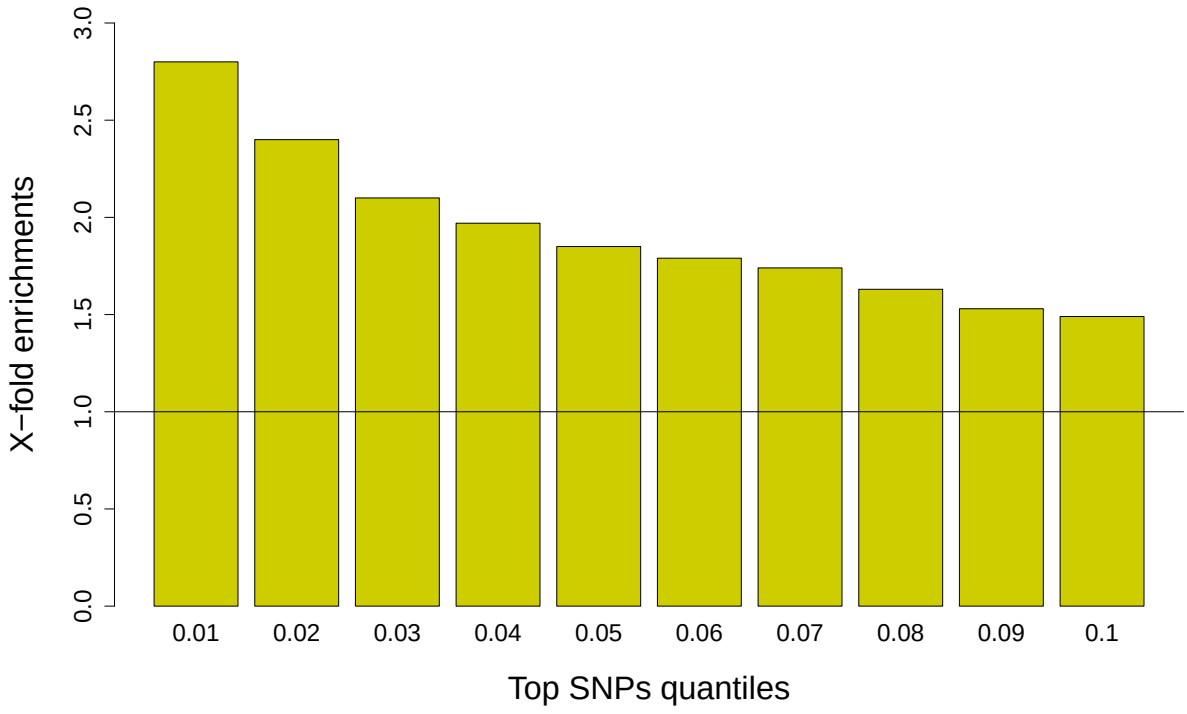


Fig. 3.5. Barplots show observed number of overlapping SNPs between performance-associated SNPs in the rearing experiment and host-associated SNPs (x-axis) in nature for the top 0.01% empirical quantile. In the figure legend, GLA-Medicago indicates larvae from GLA reared on *M. sativa*, GLA-Astragalus indicates larvae from GLA reared on *A. canadensis*, SLA-Medicago indicates larvae from SLA reared on *M. sativa*, and SLA-Astragalus indicates larvae from SLA reared on *A. canadensis*. * indicates x-fold enrichments with $P \leq 0.05$.

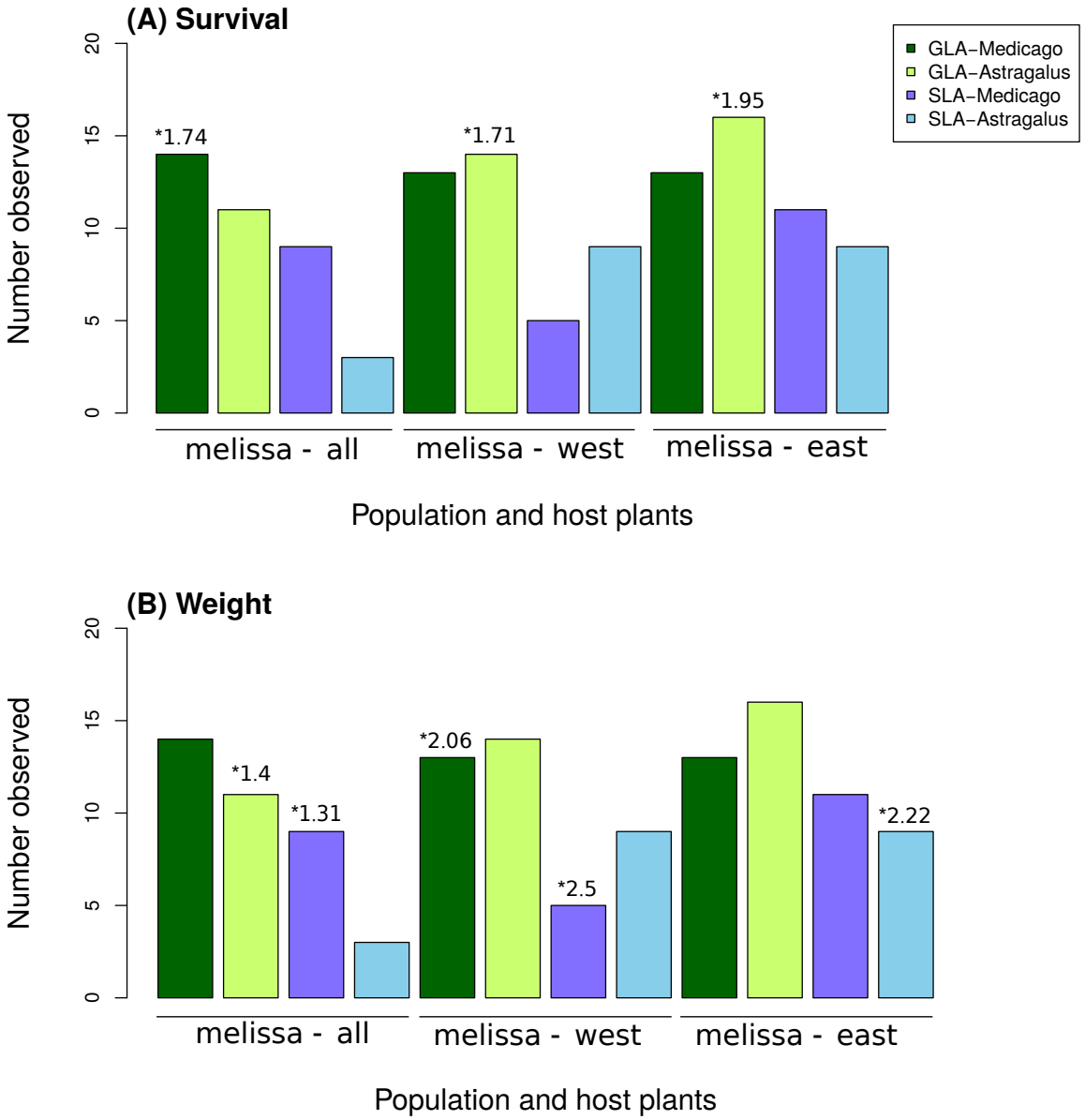
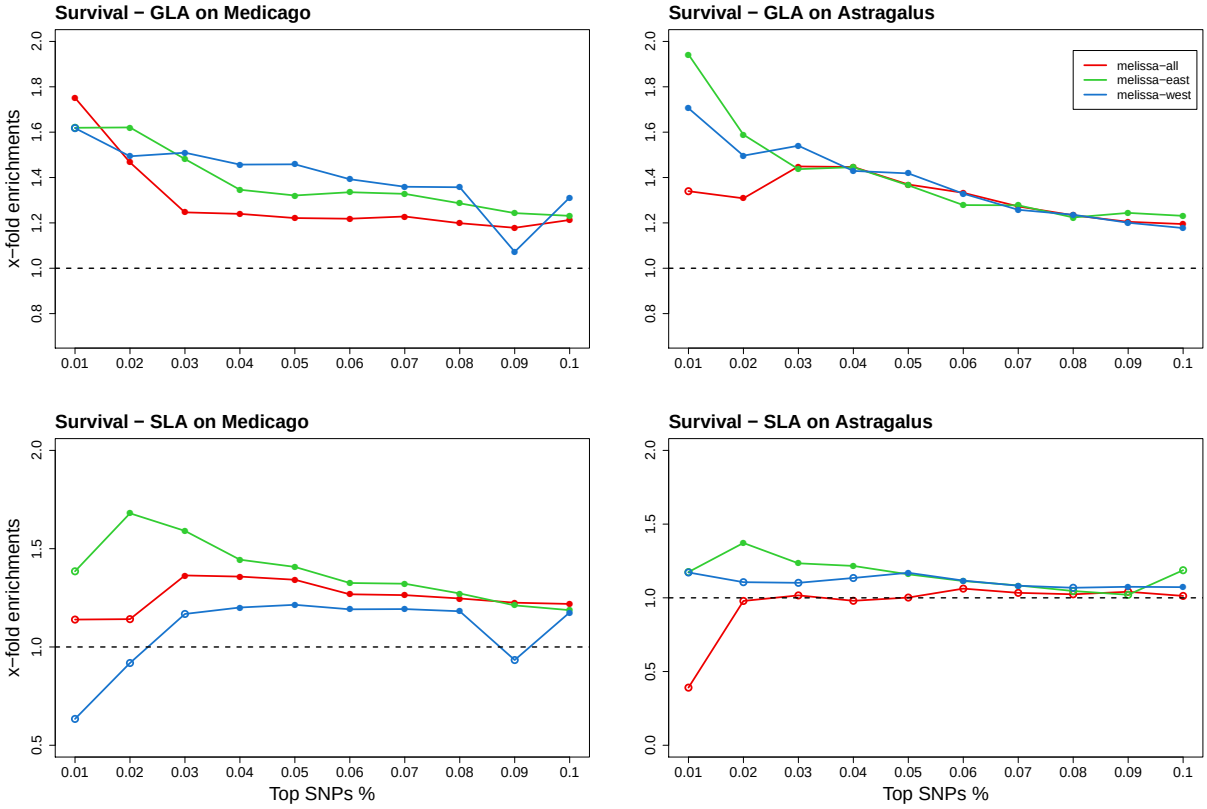


Fig. 3.6. Line plots show x-fold enrichments across quantiles for overlapping SNPs between survival-associated SNPs in the rearing experiment and host-associated SNPs in nature. Open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.



Supplemental tables and figures

Table 3.2. S1 Table shows summary of randomization tests for top 0.01% host-associated SNPs and top 0.01% parallel host-associated SNPs for presence on Z-chromosome (No. observed = number of SNPs observed on the sex chromosome; x-fold = number of observed is how much more than chance; number of SNPs observed on Z-chromosome and tests for randomizations; *P* = randomization-based *P*-values for the null hypothesis that the proportion of top SNPs observed on the Z-chromosome is not greater than the genomic proportion).

Set	**No. observed**	**P**	***x-fold***
All populations			
top 0.01% host-associated	195	<**0.01**	2.26
top 0.01% parallelism	6	0.03	2.48
melissa-east			
top 0.01% host-associated	193	<**0.01**	2.23
top 0.01% parallelism	6	0.03	2.48
melissa-west			
top 0.01% host-associated	134	<**0.01**	1.55
top 0.01% parallelism	6	0.04	2.48

Table 3.3. S2 Table shows summary of randomization tests for presence of top (0.01%) host-use associated SNPs on gene region of the genome (Top SNP% = Quantiles cut off for analysis; x-fold = Number of SNPs observed is how much more than chance; P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion; Mean = mean for the null hypothesis). $P \leq 0.05$ are in bold.

Top SNP %	**No. observed**	***P***	**x-fold enrichment**
All populations			
top 0.001%	0.13	**0.01**	1.30
top 0.01%	0.42	0.82	0.96
top 0.1%	0.34	1.00	0.51
melissa-east			
top 0.001%	0.15	< **0.01**	1.30
top 0.01%	0.42	0.86	0.96
top 0.1%	0.31	0.26	1.01
melissa-west			
top 0.001%	0.14	< **0.01**	1.30
top 0.01%	0.47	0.91	0.96
top 0.1%	0.32	1.00	0.51
top 0.01% parallelism			
top 0.001%	0	0.99	0.5
top 0.01%	0	1.00	0

Table 3.4. S3 Table shows summary of randomization tests for presence of top (0.01%) host-use associated SNPs on coding region of the genome (To SNP% = quantiles cut off for analysis; x-fold = Number of SNPs observed is how much more than chance; *P* = randomization-based *P*-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion; Mean = mean for the null hypothesis). *P* significant at 0.05 are in bold.

Set	***No. observed***	**P**	**x-fold**
All populations			
top 0.001%	1.00	0.12	0.54
top 0.01%	< **0.01**	0.43	1.46
top 0.1%	< **0.01**	0.31	1.07
melissa-east			
top 0.001%	1.00	0.14	0.54
top 0.01%	< **0.01**	0.42	1.46
top 0.1%	< **0.01**	0.30	1.07
melissa-west			
top 0.001%	1.00	0.14	0.54
top 0.01%	< **0.01**	0.42	1.46
top 0.1%	< **0.01**	0.34	1.07
top 0.01% parallelism			
top 0.001%	0.99	< **0.01**	0.51
top 0.01%	1.00	< **0.01**	0.00

Table 3.5. S4 Table shows summary of randomization tests for determining molecular functions of top (0.01%) host-use associated SNPs (GO ID = Gene ontology reference ID; GO name = Name of the function associated with GO ID, No. = number of top 0.01% SNPs enriched for the GO function, P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion, x-fold = number of SNPs observed is how much more than chance. $P \leq 0.05$ are in bold.

Molecular functions				
GO ID	**GO name**	**No.**	***P***	**x-fold**
GO:0004519	endonuclease activity	8	< **0.01**	4.92
GO:0004177	aminopeptidase activity	7	**0.01**	2.41
GO:0008234	Cysteine-type peptidase activity	7	**0.04**	2.25
GO:0003684	damaged DNA binding	7	< **0.01**	6.23
GO:0005452	inorganic anion exchanger activity	6	**0.04**	2.16
GO:0004047	aminomethyltransferase activity	6	< **0.01**	2.92
GO:0005272	sodium channel activity	5	**0.01**	2.68
GO:0008889	glycerophosphodiester phosphodiesterase activity	5	**0.05**	2.12
GO:0005544	Calcium-dependent phospholipid binding	4	**0.04**	2.33
GO:0004768	Stearoyl-CoA 9-desaturase activity	4	0.06	2.25
GO:0004044	amidophosphoribosyltransferase activity	4	**0.01**	2.79
GO:0016709	oxidoreductase activity*	4	**0.01**	3
GO:0005247	Voltage-gated chloride channel activity	3	0.08	2.32
GO:0033897	ribonuclease T2 activity	3	**0.02**	3.61
GO:0004452	Isopentenyl-diphosphate delta-isomerase activity	3	0.12	2.56
GO:0030429	kynureninase activity	3	0.13	2.54
GO:0016844	strictosidine synthase activity	3	< **0.01**	4.23
GO:0016538	Cyclin-dependent protein serine/threonine kinase regulator activity	3	0.14	2.17
GO:0003867	4-aminobutyrate transaminase activity	3	**0.02**	3.32
GO:0004594	pantothenate kinase activity	2	**0.04**	3.34
GO:0017172	cysteine dioxygenase activity	2	**0.01**	4.91
GO:0005201	extracellular matrix structural constituent	2	**0.04**	2.84
GO:0008565	protein transporter activity	2	0.12	2.48
GO:0004066	asparagine synthase (glutamine-hydrolyzing) activity	2	**0.02**	3.33
GO:0003950	NAD+ ADP-ribosyltransferase activity	2	**0.02**	4.14

Table 3.6. S5 Table shows summary of randomization tests for determining biological functions of top (0.01%) host-use associated SNPs (GO ID = Gene ontology reference ID; GO name = Name of the function associated with GO ID, No. = Number of top 0.01% SNPs enriched for the GO function, P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion, x-fold = Number of SNPs observed is how much more than chance. $P \leq 0.05$ are in bold.

Biological functions				
GO ID	**GO name**	**No.**	***P***	**x-fold**
GO:0006741	NADP biosynthetic process	9	**0.01**	2.13
GO:0019674	NAD metabolic process	9	**0.01**	2.15
GO:0015991	ATP hydrolysis coupled proton transport	9	**0.02**	2.14
GO:0006546	glycine catabolic process	6	**0.01**	2.68
GO:0006820	anion transport	6	**0.04**	2.12
GO:0006071	glycerol metabolic process	5	**0.04**	2.31
GO:0019752	carboxylic acid metabolic process	5	**0.02**	2.54
GO:0015937	coenzymeA biosynthetic process	5	0.06	2.77
GO:0006542	glutamine biosynthetic process	5	**0.05**	2.32
GO:0042318	penicillin biosynthetic process	4	0.12	2.23
GO:0006506	GPI anchor biosynthetic process	4	0.14	2.21
GO:0009116	nucleoside metabolic process	4	**0.03**	2.81
GO:0009113	purine nucleobase biosynthetic process	4	**0.02**	2.83
GO:0006633	fatty acid biosynthetic process	4	0.09	2.32
GO:0006744	ubiquinone biosynthetic process	4	**0.01**	2.91
GO:0006821	chloride transport	3	0.07	2.23
GO:0009448	gamma-aminobutyric acid metabolic process	3	**0.02**	3.23
GO:0006397	mRNA processing	3	**0.03**	2.91
GO:0006569	tryptophan catabolic process	3	0.12	2.44
GO:0030071	regulation of mitotic metaphase/anaphase transition	2	**0.04**	3.21
GO:0006904	vesicle docking involved in exocytosis	2	**0.02**	3.92
GO:0046439	L-cysteine metabolic process	2	**0.01**	4.95
GO:0042254	ribosome biogenesis	2	0.13	2.1
GO:0000398	mRNA splicing via spliceosome	2	**0.03**	3.25
GO:0006529	asparagine biosynthetic process	2	**0.04**	3.22

Table 3.7. S6 Table shows summary of randomization tests for determining cellular functions of top (0.01%) host-use associated SNPs (GO ID = Gene ontology reference ID; GO name = Name of the function associated with GO ID, No. = Number of top 0.01% SNPs enriched for the GO function, P = randomization-based P-values for the null hypothesis that the proportion of top SNPs observed is not greater than the genomic proportion, x-fold = Number of SNPs observed is how much more than chance. $P \leq 0.05$ are in bold.

Cellular functions				
GO ID	**GO name**	**No.**	***P***	**x-fold**
GO:0005643	nuclear pore	10	< **0.01**	2.33
GO:0000139	Golgi membrane	6	**0.01**	2.72
GO:0005795	Golgi stack	4	< **0.001**	9.61
GO:0033180	proton-transporting V-type ATPase V1 domain	2	**0.01**	4.83
GO:0005581	collagen trimer	2	**0.05**	2.75
GO:0030126	COPI vesicle coat	2	**0.02**	3.24
GO:0005960	glycine cleavage complex	2	**0.01**	4.81

Table 3.8. S7 Results from randomization tests for overlap of the top (0.01%) host-use associated SNPs in nature and the top (0.01%) survival-associated SNPs in rearing experiment (based on *ran1*). Population-plant = population and plant treatment in the laboratory experiment; No. observed = number of SNPs associated with both host use in wild and performance in the lab; x-fold = enrichment relative to null expectations; P = randomization-based P-values for the null hypothesis ($P \leq 0.05$ are in bold). Results are shown based on raw Bayes factors and model-averaged effect sizes, and based on residuals controlling these metrics for allele frequencies.

	Raw values			**Residual values**		
Population-plant	**No. observed**	***P***	**x-fold**	**No. observed**	***P***	**x-fold**
All populations						
GLA-Ms	4	0.96	0.49	14	**0.03**	1.74
GLA-Ac	6	0.84	0.72	11	0.21	1.34
SLA-Ms	2	0.99	0.25	9	0.39	1.14
SLA-Ac	0	1.00	0.00	3	0.39	0.98
melissa-east						
GLA-Ms	6	0.81	0.74	13	0.06	1.62
GLA-Ac	7	0.71	0.85	16	**0.01**	1.95
SLA-Ms	2	0.99	0.25	11	0.17	1.39
SLA-Ac	4	0.95	0.52	9	0.36	1.17
melissa-west						
GLA-Ms	9	0.41	1.12	13	0.07	1.61
GLA-Ac	13	0.07	1.58	14	**0.04**	1.71
SLA-Ms	4	0.95	0.50	5	0.89	0.63
SLA-Ac	4	0.94	0.52	9	0.35	1.18

Table 3.9. S8 Results from randomization tests for overlap of the top (0.01%) host-use associated SNPs in nature and the top (0.01%) weight-associated SNPs in rearing experiment (based on *ran1*). Population-plant = population and plant treatment in the laboratory experiment; No. observed = number of SNPs associated with both host use in wild and performance in the lab; x-fold = enrichment relative to null expectations; P = randomization-based P-values for the null hypothesis ($P \leq 0.05$ are in bold). Results are shown based on raw Bayes factors and model-averaged effect sizes, and based on residuals controlling these metrics for allele frequencies.

	Raw values			**Residual values**		
Population-plant	**No. observed**	***P***	**x-fold**	**No. observed**	***P***	**x-fold**
All populations						
GLA-Ms	5	0.89	0.64	11	0.16	1.42
GLA-Ac	21	< **0.01**	2.67	11	0.17	1.39
SLA-Ms	22	< **0.01**	2.87	10	0.23	1.31
SLA-Ac	2	0.99	0.26	6	0.78	0.78
melissa-east						
GLA-Ms	7	0.66	0.90	8	0.52	1.03
GLA-Ac	5	0.89	0.64	10	0.25	1.28
SLA-Ms	5	0.95	0.52	4	0.95	0.53
SLA-Ac	8	0.49	1.05	17	< **0.01**	2.23
melissa-west						
GLA-Ms	16	**0.01**	2.08	16	**0.01**	2.06
GLA-Ac	12	0.09	1.53	13	**0.05**	1.66
SLA-Ms	10	0.23	1.32	19	< **0.01**	2.50
SLA-Ac	10	0.24	1.31	12	0.09	1.56

Table 3.10. S9 Table shows summary of randomization tests for concordance in effect signs of overlapping host-associated SNPs and survival-associated SNPs in the rearing experiment for the top 0.01% empirical quantile ($P \leq$ are in bold). Results are shown for randomization tests *ran2A* and *ran2B*.

		ran2A		***ran2B***	
Population-plant	**No. observed**	***P***	**x-fold**	***P***	**x-fold**
All populations					
GLA-Ms	8	0.16	1.07	0.23	1.11
GLA-Ac	4	0.39	0.87	0.69	0.75
SLA-Ms	5	1.00	1.00	0.27	1.08
SLA-Ac	1	0.36	0.58	0.48	0.67
melissa-east					
GLA-Ms	9	**0.05**	1.21	**0.04**	1.38
GLA-Ac	5	0.81	0.64	0.88	0.64
SLA-Ms	9	< **0.01**	1.39	**0.01**	1.59
SLA-Ac	4	1.00	1.00	0.49	0.89
melissa-west					
GLA-Ms	7	0.54	0.85	0.29	1.07
GLA-Ac	9	0.28	1.01	0.36	1.03
SLA-Ms	4	< **0.01**	1.42	**0.03**	1.53
SLA-Ac	5	0.12	1.15	0.26	1.11

Table 3.11. S10 Table shows summary of randomization tests for concordance in effect signs of overlapping host-associated SNPs and weight-associated SNPs in the rearing experiment for the top 0.01% empirical quantile (significant *P*-values at 0.05 are in bold). Results are shown for randomization tests *ran2A* and *ran2B*.

		ran2A		***ran2B***	
Population-plant	**No. observed**	***P***	**x-fold**	***P***	**x-fold**
All populations					
GLA-Ms	7	0.13	1.07	0.13	1.25
GLA-Ac	5	0.45	0.85	0.46	0.92
SLA-Ms	2	0.71	0.58	0.93	0.41
SLA-Ac	1	0.66	0.43	0.87	0.34
melissa-east					
GLA-Ms	4	0.17	1.07	0.36	1.02
GLA-Ac	2	0.94	0.39	0.93	0.41
SLA-Ms	1	0.48	0.51	0.69	0.49
SLA-Ac	2	0.99	0.25	0.99	0.24
melissa-west					
GLA-Ms	7	0.61	0.82	0.59	0.87
GLA-Ac	5	$<$ **0.01**	1.21	0.68	0.78
SLA-Ms	7	0.59	0.85	0.82	0.74
SLA-Ac	6	0.35	0.93	0.39	1.01

Table 3.12. S11 Table shows summary of randomization tests for overlapping high F_{ST} SNPs and performance-associated SNPs in the rearing experiment for the top 0.01% empirical quantile ($P \leq 0.05$ are in bold). Results are shown for randomization tests *ran1*

Population-plant	**No. observed**	***P***	**x-fold**
GLA-SLA Pairwise F_{ST}			
GLA-Ms-weight	20	<**0.01**	2.58
GLA-Ac-weight	14	**0.02**	1.81
GLA-Ms-survival	14	**0.02**	1.81
GLA-Ac-survival	29	<**0.01**	3.75
SLA-Ms-weight	21	< **0.01**	2.71
SLA-Ac-weight	35	< **0.01**	4.53
SLA-Ms-survival	13	**0.04**	1.68
SLA-Ac-survival	11	0.16	1.42
GLA-ABC Pairwise F_{ST}			
GLA-Ms-weight	12	0.09	1.55
GLA-Ac-weight	13	**0.05**	1.68
GLA-Ms-survival	15	**0.01**	1.93
GLA-Ac-survival	13	**0.04**	1.68
SLA-VCP Pairwise F_{ST}			
SLA-Ms-weight	13	**0.05**	1.67
SLA-Ac-weight	16	**0.01**	2.06
SLA-Ms-survival	6	0.79	0.77
SLA-Ac-survival	7	0.65	0.91

Table 3.13. S12 Table shows summary of randomization tests for concordance in effect signs of overlapping pairwise high F_{ST} SNPs and performance-associated SNPs in the rearing experiment for the top 0.01% empirical quantile ($P \leq 0.05$ are in bold). Results are shown for randomization tests *ran2A* and *ran2B*.

		ran2A		***ran2B***	
Population-plant	**No. observed**	***P***	**x-fold**	***P***	**x-fold**
GLA-SLA Pairwise F_{ST}					
GLA-Ms-weight	15	< **0.01**	1.37	**0.01**	1.55
GLA-Ac-weight	3	0.93	0.44	0.94	0.47
GLA-Ms-survival	9	0.27	0.94	0.09	1.28
GLA-Ac-survival	5	1.00	0.34	0.99	0.38
SLA-Ms-weight	11	0.33	1.00	0.32	1.07
SLA-Ac-weight	1	1.00	0.08	1.00	0.06
SLA-Ms-survival	11	0.09	1.16	0.12	1.21
SLA-Ac-survival	6	< **0.01**	1.17	0.25	1.13
GLA-ABC Pairwise F_{ST}					
GLA-Ms-weight	3	0.92	0.48	0.89	0.52
GLA-Ac-weight	3	0.89	0.54	0.94	0.49
GLA-Ms-survival	8	< **0.01**	1.34	0.29	1.07
GLA-Ac-survival	4	0.78	0.63	0.82	0.66
SLA-VCP Pairwise F_{ST}					
SLA-Ms-weight	4	0.78	0.63	0.81	0.66
SLA-Ac-weight	3	0.98	0.38	0.97	0.40
SLA-Ms-survival	6	0.15	1.09	0.29	1.08
SLA-Ac-survival	3	< **0.01**	1.22	0.44	0.92

Fig. 3.7. S1 Diagram shows a schematic representation of the analyses conducted to test for concordance between direction of allele frequency differences between alfalfa-feeding and native-feeding populations and signs for model average effects for performance-associated SNPs in rearing experiment. Each box represents an analysis conducted in the study. SAF = standardized allele frequencies for host-associated SNPs in natural populations, MAE = model average effects for performance-associated SNPs.

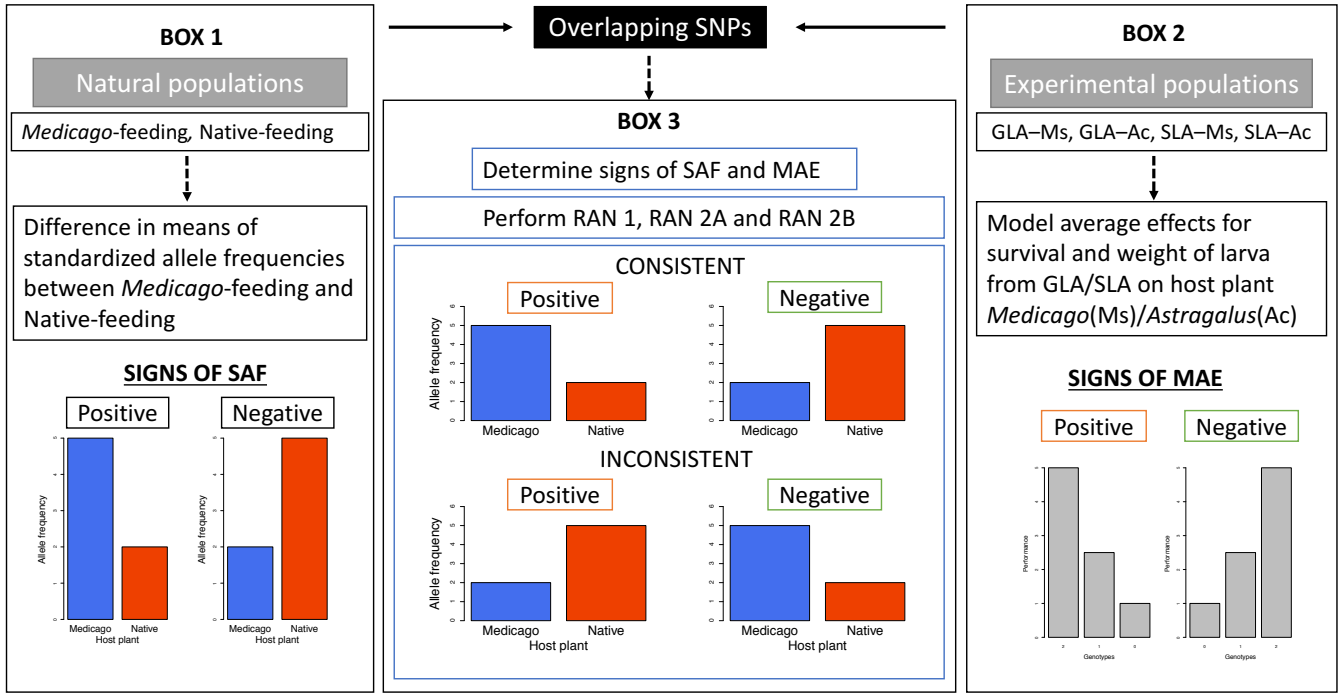


Fig. 3.8. S2 Plot shows proportion of variation explained by the TREEMIX population graph with different numbers of migration edges.

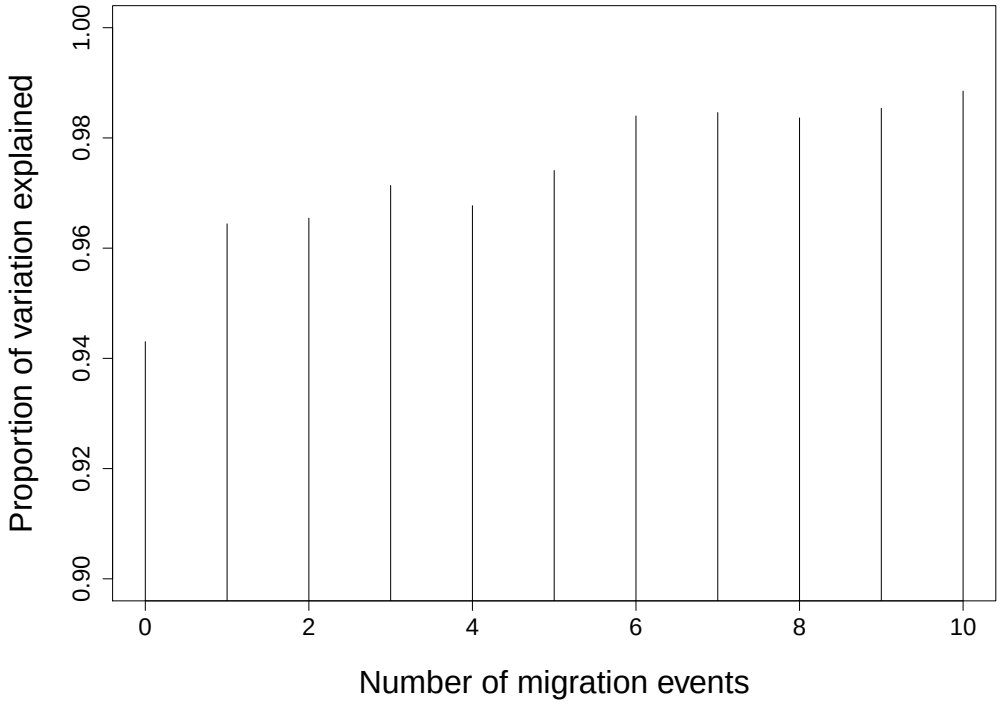


Fig. 3.9. S3 Tree shows ancestral state recontruction of the mutations that lead to colonization shifts from native host to novel host *Medicago sativa*. Terminal nodes are labeled by abbreviations for geographical locations from where samples were collected and circles beside the terminal locations are colored according to host-plant association. Inferred ancestral states are denoted by pie-charts that indicate the posterior probability of being associated with native host (orangered) versus being associated with *Medicago* (blue).

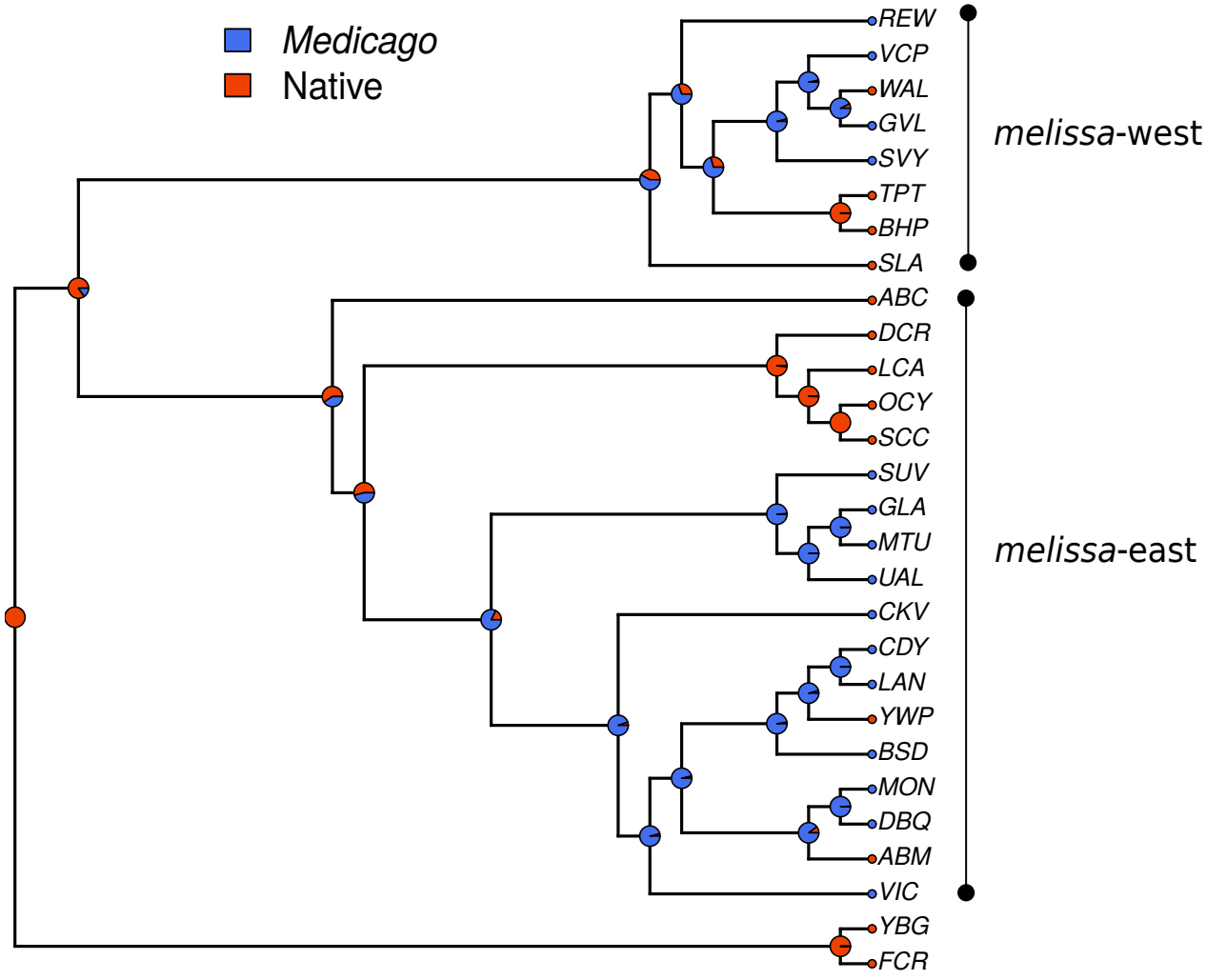


Fig. 3.10. S4 Plot shows the mean assignment probability to the correct population (i.e., the one that an individual was sampled from) across all 300 pairs as a function of log geographic distance and whether the pair of populations feed on the same or different host plants. Note that average assignments to the collected populations were very similar for same (0.964, sd = 0.0524) and different (0.984, sd = 0.953) host comparisons.

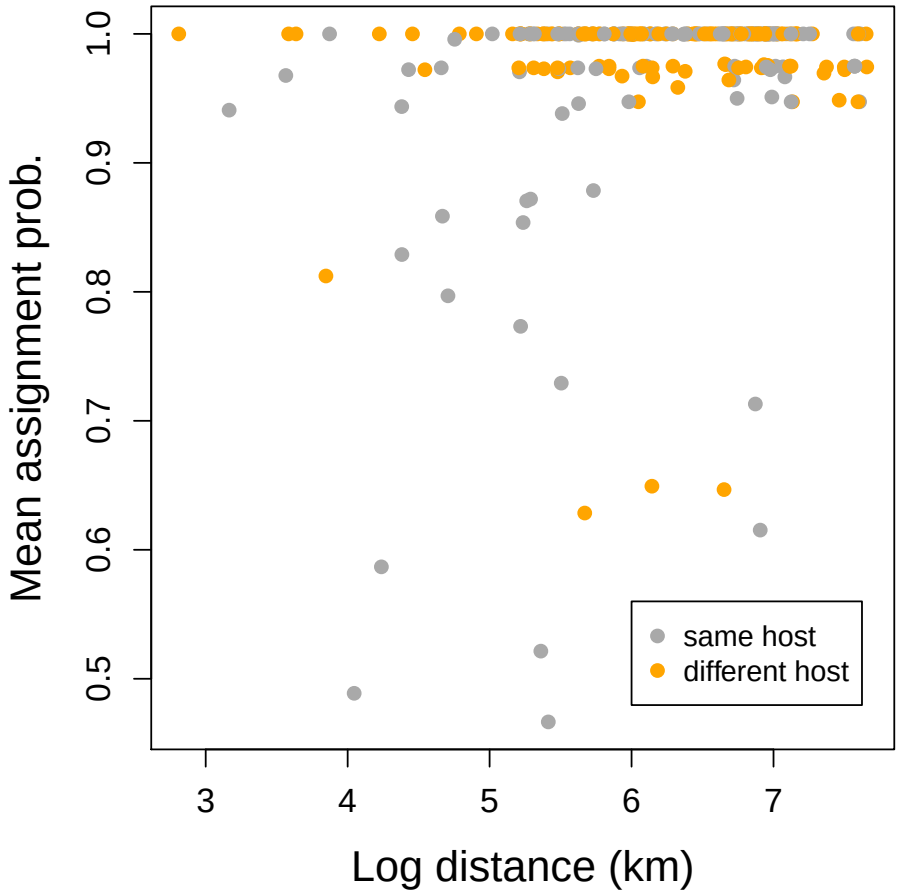


Fig. 3.11. S5 Barplots show individual assignment probabilities for four of the nearest population pairs that fed on different host plants. In panels (A) and (C) all individuals were confidently assign to the population they were collected from. (B) shows a case where that there is much more uncertainty in general (i.e., genetic differentiation between these populations is low), but two likely migrants. (D) shows a single individual that is most likely a migrant from SUV (or a similar population) to SLA.

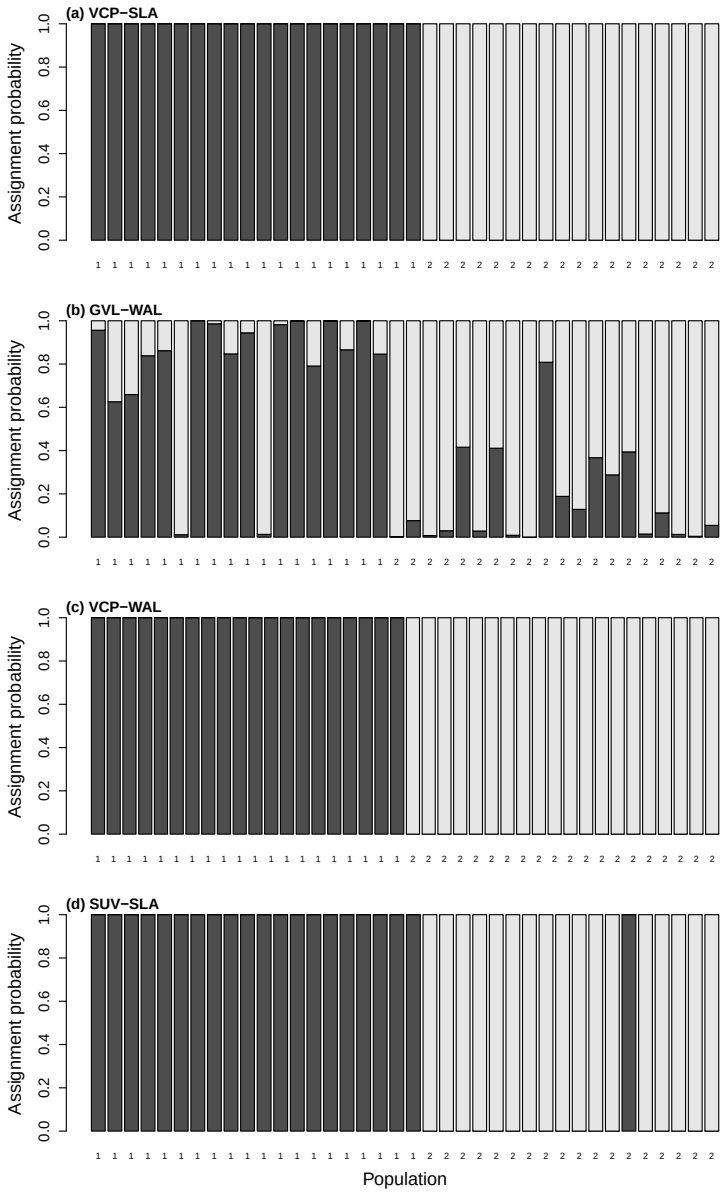


Fig. 3.12. S6 Manhattan plot for all populations (N=25) shows SNPs (N=206,028) as points mapped along linkage groups (1-Z). Z indicates the sex-chromosome. NA indicates SNPs which have not been assigned to a linkage group. Straight line separates the top 0.01% SNPs with high Bayes factor values.

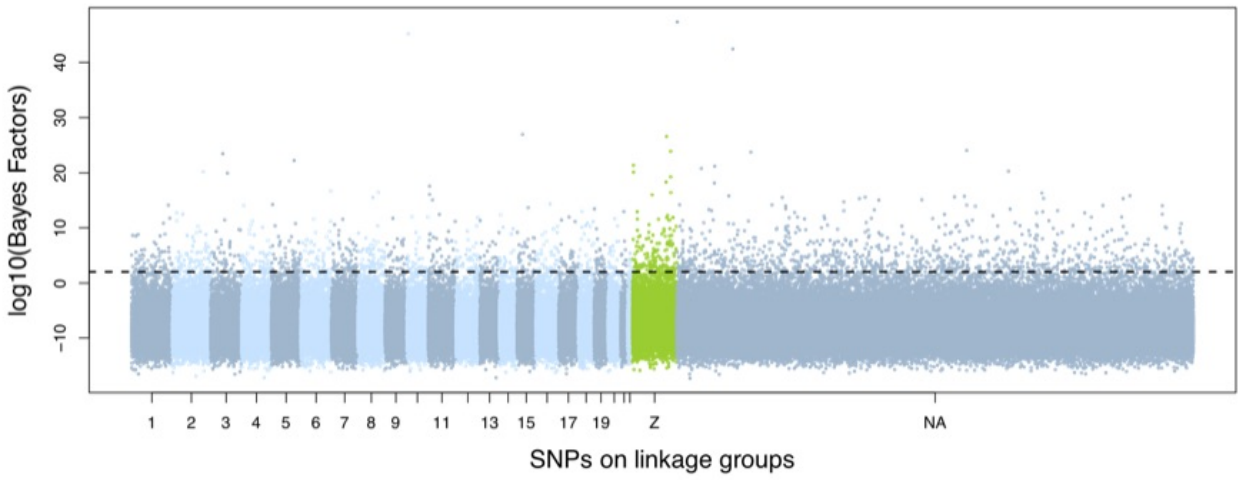


Fig. 3.13. S7 Line plots show x-fold enrichments across quantiles for overlapping SNPs between weight-associated SNPs in the rearing experiment and host-associated SNPs in nature. Open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

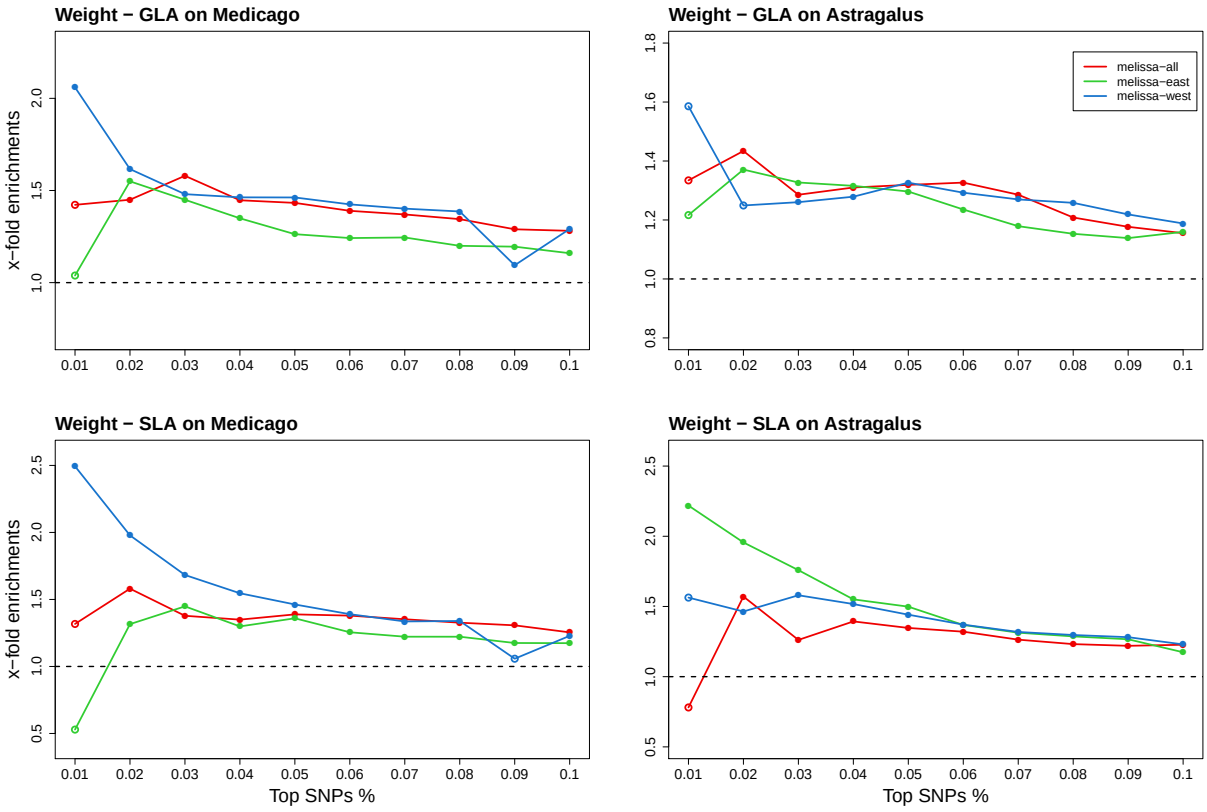


Fig. 3.14. S8 Line plot shows x-fold enrichments across quantiles for overlapping SNPs between survival-associated SNPs in rearing experiment and host-associated SNPs in nature for *ran2A*. Open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

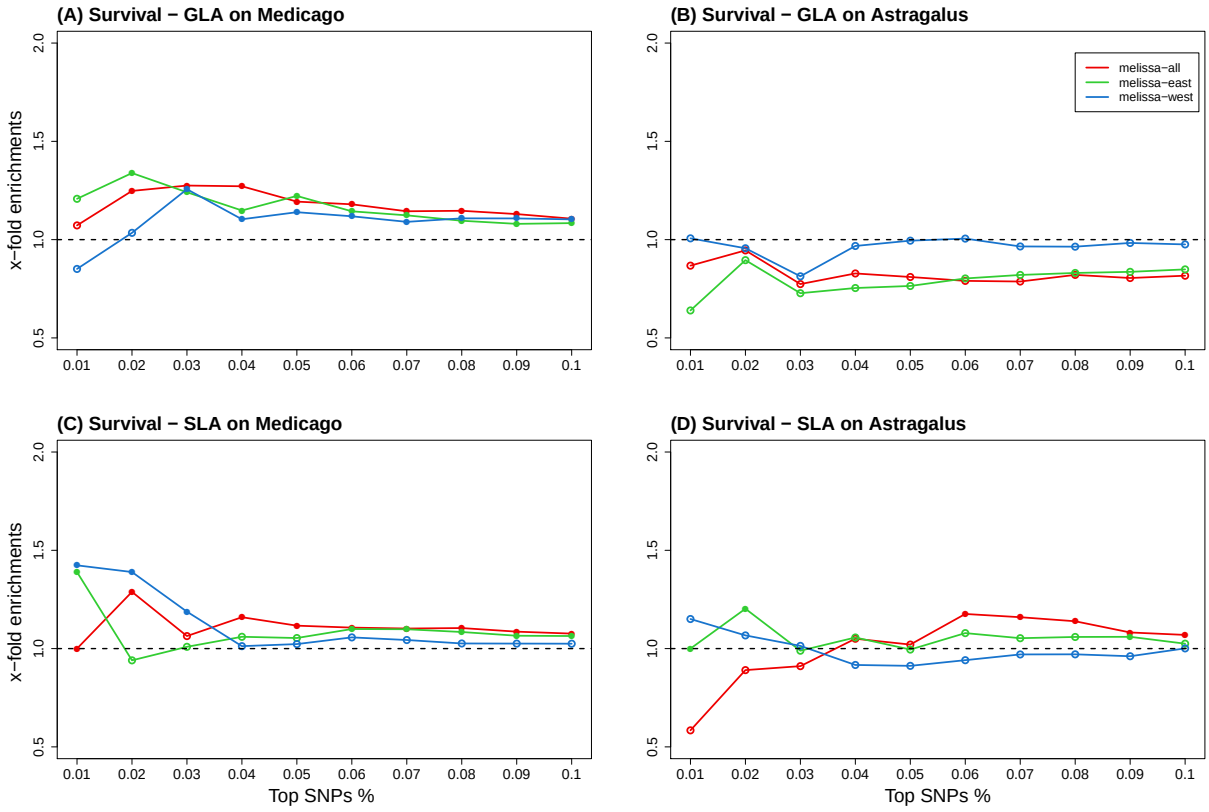


Fig. 3.15. S9 Line plots show x-fold enrichments across quantiles for overlapping SNPs between weight-associated SNPs in the rearing experiment and host-associated SNPs in nature for *ran2A*. Open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

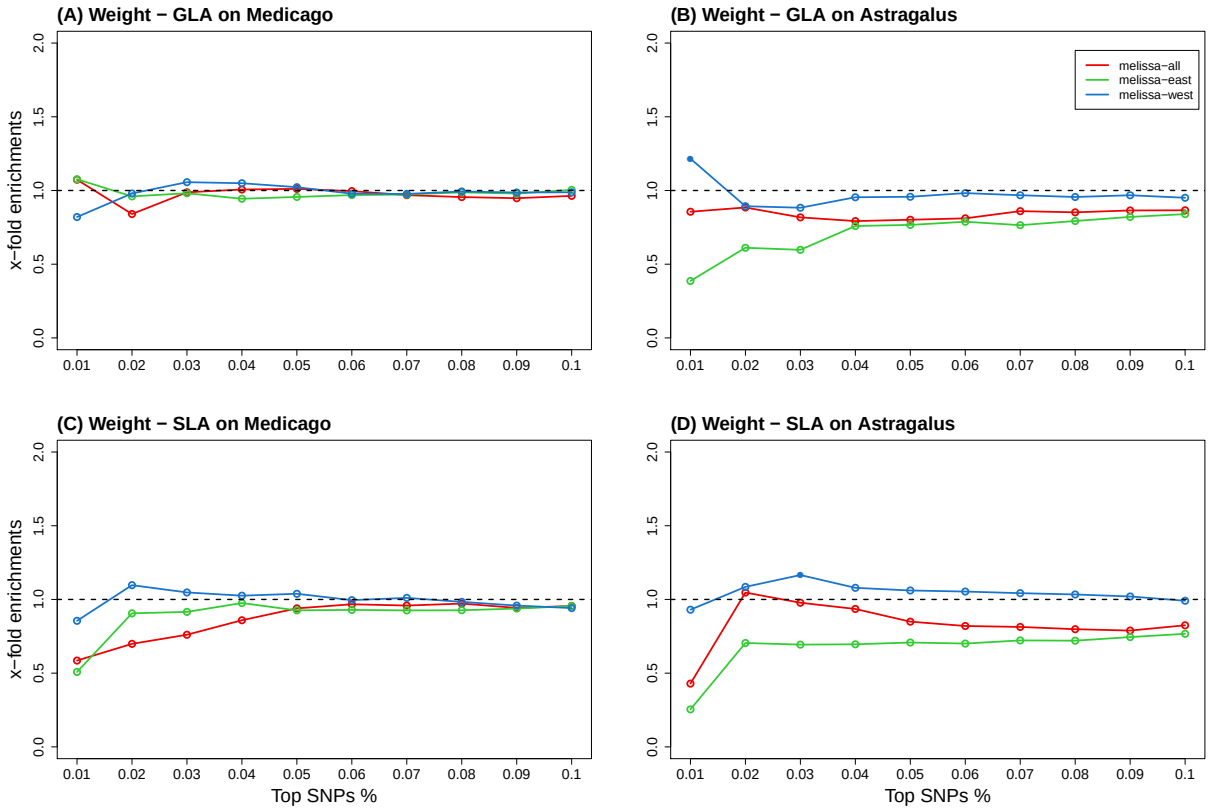


Fig. 3.16. S10 Line plot shows x-fold enrichments across quantiles for overlapping SNPs between survival-associated SNPs in rearing experiment and host-associated SNPs in nature for *ran2B*. Open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

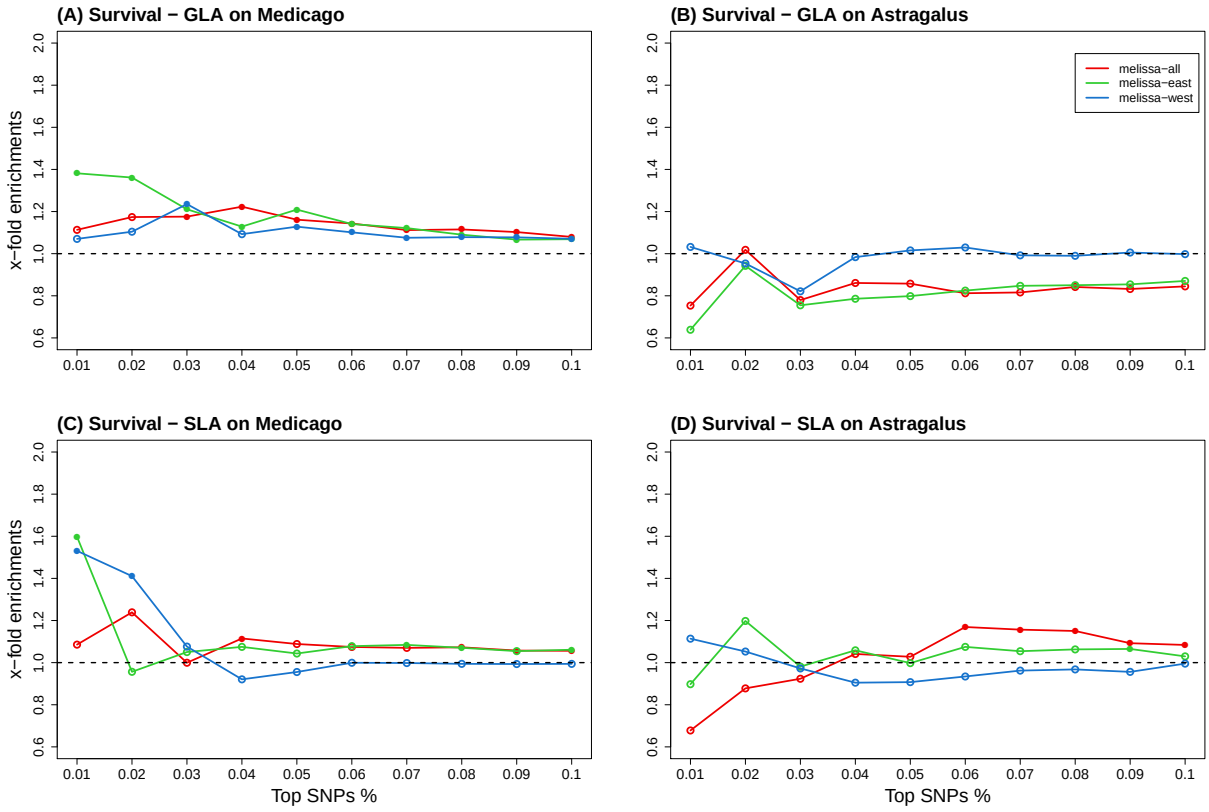


Fig. 3.17. S11 Line plots show x-fold enrichments across quantiles for overlapping SNPs between weight-associated SNPs in the rearing experiment and host-associated SNPs in nature for *ran2B*. Open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

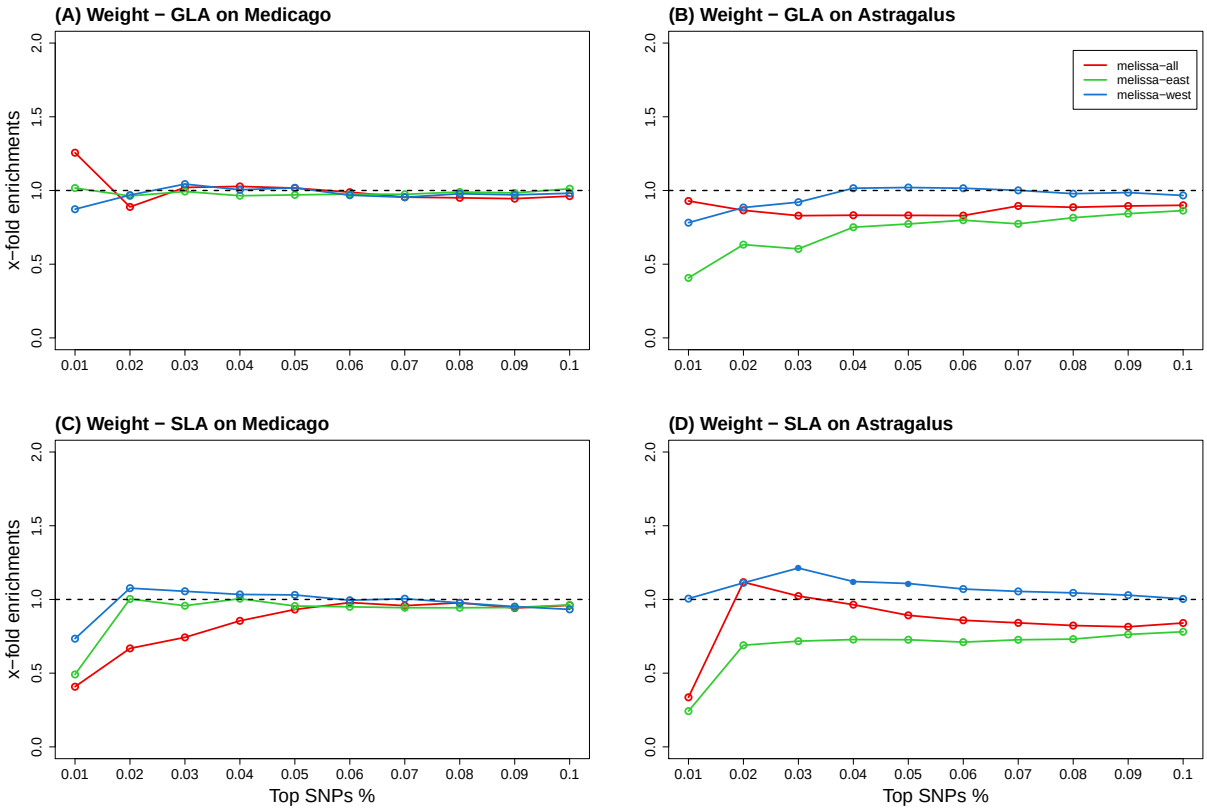


Fig. 3.18. S12 Line plots show x-fold enrichments across quantiles for overlapping SNPs between performance-associated SNPs in the rearing experiment and pairwise Fst-associated SNPs in nature for *ran1*. Open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

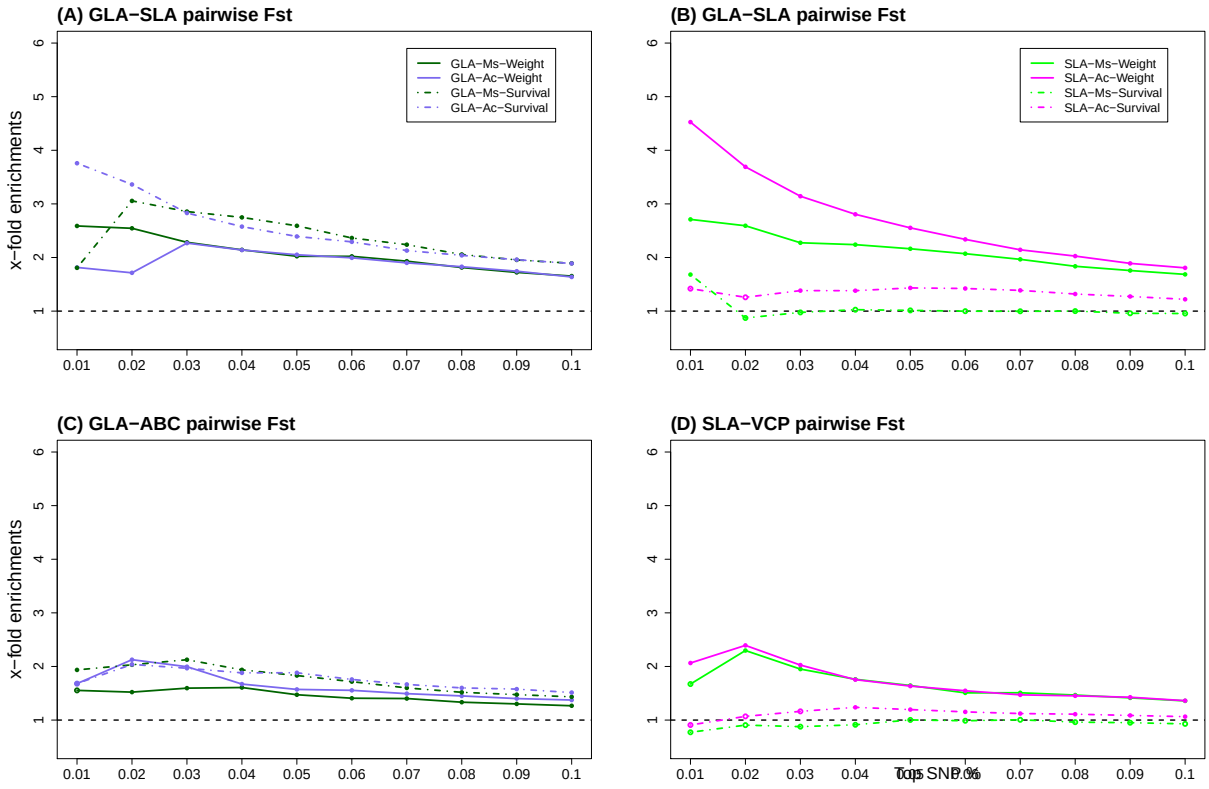


Fig. 3.19. S13 Line plot shows x-fold enrichments across quantiles for concordance in effect signs for overlapping SNPs between performance-associated SNPs in rearing experiment and pairwise Fst-associated SNPs in nature for *ran2A*. Open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.

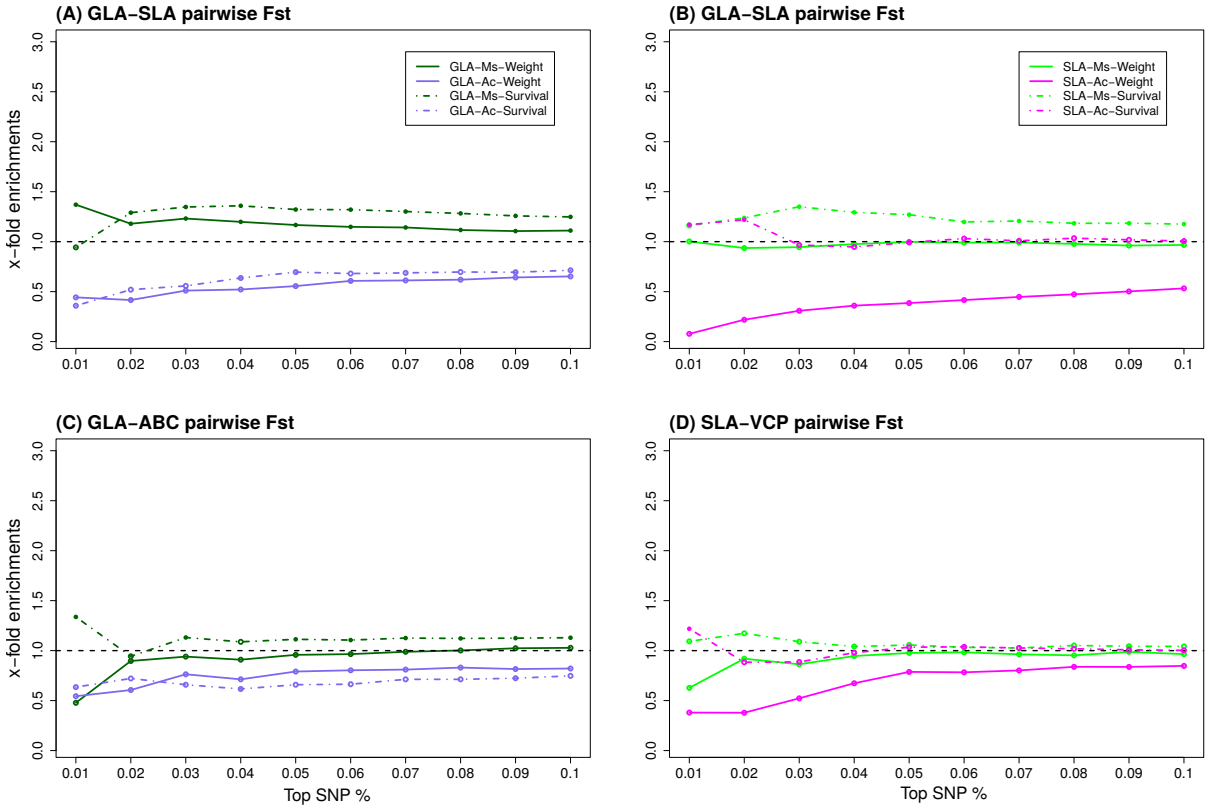
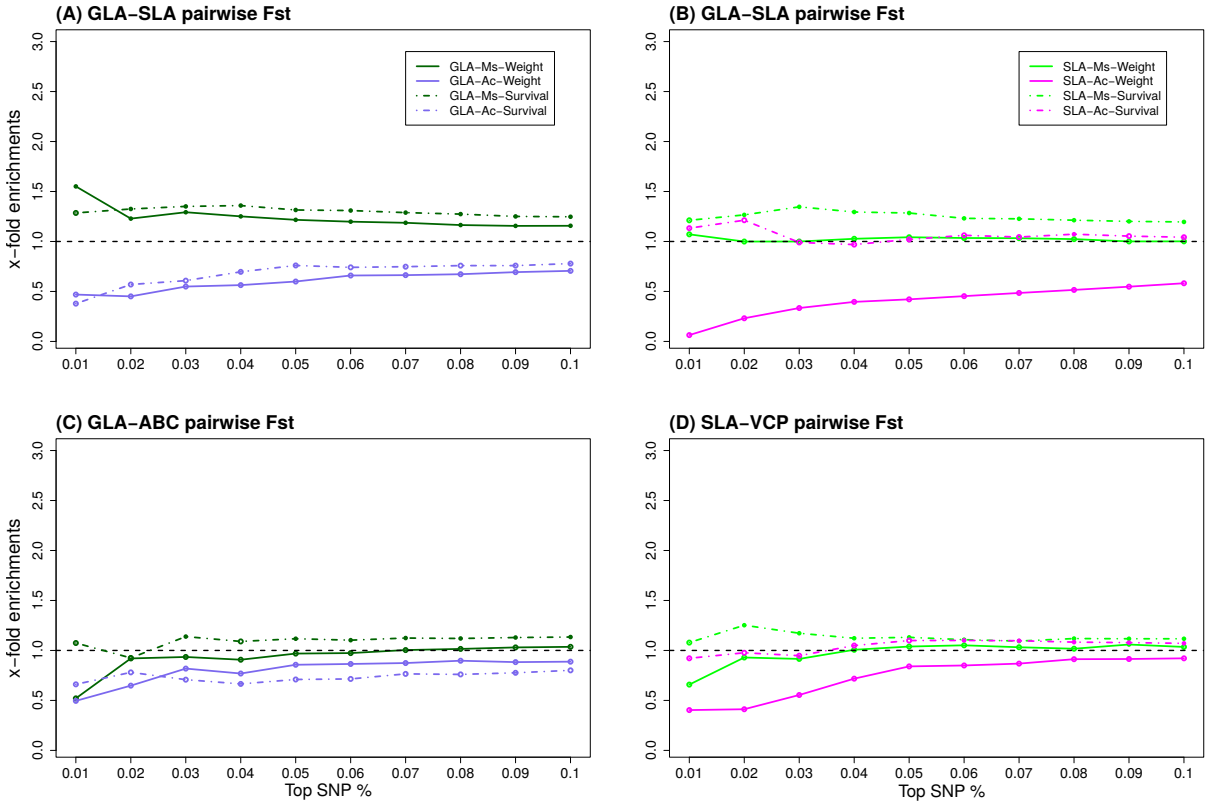


Fig. 3.20. S14 Line plots show x-fold enrichments across quantiles for concordance in effect signs for overlapping SNPs between performance-associated SNPs in the rearing experiment and pairwise Fst-associated SNPs in nature for *ran2B*. Open circles indicate $P > 0.05$ and filled circles indicate $P < 0.05$.



CHAPTER 4

SOURCES OF VARIATION IN THE GUT MICROBIAL COMMUNITY OF *LYCAEIDES MELISSA* CATERPILLARS [2]

Abstract

Microbes can mediate insect-plant interactions and have been implicated in major evolutionary transitions to herbivory. Whether microbes also play a role in more modest host shifts or expansions in herbivorous insects is less clear. Here we evaluate the potential for gut microbial communities to constrain or facilitate host plant use in the Melissa blue butterfly (*Lycaeides melissa*). We conducted a larval rearing experiment where caterpillars from two populations were fed plant tissue from two hosts. We used 16S rRNA sequencing to quantify the relative effects of sample type (frass versus whole caterpillar), diet (plant species), butterfly population and development (caterpillar age) on the composition and diversity of the caterpillar gut microbial communities, and secondly, to test for a relationship between microbial community and larval performance. Gut microbial communities varied over time (that is, with caterpillar age) and differed between frass and whole caterpillar samples. Diet (host plant) and butterfly population had much more limited effects on microbial communities. We found no evidence that gut microbe community composition was associated with caterpillar weight, and thus, our results provide no support for the hypothesis that variation in microbial community affects performance in *L. melissa*.

Introduction

Despite the low nutrient content, indigestibility and toxicity of many plant tissues, plant-feeding insects are among the most abundant and diverse groups of organisms on earth [1]. Herbivorous insects possess numerous morphological, behavioral and physiological traits that allow them to overcome these dietary obstacles [2]. Insect species and populations are often highly specialized [3, 4], feeding on one or a few families or even species of plants. Therefore, evolutionary shifts to

[2]This manuscript has been published in Scientific Reports and was coauthored by Lauren K Lucas, Alexandre Rego and Zachariah Gompert. Permission has been granted by the required coauthors for this research to be included in my dissertation (Appendix C).

new plant hosts can lead to speciation and catalyze further diversification [5, 6, 7, 8, 9]. Specific adaptations that allow insects to utilize novel plant hosts have been identified, and include changes in the structure or abundance of gut enzymes that reduce the toxicity of plant allelochemicals [10, 11, 12]. Changes in gut microbial communities could facilitate host plant shifts in a similar manner (e.g., [13]), but data in support of this are mostly lacking [14, 15, 16, 17].

Vertically transmitted microorganisms provide necessary nutritional benefits for some insects that have poor or unbalanced diets, such as phloem-sap or wood feeders [18, 19, 20]. For example, microorganisms could provision insects with essential amino acids which were missing in their diet [20]. Moreover, the acquisition of symbiotic microbes has been associated with major evolutionary shifts from non-plant-based to plant-based diets [16, 21]. Whether microbial symbionts facilitate more modest host plant shifts, that is host shifts or expansions to novel plant species or genera, is less clear [16, 17]. Gut microbes in particular have been hypothesized to shape host use and diet breadth by allowing herbivorous insects to detoxify specific plant allelochemicals (hereafter the “gut microbial facilitation hypothesis" [14, 17]). The high diversity of catabolic pathways available to microbes and the potential for gut microbes to interact with plant toxins make the gut microbial facilitation hypothesis an intriguing possibility [17]. Perhaps the strongest support for this hypothesis comes from studies of wild populations of pea aphids that found an association between the presence of specific gut microbes and host use [22]. However, experimental tests of the effects of these microbes on pea aphid performance have been inconsistent [22, 23, 24]. More generally, the lack of empirical support for the gut microbe facilitation hypothesis could be the result of a paucity of experimental studies designed to test it [16, 17], and additional evidence in support of this hypothesis is beginning to emerge [25]. Here we evaluate the gut microbial facilitation hypothesis in the Melissa blue butterfly (*Lycaeides melissa*).

Lycaeides melissa (Lepidoptera: Lycaenidae) occurs in the western US and southern Canada where it feeds exclusively on the leaves and flowers of legumes (Fabaceae); common native hosts include members of the *Astragalus* and *Lupinus* genera [26, 27]. Alfalfa (*Medicago sativa*) was introduced to the western US in the mid 1800s as a forage crop [28], and has since been colonized by *L. melissa*. Despite evidence of adaptation to this novel resource, *M. sativa* remains a poor host

relative to known native hosts [29, 30, 31]. For example *Lycaeides melissa* butterflies reared on *M. sativa* are smaller and suffer higher larval mortality than caterpillars reared on native hosts [29, 31]. Indeed, population persistence on alfalfa in the wild may depend on the presence of mutualistic ants that tend *L. melissa* caterpillars [29].

Variation exists within and among *L. melissa* populations for host acceptance and larval performance [29, 31]. Likewise, host plant populations and species differ in their average palatability to *L. melissa* [29, 32]. However, this variability in host acceptance and performance is poorly predicted by plant phytochemistry or protein content [32]. Microbial symbionts could explain some of this variation, and this is the focus of our current study. We consider only bacterial microbes at present, though complementary work investigating the role of fungal endophytes is underway [33].

Microbes could influence *L. melissa* host plant use in several ways. If endophytic or epiphytic microbial communities vary among host plants and caterpillars acquire their gut microbiome from their diet (e.g., [21, 34]), butterflies feeding on different populations or species of plant should have different gut microbes. Additionally, genetic differences among individual butterflies or populations could affect the gut environment in such a way that favors different gut microbes and thereby alters the gut microbial community. In either case, the resulting differences in gut microbiomes could be beneficial, detrimental, or have no effect on *L. melissa* fitness and population persistence. If caterpillar gut microbiome is mostly determined by diet (i.e., if it has a low heritability), then the ecological consequences of microbes on a host shift would be immediate. Caterpillars would acquire a new microbial community upon colonizing a new population or species of plant and experience any fitness consequences that follow. Alternatively, if gut variation in microbiome has a substantial genetic component (i.e., a non-negligible heritability), shifts in gut microbiome could occur over multiple generations due to evolution by genetic drift or selection and could be a key component of host-associated adaptation and specialization.

Herein we describe a larval rearing experiment conducted to measure the effects of diet (host plant) and population (a surrogate for genotype) on *L. melissa* caterpillar microbiomes. The experiment involved two *L. melissa* butterfly populations from northern Utah (USA): one in Blacksmith Fork Canyon, near Hardware Ranch (HWR; latitude = 41.6188° N, longitude = 111.5647° W) and

one along the Bonneville Shoreline Trail (BST; latitude = 41.7428° N, longitude = 111.7885° W). We lack data on the genetic similarity of these butterfly populations, but results from large genomic surveys of many *Lycaeides* populations show that even populations separated by short distances are genetically differentiated [35]. *Lycaeides melissa* feeds on alfalfa (*Medicago sativa*) and a native lupine (*Lupinus argenteus*) at HWR, whereas alfalfa is the only available host at BST (personal obs.). At HWR, the two host plant species are intermixed and butterflies can readily fly from one plant species to the other while laying eggs. Dispersal in caterpillars is much more limited.

Microbial communities were measured from plants, caterpillars and frass (caterpillar excrement) using high-throughput DNA sequencing of 16S rRNA. Because *L. melissa* at these two populations differ in host use, host-associated selection could differ between sites and lead to local adaptation, which could include adaptive differences in the caterpillar gut environment and consequently in gut microbial communities, or gut microbial communities could be determined mostly by diet and not be affected by genetic differences between these populations. We tested these alternatives. We also test for effects of microbial community composition and diversity on caterpillar performance (i.e., caterpillar weight). We show that caterpillars harbor a microbial community that varies over time, and that is minimally affected by source population or diet. We fail to find compelling evidence for an association between microbial community composition and larval performance in general, or in a host-specific manner.

Results

Microbial community structure

After removing chloroplast and mitochondrial sequences and performing rarefaction, we retained 59 samples (frass = 41, caterpillar = 10 and plant [endophytes and epiphytes] = 8) (sequence depth prior to rarefaction was not strongly associated with measures of OTU richness or diversity in the samples; Fig. 4.1). Microbial communities from frass, whole caterpillar and plant samples were mostly dominated by Betaproteobacteria, Gammaproteobacteria, Actinobacteria and Firmicutes (order Bacilli) (Fig. 4.2). The first two principal components (PCs) of the chord-transformed relative abundance matrix captured most of the variation in microbial community composition among the

caterpillar, frass and plant samples (57% of the total variation) (Fig. 4.3a). Caterpillar, frass and plant microbial communities overlapped in PC space, but sample types differed in their average PC scores and degree of variability (Table 4.1). Most notably, average caterpillar communities differed from frass and plant communities with respect to PC1 scores (Bayesian posterior prob. [pp] $\mu_{\text{Larvae}} > \mu_{\text{Frass}} > 0.99$; pp $\mu_{\text{Larvae}} > \mu_{\text{Plant}} > 0.99$). Frass and plant microbe communities were generally more similar. Complementary analyses based on a principal coordinate analysis (PCOA) of Bray-Curtis community dissimilarities gave similar results (PCs and PCOs were highly correlated: $|r_{PC1,PCO1}| = 0.98$, $|r_{PC2,PCO2}| = 0.64$, both $p < 0.0001$; Table 4.1; Fig. 4.7a). The effective number of phylotypes (that is "true diversity" = ${}^2D$) was higher and differed more among samples for the plant microbial communities (mean ${}^2D$, posterior median [pm] = 6.13, 95% ETPIs = [4.09, 8.06]; s.d. ${}^2D$, pm = 2.59, 95% ETPIs [1.63, 4.97]) than the caterpillar microbial communities (mean ${}^2D$, pm = 2.54, 95% ETPIs = [1.48, 3.64]; s.d. ${}^2D$, pm = 1.58, 95% ETPIs = [1.05, 2.75]). Intermediate diversity levels were observed in the frass microbial communities(mean ${}^2D$, pm = 4.14, 95% ETPIs = [3.52, 4.77]; s.d. ${}^2D$, pm = 1.98, 95% ETPIs = [1.62, 2.51]).

Random Forest (RF) (a decision tree classification method) was able to correctly classify most frass and caterpillar samples (as frass and caterpillars, respectively), whereas most plant samples were incorrectly classified as frass (Table 4.6). Discrimination between frass and caterpillar samples was mostly due to the fact that *Wolbachia* (Order Rickettsiales) was very common in the whole caterpillar samples but largely absent from the frass samples (GINI index = 4.89; whole caterpillar relative abundance: mean = 0.37, s.d. = 0.36; frass relative abundance: mean = 0.0007, s.d. = 0.003; Table 4.4). This pattern is unsurprising, as *Wolbachia* is a common intracellular symbiont in arthropods and has been found in *Lycaeides melissa* [36] and other Lycaenid butterflies [37].

Determinants of frass and caterpillar microbial communities

After removing the *Wolbachia* sequences and re-rarifying the OTU table, we retained 53 samples (42 frass and 11 whole caterpillars) for subsequent analyses of frass and caterpillar microbial communities. Hierarchical clustering of the frass and whole caterpillar samples based on differences in microbial communities did not show distinct clusters based on age, source population, host plant (diet) or sample type. However, several small clusters or groups in the dendrogram consisted of frass

or whole caterpillars of the same age or reared on the same host-plant (Figs. 4.4, 4.8).

The first two PCs from an ordination of only the frass and whole caterpillar microbiomes (chord-transformed relative abundances) captured most of the variation in microbial community composition among these samples (61% of the total variation in the chord transformed relative abundances) (Fig. 4.3b). PC1 and PC2 reflected variation in the relative abundance of several Proteobacteria (Table 4.3). Bayesian linear models showed that microbial communities (as measured by PCs) were mostly affected by caterpillar age (days since hatching) and sample type (frass vs. whole caterpillar) (PCOs and PCs were highly correlated, $|r_{PC1,PCO1}| = 0.99$, $|r_{PC2,PCO2}| = 0.98$, both $p < 0.0001$, and thus analyses based on PCOs gave similar results; Fig. 4.7b). Specifically, even after removing *Wolbachia* sequences, frass and whole caterpillars contained different microbiomes (β_{type} for PC1, pm = -0.418, 95% ETPIs = [-0.685, -0.150]). The microbial communities of caterpillar guts (measured from frass and whole caterpillar samples) also changed over time with respect to PC1 scores (β_{age} for PC1, pm = -0.048, 95% ETPIs = [-0.076, -0.020]). Similarly, phylotype diversity was lower in frass and whole caterpillar microbial communities from older larvae (β_{age}, pm = -0.151, 95% ETPIs = [-0.296, -0.005]). Based on these estimates, diversity dropped by almost two effective species between the 15 and 25 day samples, with the most pronounced shift occurring between the 20 and 25 days, that is, late in larval development (pupation occurred at between 23 and 28 days of development). We failed to detect credible effects of butterfly population or plant species on community composition or diversity, and more generally, RF failed to accurately discriminate between samples from different sources (frass vs. whole caterpillar), host plant treatments, populations or samples collected at different larval ages (Tables 4.7, 4.8). Instead, RF generally assigned most samples to the more common group.

Despite the lack of a clear effect of our primary treatments (butterfly population and host plant) on community composition, we identified several microorganisms with significantly different relative abundances in different butterfly population × plant species treatment combinations. This was done using a Bayesian multinomial-Dirichlet for relative abundance counts and considering only frass samples from 15 or 20 day old larvae (we focused on this subset of samples to maximize the sample size while minimizing the confounding effects of caterpillar age and sample type documented above;

Fig. 4.6). Lactobacillales (Bacilli) and Rhodospirillales (Alphaproteobacteria) were more abundant in frass samples from caterpillars reared on *L. argenteus* (Lactobacillales: pm, *L. argenteus*-HWR [L-HWR] = 0.082, *M. sativa*-HWR [M-HWR] = 0.011, *M. sativa*-BST [M-BST] = 0.006; Rhodospirillales: pms, L-HWR = 0.0030, M-HWR = $2.3e^{-4}$, M-BST = $3.6e^{-}5$), whereas Pseudomonadales (Gammaproteobacteria) were more abundant in frass from caterpillars fed *M. sativa* (pm, L-HWR = 0.087, M-HWR = 0.193, M-BST = 0.193). Rhizobiales (Alphaproteobacteria) were more abundant in frass from BST caterpillars (pm, L-HWR = 0.010, M-HWR = 0.011, M-BST = 0.18), and Enterobacteriales (Gammaproteobacteria) and Sphingomondadales (Alphaproteobacteria) differed in relative abundance based on host plant and butterfly population (Enterobacteriales: pm, L-HWR = 0.087, M-HWR = 0.013, M-BST = 0.065); Sphingomondadales, pm, L-HWR = 0.178, M-HWR = 0.117, M-BST = 0.161)). In each of these cases, the posterior probability for sample differences was ≥ 0.99, and the posterior predictive root-mean square error (RMSE) was significantly lower for a model allowing for different microbe relative abundances for each treatment combination than a constrained null model (pp = 0.969).

Microbial community and larval performance

Thirty-one percent of the 181 caterpillars survived to 15 days, that is, to when the first frass samples were collected and larval weight was measured. We found greater evidence for an effect of population on survival than plant (β_{pop}, pm = -0.56, 95% ETPIs = [-1.22, 0.07]; β_{plant}, pm = -0.15, 95% ETPIs = [-0.84, 0.51]), such that probabilities of survival were 0.38 (95% ETPIs = 0.27, 0.50) for HWR caterpillars on *M. sativa*, 0.35 (95% ETPIs = [0.22, 0.49]) for HWR caterpillars on *L. argenteus*, 0.26 (95% ETPIs = [0.17, 0.37]) for BST caterpillars on *M. sativa*, and 0.23 (95% ETPIs = 0.13, 0.37) for BST caterpillars on *L. argenteus*.

We next tested for an association between microbial communities and caterpillar weigth (a metric of performance). We focused on PC1, PC2, PCO1 and PCO2 (measures of community composition) and phylotype diversity (^2D) for frass samples from 15 and 20 day old caterpillars (that is, from caterpillars that survived long enough for the first frass samples to be taken; N = 31 samples). The best model (lowest DIC) was the base model with plant, population and age (i.e., with no effect of microbial community), but several other models that included effects of microorganisms

had only slightly worse DIC values (Table 4.5). Unsurprisingly, caterpillar weight increased with age (β_{age}, pm = 0.225, 95% ETPIs = [0.098, 0.353]) in the base model. We found that feeding on *M. sativa* reduced caterpillar weight relative to feeding on *L. argenteus* (β_{plant}, pm = -0.961, 95% ETPIs = [-1.643, -0.281], pp $< 0 = 0.99$).

Discussion

Simple models of genetic trade-offs have mostly failed to explain host specialization in herbivorous insects [31, 38, 39]. However, experimental tests have rarely considered symbionts, such as gut microbes, which can mediate interactions between insects and their hosts [16, 40, 41]. As proposed by the gut microbial facilitation hypothesis, microbial communities have to affect fitness and have a non-zero heritability for adaptive shifts in gut microbes to contribute to host use evolution by insects [17]. Our current study represents an initial attempt to evaluate evidence for and against this hypothesis (also see [25, 42]). We failed to find a convincing association between microbial community and larval performance. Instead we found that microbes mostly varied over time and differed between frass and whole caterpillar samples, with frass samples harboring microbiomes that were more similar to the plant microbial communities. We found minimal overall effects of butterfly population or diet (host plant) on gut microbiomes, but did identify several microorganisms that differed in their relative abundances across treatments. Thus, in total, our results do not suggest that genetic differences among *L. melissa* populations contribute substantially to adaptive variation in microbial communities (at least not for the populations we studied). Nonetheless, gut microbes could contribute to host use evolution in *L. melissa*, if for example, microbial variation among plants affects whether initial colonization of a new host is possible. Additionally, as we only considered a single pair of butterfly populations, genetic variation for microbial communities could certainly exist at greater spatial scales (i.e., between more distant populations) or even among individuals within some populations. Thus, our current results neither strongly support nor refute the gut microbial facilitation hypothesis in *L. melissa*.

We discuss our results and these issues in more detail below, but first, three potential limitations of our study should be noted. First, we used microbial communities from frass as a proxy for the gut microbial communities in caterpillars. This approach has been used in previous studies of caterpillar

gut communities [37, 43] and allows for non-destructive sampling of community composition and diversity over time or developmental stages. But, even though frass and caterpillar communities were more similar after removing *Wolbachia* sequences, they still differed. This means that some differences certainly exist between caterpillar gut microbial communities (because caterpillars were surface sterilized, we expect them to be enriched for gut microbes) and the microbial communities we sampled from frass, and these differences could affect some of our conclusions. Second, we failed to amplify DNA from about half of our samples (we purified DNA from 145 samples but only successfully sequenced microbes in 69 of them). We think that this reflects variation in the abundance of microbes, particularly in the small frass samples from early instar caterpillars (where we had the least success). An effect of raw microbe abundance on amplification success could introduce some bias, but we would not expect it to bias results in terms of comparisons among treatments (i.e., raw microbe abundance does not appear to vary by treatment). Finally, it is almost certain that our sequence data include contaminant microorganisms, as (i) it is unlikely that the sterilization procedures fully eliminated non-target microbes, and (ii) microbes are also often present in DNA extraction kits and reagents [44] (which is something that we cannot currently quantify based on our existing data). Nonetheless, our main interest was in differences among samples, and thus, we do not think that contaminants have created false positive signals (but they could have obscured true signals). In other words, contamination should alter microbial communities, but not create differences across treatments, particularly as all samples were processed with the same kit and reagents and in the same lab.

We had clear evidence that microbial community composition in *L. melissa* caterpillar guts shifted over time and exhibited a decrease in diversity. A similar pattern was recently found in *L. melissa* fungal communities [33]. Temporal variation in microbial communities has also been documented in *Heliconius* butterflies, but at different developmental stages (larvae versus pupae versus adults) rather than within a single developmental stage. Despite this temporal variation, several phylotypes were common across many of the *L. melissa* frass and larvae samples, including Actinobacteria, Proteobacteria and Bacilli (Firmicutes). Proteobacteria have frequently been found in other insects, including other Lepidoptera, plataspid bugs, alydid bugs, reed beetles, bees and

termites [25, 45, 46]. Proteobacteria could play a role in nutrient provisioning and degradation of toxins [46]. Our results show that *L. melissa* harbor all three classes of Proteobacteria (Alpha, Beta and Gamma). Alphaproteobacteria was also identified as a core bacteria in mosquito species with different diets [42] and can be horizontally and vertically transferred [47, 48]. Fimicutes (Bacilli, order Lactobacillales) were also common in *L. melissa* frass and larvae. Firmicutes have been reported in other Lycaenid butterflies [37], other Lepidopteran larvae [25, 45, 49], fruit flies and ground beetles [50, 51]. Firmicutes have been identified to play a role in nutrient provisioning, food digestion and fermentation [46]. Actinobacteria have also been reported in Lepidoteran larvae [25, 45] and have been reported to play a role in nutrient provisioning [46].

Neither diet (host plant speices) nor butterfly population (BST vs. HWR) had a detectable effect on overall microbial community composition. Nonetheless, diet was associated with the relative abundance of a few common microbes (Lactobacillales, Rhodospirillale, and Pseudomonadales), and similarly, a few microbes were more abundant in specific populations (Enterobacteriales, Sphingomondadales, and Rhizobiales). Diet has been shown to affect gut microbial communities in insects, including other Lepidoptera [25, 37, 45, 49], bees [52], *Drosophila* [53] and mosquitoes [42]. An effect of diet on microbial community has also been shown in mammals, including humans[34], suggesting that this is a general mechanism shared by distantly related taxa. Thus, the fact that our results show an effect of diet on at least some microbes is unsurprising, and the limited nature of this effect might reflect similarities in the microbiomes of the two plant species (we lacked sufficient sample sizes to formally test for differences between plant species, but the communities from the *M. sativa* and *L. argenteus* plants overlapped in ordination space). In contrast, a recent study of Lycaenid butterflies failed to show a consistent effect of diet (including herbivory versus carnivory) on caterpillar gut microbiomes [37]. But, this study considered a few individuals across many different species of butterflies, so the lack of consistency is not evidence for a lack of an effect of diet within butterfly species.

Our results suggest that genetic differences between the two butterfly populations have at most a limited effect on the gut environment as perceived by most of the detected gut microbes. Perhaps this is unsurprising, as these populations occur in similar environments (mid-elevation, dry montane

environments), have one of the same host plants (*M. sativa*; *L. argenteus* is used as an additional host at HWR) and are only separated by about 20km. Thus, insufficient time, on-going gene flow or limited divergent selection could explain this lack of genetic divergence in microbial communities, and thus genetic divergence in microbial communities is still possible for more distant populations or those that differ more in host use. Likewise, genetic variation for gut microbiomes could exist within populations, but testing for this would require larger sample sizes (more families and caterpillars per family).

We failed to find compelling evidence for an association between larval performance (weight) and microbial community composition, though it is still possible that microbes are associated with other fitness components or metrics. Our results differ from a recent study that detected an association between growth and gut microbiome in the Glanville fritillary (*Melitaea cinxia*) [25]. Likewise, an association between microbial community and fitness has been detected in *Drosophila* and pea aphids [22, 23, 54]. But, in most of these cases (as would have been the case in our study) it is unclear whether the microbiome affects insect performance or insect performance affects the microbiome (i.e., different microbes could be favored in healthier insects) or both. Experiments whereby microbes are directly manipulated have been conducted in pea aphids to test these alternatives, but the results have been inconclusive [22, 23, 24]. Studies in humans suggest complex interactions between gut microbes and health that include feed-backs [55, 56, 57, 58]. Similar complexity could exist in herbivorous insects.

In conclusion, we failed to find convincing evidence that gut microbes play a role in host-plant adaptation in *L. melissa*. This might or might not be a general pattern in Lepidopterans. Additional work to elucidate the host specific effects of microbiomes on fitness is critical, as microbes can only mediate adaptation to novel hosts if their effects on performance differ across hosts. Otherwise, their role would be limited to adaptation to herbivory in general. Larger studies that consider additional components of fitness, such as those we have planned for *L. melissa*, are needed to better parse these effects.

Methods

Larval rearing experiment

In June of 2014, female *L. melissa* butterflies were captured at BST (N = 31) and HWR (N=23) and caged individually in oviposition cages to lay eggs (as in [59]). After 48 hours, eggs were collected and stored in petri dishes at room temperature under bright lights until hatching. We obtained 182 neonate larvae from 20 of the females that layed eggs (i.e., not all females layed eggs; mean number of caterpillars per family = 9.1, s.d. = 9.6). Neonate larvae were transferred individually to new petri dishes once they hatched. Each caterpillar was fed exclusively on *M. sativa* (alfalfa) from BST or HWR, or *L. argenteus* from HWR. These diet treatments were assigned in alternation when caterpillars hatched. Fresh plant material was collected from the field once a week and fed to larvae *ad libitum* as small sprigs without flowers and with leaf petioles wrapped in damp Kimwipes. Petri dishes were checked and cleaned daily. Caterpillars were reared at room temperature on lab bench tops under 12-h light:dark cycles as we have done for other experiments with *Lycaeides* butterflies [29, 31].

Petri dishes were checked daily to determine whether caterpillars were alive or dead, and survival time in days was recorded for dead caterpillars. As a second measure of performance, larval weight was quantified at 15, 20 and 25 days (weight was only measured for living caterpillars that had not yet pupated). Caterpillars were weighed on a Mettler Toledo XS64 microbalance to the nearest 0.1 mg (an average of three measurements was recorded). Frass was collected from each petri dish at 15, 20 and 25 days as well, and then stored by freezing at -80° C in 1.5 mL tubes for subsequent microbial DNA extraction. A previous study with *Heliconius* butterflies showed that frass communities are a good proxy for gut microbial communities sampled from whole caterpillars [43]. Thus, frass samples can provide a non-lethal way to sample caterpillar gut microbes over time and without contamination from cellular endosymbionts commonly found in Lepidoptera, such as *Wolbachia* [36, 37]. Nonetheless, more substantial differences between frass and gut communities could occur in some systems. Thus, 38 randomly chosen caterpillars were sacrificed and frozen at 15 (N = 14), 20 (N = 16) or 25 (N = 8) days so that gut microbial communities from caterpillars could be compared with the frass communities.

DNA extraction and sequencing

DNA was isolated from the frass (N = 93) and whole caterpillar (N = 40) samples described above and from field-collected plant samples (N = 12, details follow). Prior to extraction, frozen caterpillars were surface-sterilized by rinsing them three times for one minute each in a 1% Sodium Hypochlorite (diluted bleach), 95% ethanol and deionized water (to rinse the samples). This was done to reduce the prevalence of surface microbes while leaving gut microbes intact. This procedure is unlikely to have removed all surface microbes, but it should enrich our samples for gut rather than surface microbes. Fresh leaf tissue was collected from *M. sativa* and *L. argenteus* at BST and HWR (leaves from four plants per site and species). Leaves were collected during the rearing experiment, so that they would be representative of the age and phenology of leaves being fed to the caterpillars. Surface microbes (epiphytes) were isolated by washing plant leaves in an isotonic 0.1 M PBS buffer. Bacterial cells were then extracted from the PBS buffer by passing the buffer through a Nalgene 0.2 micron vacuum filter; the filter paper was then cut into small pieces and used as a template for DNA extraction. We then sterilized the remaining leaf tissues by washing with 1% Sodium Hypochlorite, 95% ethanol and deionized water (as described above) and retained this tissue for isolation of endophytic microbial DNA.

Genomic DNA was extracted from frass, caterpillar and plant (epiphytes and endophytes) samples using the MoBio PowerSoil kit according to the manufacturers standard protocol. We then amplified the V4 region of the 16S rRNA gene using the standard PCR primers 515F and 806R and in accordance with the recommended protocol from the Earth Microbiome Project and as described previously in [60]. The primer design used included unique barcode sequences for index reads so that samples could be multiplexed. We successfully amplified DNA from 69 of the 145 samples. 16S rRNA amplicon libraries for these samples were sequenced at the University of Texas Genomic Sequencing and Analysis Facility (Austin, TX) on the Illumina MiSeq platform. We obtained 13.4 million 250-bp, paired-end reads.

Identification of OTUs

We used the Quantitative Insight into Microbial Ecology (QIIME) pipeline to assign phylotypes

(operational taxonomic units or OTUs) to 16S rRNA sequences and to estimate the relative abundance of each OTU in each sample [61]. We specifically used QIIME's open reference-based OTU picking strategy [61], which uses UCLUST to cluster sequences [62]. Sequences were first clustered against the Green Genes database (ver. 13-08) with a minimum of 97% sequence similarity [63]. A subset of sequences that did not cluster in this first step were clustered *de novo*, and the centroids of the new clusters were used to generate an additional sequence set for reference-based clustering (also at 97% sequence similarity). A final round of *de novo* clustering was conducted with sequences that did not match this reference sequence set. Taxonomic identifications were then assigned from the Green Genes database based on the centroid of each cluster [64].

The 16S rRNA primers 515F and 806R also amplify chloroplast and mitochondrial DNA [65, 66]. Thus, sequences identified by the Green Genes database as chloroplast or mitochondrial 16S rRNA were removed before downstream analysis. We used rarefaction to ensure comparisons among microbial communities were not biased based on differences in the number of sequences obtained. Specifically, we randomly retained 1311 16S rRNA sequences from each sample for downstream analysis (this number was chosen as a compromise between removing samples and sequences). The samples which had fewer than 1311 bacterial sequences were removed from the analysis, which decreased our sample size from 69 samples to 60 samples: 41 frass communities, 10 caterpillar communities, and nine plant communities (five epiphyte and four endophyte samples). We also dropped unassigned OTUs before proceeding with our downstream analysis.

Statistical analyses of community structure

All statistical analyses were conducted using the R statistical computing environment (R version 3.3.1 [67]). We used two complementary ordination methods to summarize patterns of microbial community composition across the frass, caterpillar and plant (epiphytes and endophytes) samples (N = 60 communities). First, we conducted a principal component analysis (PCA) on the centered (but not scaled) chord-transformed relative abundance matrix. Chord transformation prior to PCA reduces the tendency for patterns to be driven by shared absences of microbes, while still allowing for ordination of (transformed) relative abundances [68, 69]. Second, we used principal coordinate analysis (PCOA) to ordinate samples based on pair-wise Bray-Curtis dissimilarities. The two

methods gave very similar results (e.g., the Pearson correlation between PC1 and PCO1 scores was 0.98), and thus we mostly focus on the PC scores in the main text. Results based on PCOA are provided in the supplemental material. The `prcomp` function in `R` was used for PCA, the `vegdist` function in `Vegan` package was used to calculate the Bray-Curtis dissimilarities [70], and the `pcoa` function in `ape` was used for PCOA [71].

We next quantified the OTU diversity of microbial communities using Hill numbers [72, 73]. Specifically, we calculated the "true" diversity or effective number of phylotypes in each sample as ${}^{q}D = \left(\sum_i p_i^q\right)^{1/(1-q)}$; p_i is the relative abundance of OTU i in the sample, and q determines the weight given to rare phylotypes. We chose $q = 2$, as lower values are unlikely to yield reliable estimates of diversity for microbial communities [74]. In addition we calculated effective number of species per sample using the same diversity index to determine the magnitude of diversity retained for each sample as sequencing depth increases.

We used Bayesian models to quantify differences in microbial community composition (PC1, PC2, PCO1 and PCO2 scores) and diversity (${}^{q=2}D$) among sample types, that is, frass, caterpillar and plant communities (ephiphytes and endophytes were pooled because of small sample sizes and because visual inspection of PCA plots suggested these communities were similar). The model we used can be viewed as a Bayesian analog to a single factor ANOVA, but without the constraint of equal variances across treatments. For each metric (PC1, PC2, PCO1, PCO2 and ${}^{2}D$) we assumed that the scores or values from each sample type could be characterized by a Normal distribution with an unknown mean and standard deviation (s.d.) that we estimated from the data. We placed uninformative priors on the mean (Normal, with $\mu = 0$, and $\tau = \frac{1}{\sigma^2} = 1e-6$) and precision (that is, the inverse of the variance; gamma with shape and rate parameters equal to 0.01) for each group.

We then used a classification method, Random Forest (RF), to determine whether frass, whole caterpillar and plant communities could be assigned to their respective sample type and if so, to identify the microbes or combination of microbes that were most important for accurate classification. RF was performed in R with the `randomForest` function using 50,000 trees [75]. An advantage of this approach (e.g., compared to discriminant analysis) is that the number of observations (samples) does not have to be larger than the number of variables used for classification (i.e., the number of

microbial OTUs).

We found that the microbial communities from frass and whole caterpillar samples differed, mostly because the intracellular endoysmbiont *Wolbachia* (Order Rickettsiales) was highly abundant in the caterpillars but almost completely absent from the frass (see Results for details). *Wolbachia* is not a free-living member of the gut microibal community, but rather a vertically transmitted intracellular endosymbiont found in germ and somatic tissues [76]. Thus, we removed the *Wolbachia* sequences before continuing with tests of the determinants of gut microbial community. Specifically, we re-rarified the original OTU table after (i) dropping chloroplast, mitochondrial and *Wolbachia* sequences, (ii) removing the plant samples (which were not needed for additional analyses), and (iii) dropping rare OTUs with a relative abundance of $< 1\%$ in all samples. For this rarefaction, we randomly retained 500 sequences from each sample for downstream analysis. This decreased our sample size to 53 samples (42 frass communities and 11 caterpillar communities). Note that one frass and one caterpillar retained in this round of rarefaction were dropped from the first set of analysis as the minimum number of sequences required for retention was different. We again removed any unassigned OTUs before proceeding with downstream analysis.

Determinants of frass and caterpillar microbial communities

We investigated the effect of diet, population and caterpillar age on the microbial communities detected in the frass and whole caterpillar samples (N = 53) using the re-rarefied OTU data. Using both frass and whole caterpillar samples allowed us to increase our sample size and assess remaining differences between these sample types after removing *Wolbachia* sequences. We first used a hierarchical clustering approach to test the relatedness of bacterial communities in frass and larvae. We used Bray-Curtis dissimilarities (recommended for relative abundance data) and the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method for clustering to determine relatedness in bacterial communities in frass and larvae samples [77, 78]. The `vegdist` function in `Vegan` package in R was used to calculate the dissimilarity matrix [70]. Hierarchical Clustering was performed by using the `hclust` function in `R` [79]. `heatmap.2` function from `gplots` package in `R` was used to visualize the distance matrix and clustering results [80].

We then calculated phylotype diversity (^2D) for each frass and whole caterpillar sample and

conducted a second PCA and PCOA to summarize microbial community patterns across these samples. Bayesian linear models were used to quantify the effect of sample type (frass or whole caterpillar), plant, population (all binary covariates), and days since hatching (i.e., larval age or development) on microbial community composition (PC1, PC2, PCO1 and PCO2 scores) and diversity (^2D). Uninformative priors were placed on the regression coefficients, $\beta_{type}, \beta_{plant}, \beta_{pop}$ and β_{age} (Normal, with $\mu = 0$, and $\tau = \frac{1}{\sigma^2} = 1e-6$) and on the precision term (gamma with shape = 0.01 and rate = 0.01).

We also tested for differences in the relative abundance of individual microbes. We did this by fitting Bayesian multinomial-Dirichlet models for the rarefied OTU count data. This model assumes that microbe counts for each sample are drawn from a multinomial distribution with success probabilities (π) give by the true relative abundances. We placed an uninformative prior on the vector of true relative abundances ($\pi \sim$ Dirichlet$(1, \ldots, 1)$). As this model does not readily incorporate covariates, we focused on a more homogeneous subset of the data: frass samples at 15 to 20 days. This minimized the effects of age and sample type while maximizing the sample size: $n = 35$, which includes frass from 12 HWR caterpillars reared on *L. argenteus*, 10 HWR caterpillars reared on *M. sativa* and 13 BST caterpillars reared on *M. sativa* (we removed the single sample from a BST caterpillar reared on *L. argenteus* that otherwise met the criteria for inclusion as the sample size, $n = 1$, was too low for valid inference). We fit and compared models (i) allowing the true relative abundance vector (π) to vary among the three source population $\times$ host plant (diet) treatments, and (ii) constraining the true relative abundances to be the same (i.e., a null model assuming no differences in microbiome among samples). We used the posterior predictive distribution of the count data (predicted OTU counts from the model) to measure model performance by calculating the root-mean square error between the observed and predicted counts.

Bayesian parameter estimates were obtained using Markov Chain Monte Carlo (MCMC) via the `R` interface with `JAGS` provided by the `rjags` package [81] (most analyses) or by direct simulation from the closed form posterior in `R` (the multinomial-Dirichlet model has a closed form solution and can be directly sampled). Three replicate MCMC runs (chains) were used for each analysis. Each chain included a 1000 iteration burn-in followed by 10,000 iterations where every third sample was

retained for inference. Adequate MCMC mixing and likely convergence to the stationary distribution were verified by quantifying the effective sample size and calculating the Gelman Rubin convergence diagnostic. We used the median of the marginal posterior distribution for each parameter as a point estimate (denoted 'pm' for posterior median). Uncertainty in parameter estimates was quantified based on marginal 95% equal-tail probability intervals [ETPIs], and, in some cases, based on the posterior probability (hereafter, 'pp') that a parameter or the difference between two parameters was greater than or less than 0. Finally, we re-ran the RF classification analysis to determine whether frass and whole caterpillar communities could be assigned to sample type (frass vs. whole caterpillar), diet/host plant (*M. sativa* vs. *L. argenteus*), population (BST vs. HWR) and caterpillar age (15 vs. 20 vs. 25 days), and if so, to identify the microbes or combination of microbes that were most important for accurate classification.

Tests for an effect of gut microbiome on larval performance

We fit Bayesian models to quantify the effect of food plant (diet) and butterfly population on larval performance and to determine whether microbial community explained additional variation in performance. First, we used a Bayesian generalized linear model with a Bernouli error distribution and logit link function to quantify the effect of plant (β_{plant}) and population (β_{pop}) on caterpillar survival to 15 days (that is the time when the first frass sample was taken). Survival data from 181 caterpillars were used for this analysis. Next, we fit and compared alternative models for caterpillar weight. All models included a potential effect of plant (β_{plant}), population (β_{pop}) and time since hatching (larval age; only 15 and 20 day-old caterpillars were included, as the sample size for 25 day caterpillars was very small and these caterpillars were much larger; β_{age}). We considered models with just these effects or these effects plus measures of microbial community composition from frass samples collected at 15 or 20 days (PC1, PC2, PCO1 and PCO2 scores) or community diversity (^2D). We tested for possible interactions between microbial community and diet (plant) on weight, which would suggest a potential role of microbes in host-specific adaptation. For all models, uninformative priors were placed on the regression coefficients, β_{type}, β_{plant}, β_{pop}, β_{age} and β_{PC}, β_{PCO} or $\beta_{^2D}$ (Normal, with $\mu = 0$, and $\tau = \frac{1}{\sigma^2} = 1e-6$) and on the precision term (gamma with shape = 0.01 and rate = 0.01). Model comparisons were based on deviance information criterion (DIC), which is

similar to Akaikie information criterion but appropriately penalizes Bayesian models based on the effective number of parameters (this is readily calculated from MCMC output).

REFERENCES

[1] May, R. M. How many species are there on earth? *Science* **241**, 1441–1449 (1988). URL http://science.sciencemag.org/content/241/4872/1441. DOI 10.1126/science.241.4872.1441. http://science.sciencemag.org/content/241/4872/1441.full.pdf.

[2] Schoonhoven, L. M., van Loon, J. J. A. & Dicke, M. *Insect-Plant Biology* (Oxford University Press, 2010), second edn.

[3] Jaenike, J. Host specialization in phytophagous insects. *Annual Review of Ecology and Systematics* 243–273 (1990).

[4] Forister, M. L. *et al.* The global distribution of diet breadth in insect herbivores. *Proceedings of the National Academy of Sciences* **112**, 442–447 (2015). URL http://www.pnas.org/content/112/2/442.abstract. DOI 10.1073/pnas.1423042112. http://www.pnas.org/content/112/2/442.full.pdf.

[5] Farrell, B. D. "inordinate fondness" explained: Why are there so many beetles? *Science* **281**, 555–559 (1998).

[6] Drès, M. & Mallet, J. Host races in plant–feeding insects and their importance in sympatric speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **357**, 471–492 (2002).

[7] Rundle, H. D. & Nosil, P. Ecological speciation. *Ecology letters* **8**, 336–352 (2005).

[8] Fordyce, J. A. Host shifts and evolutionary radiations of butterflies. *Proceedings of the Royal Society of London B: Biological Sciences* **277**, 3735–3743 (2010).

[9] Hood, G. R. *et al.* Sequential divergence and the multiplicative origin of community diversity. *Proceedings of the National Academy of Sciences* **112**, E5980–E5989 (2015).

[10] Stevens, J. L., Snyder, M. J., Koener, J. F. & Feyereisen, R. Inducible {P450s} of the {CYP9} family from larval *Manduca sexta* midgut. *Insect Biochemistry and Molecular Biology* **30**, 559 – 568 (2000). URL http://www.sciencedirect.com/science/article/pii/S0965174800000242. DOI http://dx.doi.org/10.1016/S0965-1748(00)00024-2.

[11] Zhu-Salzman, K., Koiwa, H., Salzman, R., Shade, R. & Ahn, J.-E. Cowpea bruchid *Callosobruchus maculatus* uses a three-component strategy to overcome a plant defensive cysteine protease inhibitor. *Insect Molecular Biology* **12**, 135–145 (2003).

[12] Despres, L., David, J.-P. & Gallet, C. The evolutionary ecology of insect resistance to plant chemicals. *Trends in Ecology & Evolution* **22**, 298–307 (2007).

[13] Tsuchida, T., Koga, R. & Fukatsu, T. Host plant specialization governed by facultative symbiont. *Science* **303**, 1989–1989 (2004).

[14] Berenbaum, M. Allelochemicals in insect-microbe-plant interactions: Agents provocateurs in the revolutionary arms race. *Novel Aspects of Insect-Plant Interactions, eds Barbosa P, Letourneau DK (Wiley, New York)* 97–123 (1988).

[15] Feldhaar, H. Bacterial symbionts as mediators of ecologically important traits of insect hosts. *Ecological Entomology* **36**, 533–543 (2011).

[16] Hansen, A. K. & Moran, N. A. The impact of microbial symbionts on host plant utilization by herbivorous insects. *Molecular Ecology* **23**, 1473–1496 (2014).

[17] Hammer, T. J. & Bowers, M. D. Gut microbes may facilitate insect herbivory of chemically defended plants. *Oecologia* **179**, 1–14 (2015). URL http://dx.doi.org/10.1007/s00442-015-3327-1. DOI 10.1007/s00442-015-3327-1.

[18] Douglas, A. E. Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria buchnera. *Annual Review of Entomology* **43**, 17–37 (1998).

[19] Moran, N. A., Plague, G. R., Sandström, J. P. & Wilcox, J. L. A genomic perspective on nutrient provisioning by bacterial symbionts of insects. *Proceedings of the National Academy of Sciences* **100**, 14543–14548 (2003).

[20] Moran, N. A., McCutcheon, J. P. & Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. *Annual Review of Genetics* **42**, 165–190 (2008).

[21] Russell, J. A. *et al.* Bacterial gut symbionts are tightly linked with the evolution of herbivory in ants. *Proceedings of the National Academy of Sciences* **106**, 21236–21241 (2009).

[22] Ferrari, J., West, J. A., Via, S. & Godfray, H. C. J. Population genetic structure and secondary symbionts in host-associated populations of the pea aphid complex. *Evolution* **66**, 375–390 (2012).

[23] Ferrari, J., Scarborough, C. L. & Godfray, H. C. J. Genetic variation in the effect of a facultative symbiont on host-plant use by pea aphids. *Oecologia* **153**, 323–329 (2007).

[24] McLean, A., Van Asch, M., Ferrari, J. & Godfray, H. Effects of bacterial secondary symbionts on host plant use in pea aphids. *Proceedings of the Royal Society of London B: Biological Sciences* **278**, 760–766 (2011).

[25] Ruokolainen, L., Ikonen, S., Makkonen, H. & Hanski, I. Larval growth rate is associated with the composition of the gut microbiota in the Glanville fritillary butterfly. *Oecologia* **181**, 1–9 (2016). URL "http://dx.doi.org/10.1007/s00442-016-3603-8. DOI 10.1007/s00442-016-3603-8.

[26] Nice, C. C., Fordyce, J. A., Shapiro, A. M. & Ffrench-Constant, R. Lack of evidence for reproductive isolation among ecologically specialised lycaenid butterflies. *Ecological Entomology* **27**, 702–712 (2002). URL http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2311.2002.00458.x/full. DOI 10.1046/j.1365-2311.2002.00458.x.

[27] Scholl, C. F., Nice, C. C., Fordyce, J. a., Gompert, Z. & Forister, M. L. Larval Performance in the Context of Ecological Diversification and Speciation in Lycaeides Butterflies. *International Journal of Ecology* **2012**, 1–13 (2012). URL http://www.hindawi.com/journals/ijecol/2012/242154/. DOI 10.1155/2012/242154.

[28] Michaud, R., Lehman, W. F. & Rumbaugh, M. D. World distribution and historical developments. In Hanson, A. A., Barnes, D. K. & Hill, R. R. (eds.) *Alfalfa and Alfalfa Improvement*, vol. 29, chap. World distribution and historical developments (Madison, 1988).

[29] Forister, M. L., Nice, C. C., Fordyce, J. a. & Gompert, Z. Host range evolution is not driven by the optimization of larval performance: the case of Lycaeides melissa (Lepidoptera: Lycaenidae) and the colonization of alfalfa. *Oecologia* **160**, 551–61 (2009). URL `http://www.ncbi.nlm.nih.gov/pubmed/19271241`. DOI 10.1007/s00442-009-1310-4.

[30] Forister, M. L. & Wilson, J. S. The population ecology of novel plant-herbivore interactions. *Oikos* **122**, 657–666 (2013). DOI 10.1111/j.1600-0706.2013.00251.x.

[31] Gompert, Z. *et al.* The evolution of novel host use is unlikely to be constrained by trade-offs or a lack of genetic variation. *Molecular Ecology* **24**, 2777–2793 (2015).

[32] Harrison, J. G. *et al.* The many dimensions of diet breadth: Phytochemical, genetic, behavioral, and physiological perspectives on the interaction between a native herbivore and an exotic host. *PloS one* **11**, e0147971 (2016).

[33] Harrison, J. G., Urruty, D. M. & Forister, M. L. An exploration of the fungal assemblage in each life history stage of the butterfly, *Lycaeides melissa* (Lycaenidae), as well as its host plant *Astragalus canadensis* (Fabaceae). *Fungal Ecology* **22**, 10–16 (2016).

[34] Muegge, B. D. *et al.* Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science (New York, N.Y.)* **332**, 970–4 (2011). URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3303602{\&}tool=pmcentrez{\&}rendertype=abstract`. DOI 10.1126/science.1198719.

[35] Gompert, Z. *et al.* Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology* **23**, 4555–4573 (2014).

[36] Gompert, Z., Forister, M. L., Fordyce, J. A. & Nice, C. C. Widespread mito-nuclear discordance with evidence for introgressive hybridization and selective sweeps in *Lycaeides*. *Molecular ecology* **17**, 5231–5244 (2008).

[37] Whitaker, M., Pierce, N., Salzman, S., Kaltenpoth, M. & Sanders, J. Microbial communities of Lycaenid butterflies do not correlate with larval diet. *Frontiers in Microbiology* **7**, 1920 (2016).

[38] Joshi, A. & Thompson, J. N. Trade-offs and the evolution of host specialization. *Evolutionary Ecology* **9**, 82–92 (1995).

[39] Forister, M., Dyer, L., Singer, M., Stireman, J. & Lill, J. Revisiting the evolution of ecological specialization, with emphasis on insect–plant interactions. *Ecology* **93**, 981–991 (2012).

[40] Russell, J. A. *et al.* A Veritable Menagerie of Heritable Bacteria from Ants, Butterflies, and Beyond: Broad Molecular Surveys and a Systematic Review. *PLoS ONE* **7** (2012). DOI 10.1371/journal.pone.0051027.

[41] Sugio, A., Dubreuil, G., Giron, D. & Simon, J.-C. Plant–insect interactions under bacterial influence: ecological implications and underlying mechanisms. *Journal of Experimental Botany* eru435 (2014).

[42] Coon, K. L., Vogel, K. J., Brown, M. R. & Strand, M. R. Mosquitoes rely on their gut microbiota for development. *Molecular Ecology* **23**, 2727–2739 (2014).

[43] Hammer, T. J., McMillan, W. O. & Fierer, N. Metamorphosis of a butterfly-associated bacterial community. *PloS one* **9**, e86995 (2014). URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3900687{\&}tool=pmcentrez{\&}rendertype=abstract`. DOI 10.1371/journal.pone.0086995.

[44] Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12**, 87 (2014).

[45] Robinson, C. J., Schloss, P., Ramos, Y., Raffa, K. & Handelsman, J. Robustness of the bacterial community in the cabbage white butterfly larval midgut. *Microbial Ecology* **59**, 199–211 (2010). DOI 10.1007/s00248-009-9595-8.

[46] Engel, P. & Moran, N. A. The gut microbiota of insects - diversity in structure and function. *FEMS Microbiology Reviews* **37**, 699–735 (2013). DOI 10.1111/1574-6976.12025.

[47] Chouaia, B. *et al.* Molecular evidence for multiple infections as revealed by typing of asaia bacterial symbionts of four mosquito species. *Applied and Environmental Microbiology* **76**, 7444–7450 (2010). DOI 10.1128/AEM.01747-10.

[48] Crotti, E. *et al.* Acetic acid bacteria, newly emerging symbionts of insects. *Applied and Environmental Microbiology* **76**, 6963–6970 (2010). DOI 10.1128/AEM.01336-10.

[49] Tang, X. *et al.* Complexity and variability of gut commensal microbiota in polyphagous lepidopteran larvae. *PloS ONE* **7**, e36978 (2012).

[50] Cox, C. R. & Gilmore, M. S. Native microbial colonization of drosophila melanogaster and its use as a model of enterococcus faecalis pathogenesis. *Infection and immunity* **75**, 1565–1576 (2007).

[51] Lehman, R. M., Lundgren, J. G. & Petzke, L. M. Bacterial communities associated with the digestive tract of the predatory ground beetle, poecilus chalcites, and their modification by laboratory rearing and antibiotic treatment. *Microbial ecology* **57**, 349–358 (2009).

[52] Martinson, V. G., Moy, J. & Moran, N. A. Establishment of characteristic gut bacteria during development of the honeybee worker. *Applied and Environmental Microbiology* **78**, 2830–2840 (2012). DOI 10.1128/AEM.07810-11.

[53] Chandler, J. A., Lang, J., Bhatnagar, S., Eisen, J. A. & Kopp, A. Bacterial communities of diverse Drosophila species: Ecological context of a host-microbe model system. *PLoS Genetics* **7** (2011). DOI 10.1371/journal.pgen.1002272.

[54] Blum, J. E., Fischer, C. N., Miles, J. & Handelsman, J. Frequent replenishment sustains the beneficial microbiome of *Drosophila melanogaster*. *MBio* **4**, e00860–13 (2013).

[55] Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* **13**, 260–270 (2012).

[56] Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: an integrative view. *Cell* **148**, 1258–1270 (2012).

[57] Greenblum, S., Turnbaugh, P. J. & Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences* **109**, 594–599 (2012).

[58] Bajaj, J. S. *et al.* Altered profile of human gut microbiome is associated with cirrhosis and its complications. *Journal of hepatology* **60**, 940–947 (2014).

[59] Gompert, Z. *et al.* Geographically multifarious phenotypic divergence during speciation. *Ecology and Evolution* **3**, 595–613 (2013).

[60] Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* **6**, 1621–1624 (2012). URL `http://dx.doi.org/10.1038/ismej.2012.8`. DOI 10.1038/ismej.2012.8. `arXiv:1408.1149`.

[61] Caporaso, J. G. *et al.* Qiime allows analysis of high-throughput community sequencing data. *Nature methods* **7**, 335–336 (2010).

[62] Edgar, R. C. Search and clustering orders of magnitude faster than blast. *Bioinformatics* **26**, 2460–2461 (2010).

[63] DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and Environmental Microbiology* **72**, 5069–5072 (2006).

[64] McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal* **6**, 610–8 (2012). URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3280142{\&}tool=pmcentrez{\&}rendertype=abstract`. DOI 10.1038/ismej.2011.139.

[65] Ghyselinck, J., Pfeiffer, S., Heylen, K., Sessitsch, A. & De Vos, P. The effect of primer choice and short read sequences on the outcome of 16s rrna gene based diversity studies. *PLoS One* **8**, e71360 (2013).

[66] Hanshew, A. S., Mason, C. J., Raffa, K. F. & Currie, C. R. Minimization of chloroplast contamination in 16s rrna gene pyrosequencing of insect herbivore bacterial communities. *Journal of microbiological methods* **95**, 149–155 (2013).

[67] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2016). URL `https://www.R-project.org/`.

[68] Legendre, P. & Gallagher, E. D. Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**, 271–280 (2001).

[69] Paliy, O. & Shankar, V. Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology* **25**, 1032–1057 (2016). URL `http://dx.doi.org/10.1111/mec.13536`. DOI 10.1111/mec.13536.

[70] Oksanen, J. *et al.* The vegan package. *Community ecology package* **10** (2007).

[71] Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

[72] Jost, L. Entropy and diversity. *Oikos* **113**, 363–375 (2006).

[73] Tuomisto, H. A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia* **164**, 853–860 (2010).

[74] Haegeman, B. *et al.* Robust estimation of microbial diversity in theory and in practice. *The ISME Journal* **7**, 1092–1101 (2013).

[75] Liaw, A. & Wiener, M. Classification and regression by randomforest. *R News* **2**, 18–22 (2002). URL `http://CRAN.R-project.org/doc/Rnews/`.

[76] Pietri, J. E., DeBruhl, H. & Sullivan, W. The rich somatic life of wolbachia. *MicrobiologyOpen* (2016).

[77] Ramette, A. Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology* **62**, 142–160 (2007).

[78] Anderson, M. J. *et al.* Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology letters* **14**, 19–28 (2011).

[79] Legendre, P. & Legendre, L. F. *Numerical ecology*, vol. 24 (Elsevier, 2012).

[80] Warnes, G. R. *et al.* gplots: Various r programming tools for plotting data. *R package version* **2** (2009).

[81] Lunn, D., Spiegelhalter, D., Thomas, A. & Best, N. The bugs project: Evolution, critique and future directions. *Statistics in medicine* **28**, 3049–3067 (2009).

Tables and Figures

Table 4.1. Bayesian estimates of microbial community composition for different sample types. Posterior medians ('pm') and 95% ETPIs are provided for the mean (μ) and standard deviation (σ) of principal coordinate (PCO) and principal component (PC) scores.

Parameter	PCO1		PCO2		PC1		PC2	
	pm	ETPIs	pm	ETPIs	pm	ETPIs	pm	ETPIs
μ_{Plant}	0.057	-0.092, 0.209	-0.056	-0.197, 0.091	0.062	-0.239, 0.358	0.013	-0.213, 0.238
μ_{Frass}	0.074	-0.011, 0.162	0.063	0.018, 0.107	0.121	-0.011, 0.250	0.032	-0.077, 0.135
μ_{Larvae}	-0.356	-0.471, -0.245	-0.205	-0.407, 0.000	-0.549	-0.679, -0.414	-0.139	-0.408, 0.137
σ_{Plant}	0.205	0.133, 0.375	0.202	0.130, 0.373	0.401	0.260, 0.724	0.310	0.200, 0.568
σ_{Frass}	0.278	0.225, 0.351	0.143	0.116, 0.181	0.423	0.344, 0.540	0.331	0.268, 0.420
σ_{Larvae}	0.163	0.108, 0.291	0.299	0.198, 0.522	0.194	0.128, 0.337	0.389	0.258, 0.689

Fig. 4.1. **OTU richness and diversity.** Scatterplots show the (A) number of OTUs and (B) effective number of species (^2D) for each sample as a function of sequencing depth prior to rarefaction but after removing chloroplast, mitochondrial and *Wolbachia* sequences. Colors and symbols denote different sample types and vertical lines show the rarefaction cutoff for each sample for downstream analysis (1311 sequences).

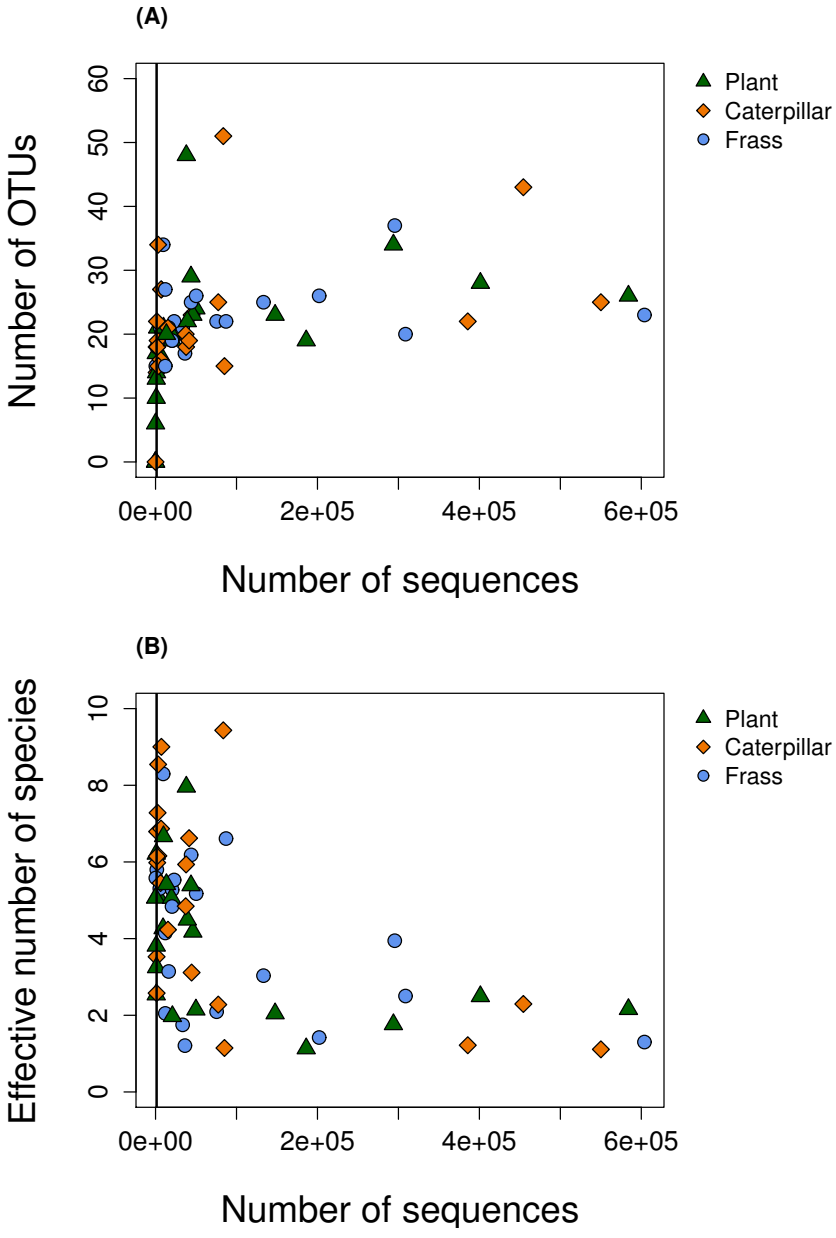


Fig. 4.2. **Relative bacterial abundances.** Relative abundances of the major microbial taxa in the plant, larval (caterpillar) and frass samples, calculated from operational taxonomic unit (OTU) counts. Samples are sorted according to sample type (plant, whole caterpillar or frass), population, plant and larval age. Sample abbreviation are: En = endophytes; Ep = epiphytes; Hardware Ranch = HWR, Bonneville Shoreline Trail = BST; Me = (*M. sativa*), and Lu = (*L. argenteus*). Numbers (15, 20 or 25) indicate caterpillar age. In the legend, OTU are identified as class (order).

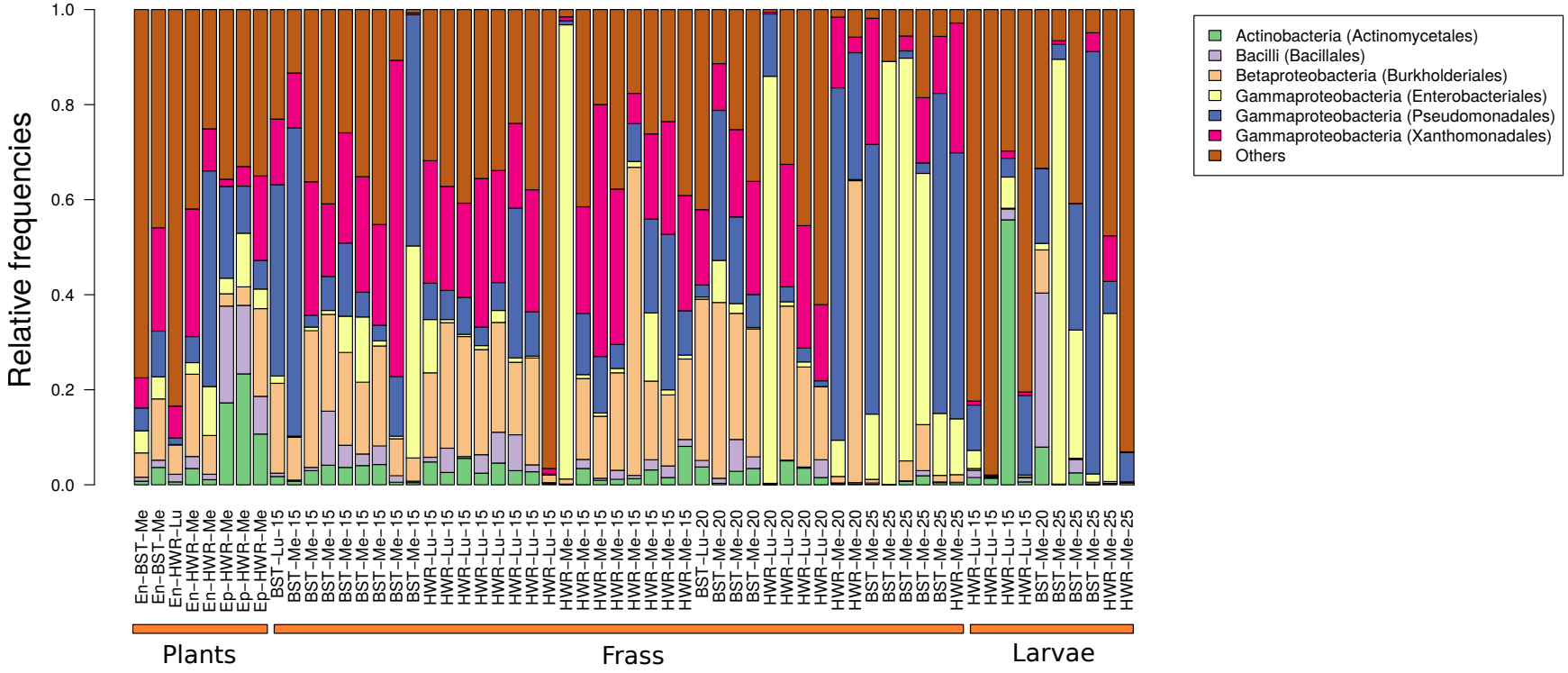


Fig. 4.3. **Principal component analysis.** Scatterplots show an ordination of microbial communities from chord-transformed relative abundance data for (A) all samples, or (B) frass and larvae. Colors and symbols denote different treatments and sample types (see legends).

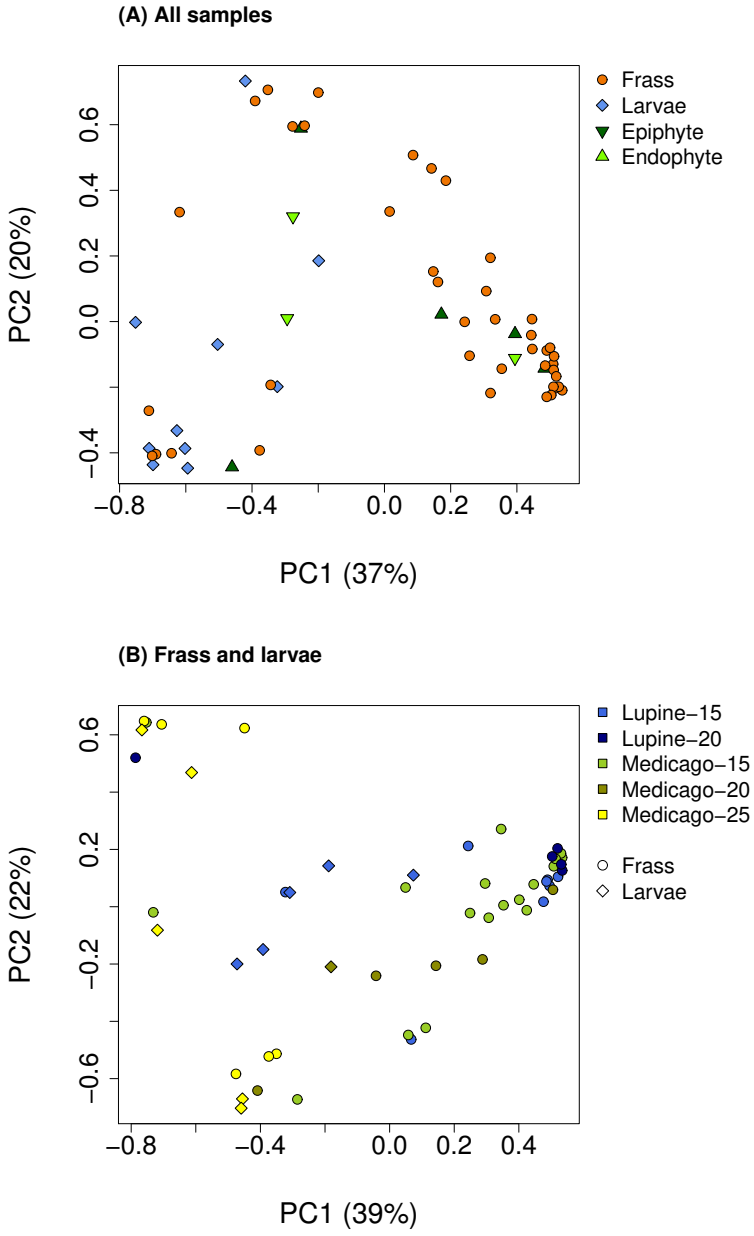


Fig. 4.4. **Hierarchical cluster analysis.** Heatmap of Bray-Curtis dissimilarities between samples and corresponding dendrograms based on bacterial OTU abundances. Each row and column represents a sample. Cell colors indicate dissimilarity values between row and column microbial communities (red = greater similarity and yellow = less similarity). The dendrogram groups samples by heirarchical clustering based on microbial community similarity. Sample abbreviations: F = frass, L = larvae, HWR = Hardware Ranch, BST = Bonneville Shoreline Trail, Me = *M. sativa*, and Lu = *L. argenteus*. Numbers with F and L indicate sample IDs, whereas the final number in each ID gives the caterpillar age (15, 20, or 25). Color bars above the heatmap indicate the age of samples, (green = 15 days, blue = 20 days, purple = 25 days). The heatmap and dendrogram show that microbiomes from different sample types, different age caterpillars and different treatments do not form distinct groups or sub-cluster, but that there is a tendency for sets of similar samples (i.e., samples from the same age caterpillar) to be more similar and cluster together.

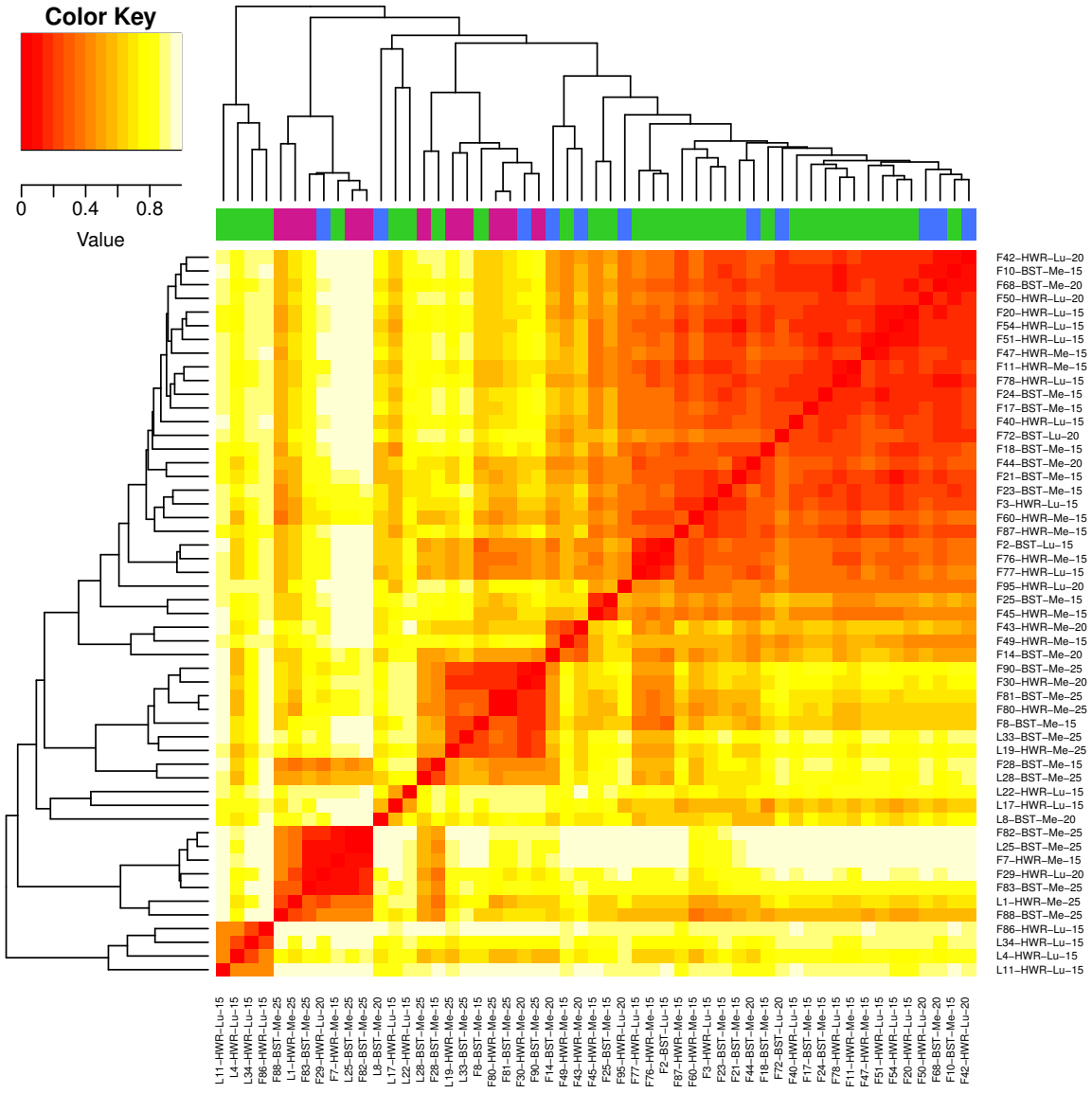


Fig. 4.5. **Phylotype diversity.** True phylotype diversity, that is Hill's effective species number with $q = 2$, is shown for all plant, caterpillar and frass samples.

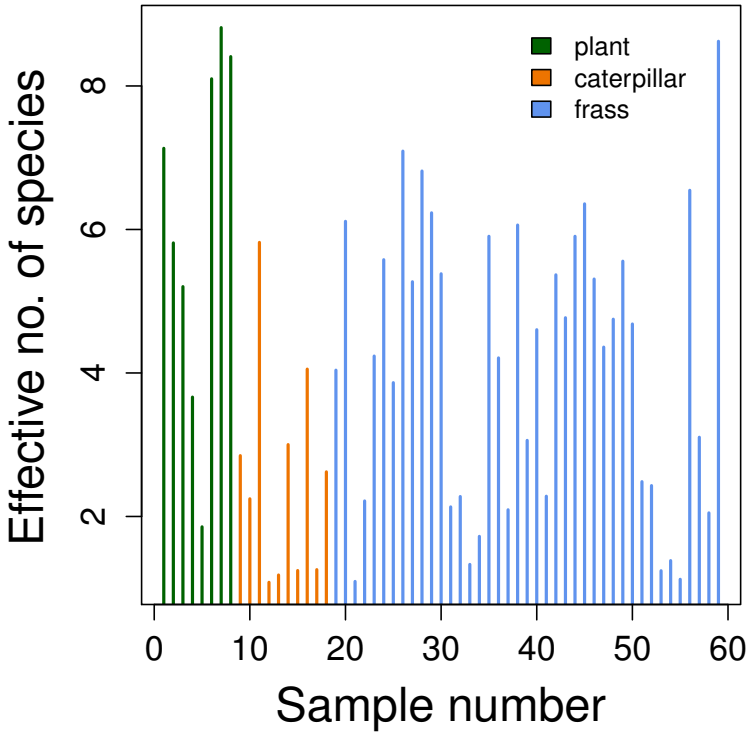
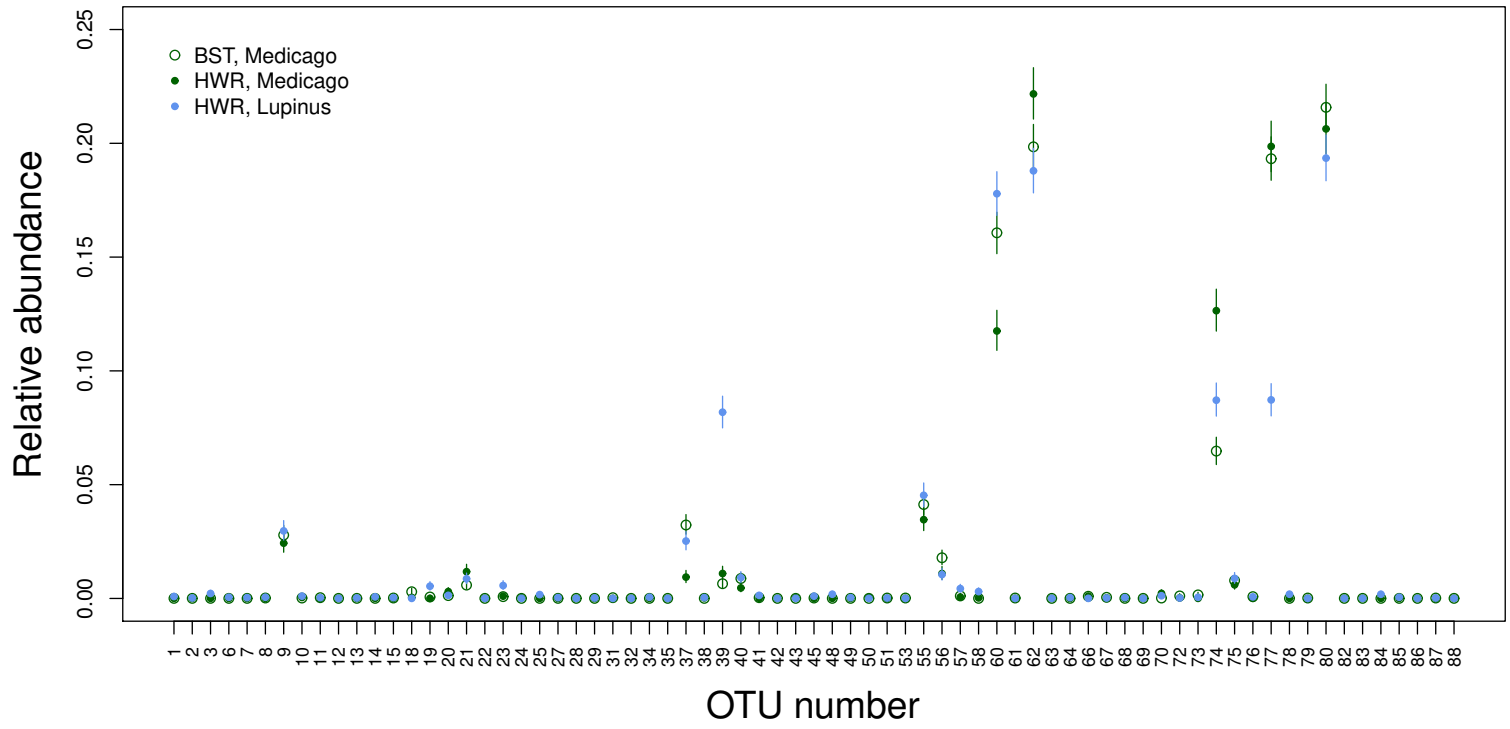


Fig. 4.6. **Microbe relative abundance from frass samples.** Points and vertical bars denote Bayesian point estimates (posterior medians) and 95% ETPIs for the relative abundance of different microbial OTUs in 15 and 20 day frass samples. Colors and symbols denote samples from different plant (*L. argenteus* or *M. sativa*) and population (BST or HWR) treatments. Estimates are from a Bayesian multinomial-Dirichlet model. Low sample sizes precluded meaningful estimates for BST on *L. argenteus*. OTU numbers are defined in Table 4.10.



Supplemental tables and figures

Table 4.2. S1 PC1 and PC2 loadings for top microbial phylotypes in epiphytes, endophytes, frass and larvae after removing chloroplast and mitochondria and following chord transformation of the relative abundance data.

Microbe class	**PC1 rotations for epiphytes, endophytes, frass and larvae**
Gammaproteobacteria (Xanthomonadales)	0.49
Betaproteobacteria (Burkholderiales)	0.49
Alphaproteobacteria (Sphingomonadales)	0.47
Alphaproteobacteria (Caulobacterales)	0.12
Bacilli (Bacillales)	0.02
Microbe class	**PC2 rotations for epiphytes, endophytes, frass and larvae**
Gammaproteobacteria (Pseudomonadales)	0.89
Bacilli (Bacillales)	0.04
Alphaproteobacteria (Rhodobacterales)	0.01
Gammaproteobacteria (Xanthomonadales)	0.006
Sphingobacteriia (Sphingobacteriales)	0.005

Table 4.3. S2 PC1 and PC2 loadings for top microbial phylotypes in frass and larvae after removing *Wolbachia* and following chord transformation of the relative abundance data.

Microbe class	**PC1 rotations for frass and larvae**
Betaproteobacteria (Burkholderiales)	0.48
Alphaproteobacteria (Sphingomonadales)	0.46
Gammaproteobacteria (Xanthomonadales)	0.45
Alphaproteobacteria (Caulobacterales)	0.12
Actinobacteria (Actinomycetales)	0.05
Microbe class	**PC2 rotations for frass and larvae**
Gammaproteobacteria (Enterobacteriales)	0.56
Alphaproteobacteria (Sphingomonadales)	0.15
Betaproteobacteria (Burkholderiales)	0.05
Actinobacteria (Actinomycetales)	0.04
Alphaproteobacteria (Caulobacterales)	0.03

Table 4.4. S3 Top five microbial phylotypes in frass, larvae and plants based on importance assigned by Random Forest (RF) GINI Indexes for class sample type.

Microbe class	**RF GINI Index - Sample type**
Alphaproteobacteria (Rickettsiales)	4.89
Betaproteobacteria (Burkholderiales)	2.46
Gammaproteobacteria (Xanthomonadales)	1.85
Alphaproteobacteria (Caulobacterales)	1.83
Alphaproteobacteria (Sphingomonadales)	1.17
Bacilli (Bacillales)	1.08

Table 4.5. S4 Model comparison for the association of microbial community with larval weight ($\bar{D}$ = mean deviance, pD = effective number of parameters, Δ DIC = difference in DIC compared to the best model).

Model	$\bar{D}$	**pD**	**Δ DIC**
null model	69.0	5.4	0
PC1	70.2	6.5	2.2
PC1 × plant	68.7	7.6	2.0
PC2	70.2	6.5	2.4
PC2 × plant	70.7	7.6	4.0
PCO1	69.8	6.5	2.0
PCO1 × plant	68.4	7.6	1.8
PCO2	70.2	6.5	2.3
PCO2 × plant	70.2	7.6	3.5
^2D	70.2	6.5	2.4
^2D × plant	71.4	7.6	4.1

Table 4.6. S5 Random Forest confusion matrix for correct assignment of plant, frass and larvae microbial communities to sample type (Out of Bag (OOB) estimate of error rate = 16.67%). Rows indicate actual class and columns indicate predicted class.

Type	**Frass**	**Larvae**	**Plant**	**error**
Frass	41	0	0	0.0
Larvae	2	8	0	0.20
Plant	8	0	1	0.88

Table 4.7. S6 Random Forest Out of Bag (OOB) estimate of error rate for classes after removing *Wolbachia*.

Class	**OOB error rate**
Type	16.98%
Age	32.08%
Age (15 + 20 days)	26.42%
Plant	39.62%
Population	45.28%

Table 4.8. S7 Confusion matrixes for Random Forest assignment of samples based on microbial community for frass and larvae after removing *Wolbachia*. Results are shown for age (with or without combining 15 and 20 days), sample type (frass or whole caterpillar), host plant species (alflafa or lupine) and populations (BST or HWR).

Age	**15days**	**20 days**	**25 days**	**Error**
15	29	0	2	0.06
20	9	0	2	1
25	4	0	7	0.36
Age	**15 + 20 days**	**25 days**	**Error**	
15 + 20	38	4	0.09	
25	10	1	0.91	
Type	**Frass**	**Larvae**	**Error**	
Frass	41	1	0.02	
Larvae	8	3	0.72	
Plant	**Lupine**	**Alfalfa**	**Error**	
Lupine	3	15	0.83	
Alfalfa	6	29	0.17	
Population	**BST**	**HWR**	**Error**	
BST	8	15	0.65	
HWR	9	21	0.30	

Table 4.9. S8 Top five microbial phylotypes in frass and larvae based on importance assigned by Random Forest (RF) GINI Indexes for classes sample type, age and plant.

Microbe class	**RF GINI Index - Sample type**
Betaproteobacteria (Burkholderiales)	1.96
Gammaproteobacteria (Xanthomonadales)	1.68
Alphaproteobacteria (Caulobacterales)	1.65
Alphaproteobacteria (Sphingomonadales)	1.11
Bacilli (Lactobacillales)	1.07
Microbe class	**RF GINI Index - Age (15 + 20 vs 25)**
Gammaproteobacteria (Enterobacteriales)	2.39
Betaproteobacteria (Burkholderiales)	1.88
Gammaproteobacteria (Pseudomonadales)	1.49
Alphaproteobacteria (Sphingomonadales)	1.19
Alphaproteobacteria (Caulobacterales)	1.12
Microbe class	**RF GINI Index - Age (15 vs 20 vs 25)**
Betaproteobacteria (Burkholderiales)	2.97
Gammaproteobacteria (Enterobacteriales)	2.77
Alphaproteobacteria (Sphingomonadales)	2.03
Gammaproteobacteria (Pseudomonadales)	1.93
Bacilli (Bacillales)	1.90
Microbe class	**RF GINI Index - Plant**
Gammaproteobacteria (Oceanospirillales)	2.51
Gammaproteobacteria (Xanthomonadales)	1.81
Gammaproteobacteria (Pseudomonadales)	1.75
Alphaproteobacteria (Sphingomonadales)	1.69
Betaproteobacteria (Burkholderiales)	1.54
Microbe class	**RF GINI Index - Population**
Gammaproteobacteria (Enterobacteriales)	2.49
Gammaproteobacteria (Xanthomonadales)	1.80
Gammaproteobacteria (Pseudomonadales)	1.58
Actinobacteria (Actinomycetales)	1.54
Alphaproteobacteria (Sphingomonadales)	1.44

Table 4.10. S9 Microbial phylotypes found across frass, larvae, endophyte and epiphyte samples after removing chloroplast and mitochondria. In some cases microbes lack formal taxonomic IDs at lower levels (e.g., Class and Order).

OTU Number	Mean relative abundance	Phylum	Class	Order
1	7.00E-5	Crenarchaeota	Thaumarchaeota	Nitrososphaerales
2	0.00003	Euryarchaeota	Halobacteria	Halobacteriales
3	0.0004	Acidobacteria	Acidobacteria	Iii1.15
4	4.00E-5	Acidobacteria	Acidobacteriia	Acidobacteriales
5	0.00002	Acidobacteria	S035	-
6	0.0002	Acidobacteria	Chloracidobacteria	RB41
7	1.00E-5	Acidobacteria	iii1.8	DS.18
8	0.0002	Actinobacteria	Acidimicrobiia	Acidimicrobiales
9	0.03	Actinobacteria	Actinobacteria	Actinomycetales
10	0.0005	Actinobacteria	Actinobacteria	Bifidobacteriales
11	0.0003	Actinobacteria	Coriobacteriia	Coriobacteriales
12	1.00E-5	Actinobacteria	Nitriliruptoria	Nitriliruptorales
13	6.00E-5	Actinobacteria	Rubrobacteria	Rubrobacterales
14	4.00E-5	Actinobacteria	Thermoleophilia	Gaiellales
15	0.0002	Actinobacteria	Thermoleophilia	Solirubrobacterales
16	2.00E-5	Bacteroidetes	-	-
17	3.00E-5	Bacteroidetes	BME43	-
18	0.001	Bacteroidetes	Bacteroidia	Bacteroidales
19	0.001	Bacteroidetes	Cytophagia	Cytophagales
20	0.002	Bacteroidetes	Flavobacteriia	Flavobacteriales
21	0.004	Bacteroidetes	Sphingobacteriia	Sphingobacteriales
22	2.00E-5	Bacteroidetes	Rhodothermi	Rhodothermales
23	0.001	Bacteroidetes	Saprospirae	Saprospirales

Continued on next page

Table 4.10 – continued from previous page

OTU Number	Mean relative abundance	Phylum	Class	Order
24	9.00E-6	Chlamydiae	Chlamydiia	Chlamydiales
25	0.0003	Chloroflexi	Anaerolineae	SBR1031
26	1.00E-5	Chloroflexi	Chloroflexi	AKIW781
27	4.00E-5	Chloroflexi	Ellin6529	-
28	7.00E-5	Chloroflexi	Gitt.GS.136	-
29	4.71E-5	Chloroflexi	TK10	AKYG885
30	0.0001	Chloroflexi	Thermomicrobia	-
31	0.0001	Chloroflexi	Thermomicrobia	JG30.KF.CM45
32	0.00008	Cyanobacteria	4C0d.2.o	MLE1.12
33	1.00E-5	Cyanobacteria	4C0d.2	YS2
34	4.00E-5	Cyanobacteria	Oscillatoriophycideae	Oscillatoriales
35	0.0001	FBP	-	-
36	1.00E-6	Firmicutes	Bacilli	-
37	0.02	Firmicutes	Bacilli	Bacillales
38	0.0002	Firmicutes	Bacilli	Gemellales
39	0.02	Firmicutes	Bacilli	Lactobacillales
40	0.006	Firmicutes	Clostridia	Clostridiales
41	0.0002	Firmicutes	Clostridia	Thermoanaerobacterales
42	0.002	Firmicutes	Erysipelotrichi	Erysipelotrichales
43	0.0003	Fusobacteria	Fusobacteriia	Fusobacteriales
44	2.00E-5	Gemmatimonadetes	Gemm.1	-
45	0.0002	Gemmatimonadetes	Gemm.3	-
46	1.00E-6	Gemmatimonadetes	Gemm.5	-
47	1.00E-5	Gemmatimonadetes	Gemmatimonadetes	-
48	0.0003	Gemmatimonadetes	Gemmatimonadetes	Gemmatimonadales

Continued on next page

Table 4.10 – continued from previous page

OTU Number	Mean relative abundance	Phylum	Class	Order
49	3.00E-5	Nitrospirae	Nitrospira	Nitrospirales
50	0.00006	OD1	ZB2	-
51	0.0001	Planctomycetes	Phycisphaerae	WD2101
52	5.00E-5	Planctomycetes	Planctomycetia	Gemmatales
53	2.00E-5	Planctomycetes	Planctomycetia	Pirellulales
54	7.00E-5	Planctomycetes	Planctomycetia	Planctomycetales
55	0.02	Proteobacteria	Alphaproteobacteria	Caulobacterales
56	0.01	Proteobacteria	Alphaproteobacteria	Rhizobiales
57	0.003	Proteobacteria	Alphaproteobacteria	Rhodobacterales
58	0.0008	Proteobacteria	Alphaproteobacteria	Rhodospirillales
59	0.05	Proteobacteria	Alphaproteobacteria	Rickettsiales
60	0.08	Proteobacteria	Alphaproteobacteria	Sphingomonadales
61	0.00003	Proteobacteria	Betaproteobacteria	-
62	0.1	Proteobacteria	Betaproteobacteria	Burkholderiales
63	0.0001690141	Proteobacteria	Betaproteobacteria	Ellin6067
64	1.00E-5	Proteobacteria	Betaproteobacteria	MND1
65	0.0001	Proteobacteria	Betaproteobacteria	Methylophilales
66	0.0005	Proteobacteria	Betaproteobacteria	Neisseriales
67	0.0003	Proteobacteria	Betaproteobacteria	Rhodocyclales
68	0.0001	Proteobacteria	Betaproteobacteria	SC.I.84
69	8.00E-5	Proteobacteria	Deltaproteobacteria	Bdellovibrionales
70	0.0007	Proteobacteria	Deltaproteobacteria	Myxococcales
71	5.00E-5	Proteobacteria	Epsilonproteobacteria	Campylobacterales
72	0.0002	Proteobacteria	Gammaproteobacteria	Aeromonadales
73	0.0006	Proteobacteria	Gammaproteobacteria	Alteromonadales

Continued on next page

Table 4.10 – continued from previous page

OTU Number	Mean relative abundance	Phylum	Class	Order
74	0.1	Proteobacteria	Gammaproteobacteria	Enterobacteriales
75	0.004	Proteobacteria	Gammaproteobacteria	Oceanospirillales
76	0.007	Proteobacteria	Gammaproteobacteria	Pasteurellales
77	0.1	Proteobacteria	Gammaproteobacteria	Pseudomonadales
78	0.0003	Proteobacteria	Gammaproteobacteria	Thiotrichales
79	2.00E-5	Proteobacteria	Gammaproteobacteria	Vibrionales
80	0.1	Proteobacteria	Gammaproteobacteria	Xanthomonadales
81	1.00E-5	Synergistetes	Synergistia	Synergistales
82	0.0003	TM7	SC3	-
83	0.00007	TM7	TM7.3	I025
84	0.0002	Verrucomicrobia	Opitutae	Opitutales
85	4.00E-5	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales
86	0.00006	Verrucomicrobia	Spartobacteria	Chthoniobacterales
87	0.0001	Thermi	Deinococci	Deinococcales
88	5.00E-5	Thermi	Deinococci	Thermales

Fig. 4.7. S1 **Principal coordinate analysis.** Scatterplots show an ordination of microbial communities based on Bray-Curtis community dissimilarities for (A) all samples, or (B) frass and larvae. Colors and symbols denote different treatments and sample types (see legends). Caterpillar, frass and plant microbial communities overlapped in principal coordinate (PCO) space, but sample types differed in their average PCO scores and degree of variability (Table 4.1). Most notably, average caterpillar communities differed from frass and plant communities with respect to PCO1 scores (Bayesian posterior prob. [pp] $\mu_{\text{Larvae}} > \mu_{\text{Frass}} > 0.99$; pp $\mu_{\text{Larvae}} > \mu_{\text{Plant}} > 0.99$), and caterpillar and frass communities differed with respect to PCO2 scores (pp $\mu_{\text{Larvae}} < \mu_{\text{Frass}} = 0.99$). Frass and plant microbe communities were more similar.

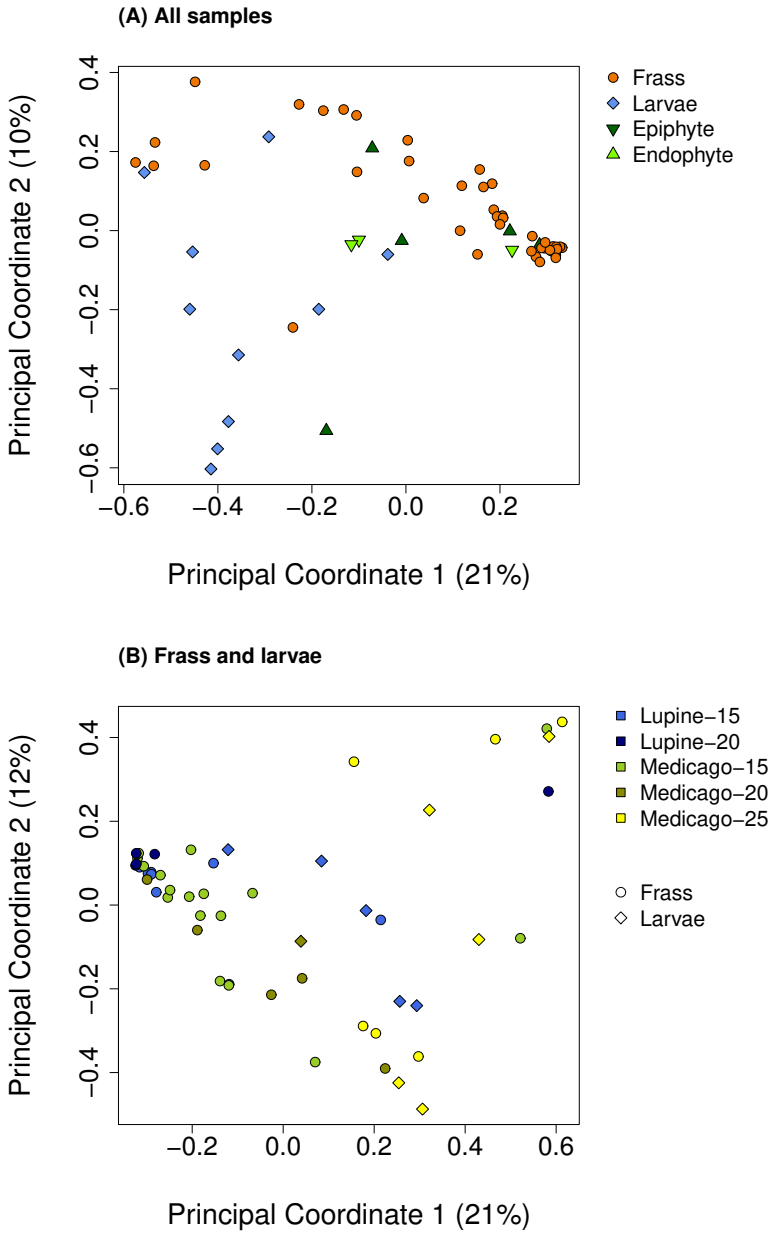
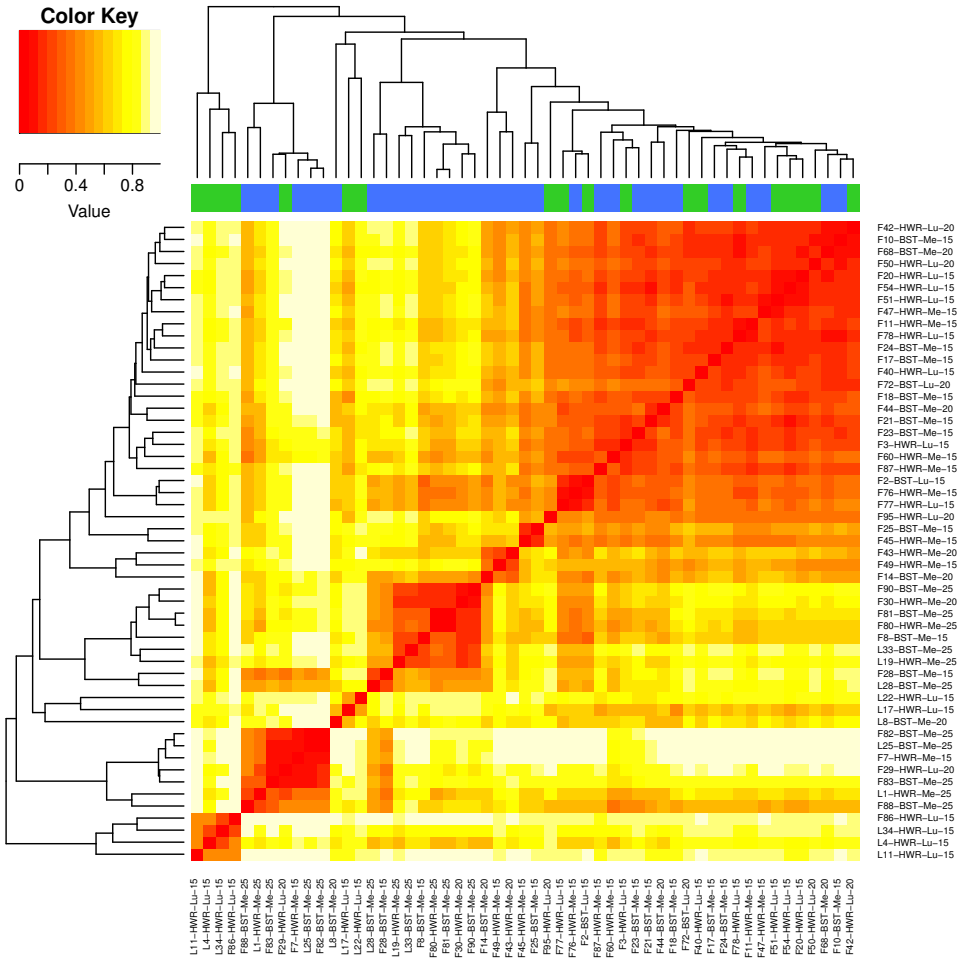


Fig. 4.8. S2 **Hierarchical Cluster Analysis.** Heatmap of Bray-Curtis dissimilarities between samples and corresponding dendrograms based on bacterial OTU abundances. Each row and column represents a sample. Cell colors indicate dissimilarity values between row and column microbial communities (red = greater similarity and yellow = less similarity). The dendrogram groups samples by heirarchical clustering based on microbial community similarity. Sample abbreviations: F = frass, L = larvae, HWR = Hardware Ranch, BST = Bonneville Shoreline Trail, Me = *M. sativa*, and Lu = *L. argenteus*. Numbers with F and L indicate sample IDs, whereas the final number in each ID gives the caterpillar age (15, 20, or 25). Color bars above the heatmap indicate the host plant each sample was reared on (green = *L. argenteus*, blue = *M. sativa*).. The heatmap and dendrogram show that microbiomes from different sample types, different age caterpillars and different treatments do not form distinct groups or clades, but that there is a tendency for sets of similar samples (i.e., samples from the same age caterpillar) to be more similar and cluster together. Note, this figure is identical to Fig. 4.4 except that colored bars denote plant rather than caterpillar age.



CHAPTER 5

SUMMARY AND CONCLUSIONS

Repeatable evolutionary changes have been demonstrated across various taxa and for several traits. However, few instances exist in which the predictability of genomic differentiation among ecologically divergent populations has been quantified. In addition, predictability of evolutionary changes can be studied in different contexts. Quantifying estimates of predictability across different contextual approaches can help investigate the contribution of different factors (for e.g. demography, selection or genetics) to evolutionary predictability. In this dissertation, I examined if evolution is predictable in *Lycaeides* butterflies. I used large scale genomic sequencing and statistical analyses to identify, quantify and infer repeatable patterns of adaptive evolution and genomic introgression in these butterflies. My results show that I can quantify predictability to understand how predictable evolution is under different circumstances and to what extent. Overall, I show that genomic changes accompanying an adaptive trait as well as the process of hybridization, are indeed predictable. However, the estimates of the degree of predictability are highly dependent on the context in which I quantify evolutionary predictability such as specific comparisons (among natural populations vs. between natural and experimental populations), spatial scale (across entire geographic range vs. pairs of populations), temporal scale (ancient vs. contemporary hybrids), and regions of the genome (autosomes vs. sex chromosomes). I summarize the results from different contextual comparisons for each chapter below.

In Chapter 2, I generated and analyzed genomic data from *Lycaeides* populations across a natural hybrid zone to ask if I can predict genomic changes (and to what extent) in a contemporary hybrid zone from a ancient hybrid zone. The results show that I can predict the overlap of genomic regions across ancient and contemporary hybrid zones and the degree of predictability is high when considering different ranges of empirical quantiles. Overall, the results demonstrate substantial repeatability in regions underlying restricted introgression between ancient and contemporary hybrids in *Lycaeides* and provide evidence that natural selection can indeed shape evolutionary predictability on a temporal scale. I first delineate genomic regions with excess local ancestry frequencies for one of the parental species in ancient hybrids which formed around 15,000 years ago. I then delineate genomic regions

exhibiting variable genomic clines in contemporary hybrids formed more recently (around 200 years ago) and identify genomic regions which experience restricted introgression and underlie speciation. Quantifying predictability on a temporal scale by comparing genomic regions under selection between ancient and contemporary hybrids, I see a really high level of predictability (50-1X overlap between SNPs depending on the genomic regions being considered). Here, I again show that the Z-chromosome has a notable enrichment of shared SNPs which are present in significant excess. However, predictability is not restricted to Z-chromosome and shared genomic regions are spread across autosomes. Lastly, these estimates hold for comparisons among natural populations on temporal as well as geographic scale wherein the ancient hybrids populations span a wide geographic range and the contemporary hybrids inhabit a very small space.

In chapter 3, I analyzed genomic data from several *L. melissa* populations to quantify the predictability of genomic change underlying a novel host plant shift. Overall, the results show that genomic changes underlying this host shift are somewhat predictable. However, the degree of predictability varies across different contextual approaches. Here, I first show that *L. melissa* butterflies have colonized a novel host plant (alfalfa, *M. sativa*) two or more times, with little to no gene flow connecting the two clades of alfalfa-feeding butterflies. Comparisons among natural populations reveal that parallel genomic changes underlie host plant use in eastern and western *L. melissa* populations which have independently colonized alfalfa. The xfold enrichments for these comparisons range between 1.8-2.7X depending on the quantiles being considered. Comparisons among source populations from the experiments and their nearest opposite-host populations reveal a significant increase in predictability (upto 4.5X excess overlap between SNPs) suggesting that geographic scale and habitat heterogeneity matters in analyses of evolutionary predictability and repeatability. Comparisons between natural and experimental populations revealed that I could partially predict genomic patterns of host use in nature from a controlled performance experiment but the degree of predictability through this comparison was consistently lower than observed in among natural populations (0.53-2.5X). However, the direction of the phenotypic effect of performance-associated SNPs from laboratory experiment was not predictive of the direction of allele frequency divergence among host-associated populations in nature. The results also show a

notable enrichment of shared SNPs on the Z-chromosome versus autosomes suggesting that locations of loci in the genome matter when quantifying predictability. In conclusion, these results emphasize on the use of different contexts to quantify predictability and how these quantitative estimates can help illuminate the prevalence and causes of evolutionary predictability of important adaptive traits.

When considering host-plant use in *Lycaeides melissa*, I was also interested in understanding the role of gut microbial communities in host-plant adaptation in herbivorous insects. In chapter 2, I found that there is no convincing evidence that gut microbes are crucial for host-plant adaptation in *L. melissa*. Gut microbioal communities vary with age but diet (host plant) itself does not have a significant effect on microbial diversity. Interestingly, this pattern has been revealed in other *Lepidopteran* species and these results indicate a convergent aspect to the role of microbiome in host plant use across different several *Lepidopteran* species.

Conclusion

Although, a plethora of repeatable evolution examples strongly suggest that evolution can be predictable, several studies also suggest that evolution is highly idiosyncratic and can be unpredictable. This dissertation highlights parallelism in genomic changes underlying an adaptive trait and the repeatability in patterns of genomic introgression between ancient and contemporary hybrid zones. This approach represents a step in a direction to quantify predictability of evolution. The results here highlight that predictability of evolutionary changes is not binary but is rather a continuum and is highly context dependent. Different comparisons and contexts to study evolutionary predictability highlight how spatial and temporal scale and locations in the genome matter when quantifying predictability. In addition, I see that as we use more and more contexts to systematically study evolutionary processes, we can possibly detect higher predictability of evolutionary changes. Although, many methodological challenges remain, quantitative estimates utilizing several contextual approaches can ultimately better inform our understanding of the prevalence and causes of predictability of evolution.

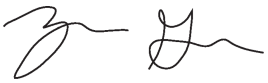
APPENDICES

APPENDIX A

Coauthor Permission Letters

I hereby give permission to Samridhi Chaturvedi to include the following articles in her thesis/dissertation, of which I am a coauthor.

1. Chaturvedi, Samridhi, Alexandre Rego, Lauren K. Lucas, and Zachariah Gompert. "Sources of Variation in the Gut Microbial Community of Lycaeides melissa Caterpillars." Scientific Reports 7 (2017).
2. Chaturvedi, Samridhi, Lauren K. Lucas, Chris C. Nice, James A. Fordyce, Matthew L. Forister, and Zachariah Gompert. "The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies." Molecular Ecology 27, no. 12 (2018): 2651-2666.

Signed: 

Date: April, 12, 2019

I hereby give permission to Samridhi Chaturvedi to include the following articles in her thesis/dissertation, of which I am a coauthor.

1. Chaturvedi, Samridhi, Alexandre Rego, Lauren K. Lucas, and Zachariah Gompert. "Sources of Variation in the Gut Microbial Community of Lycaeides melissa Caterpillars." Scientific Reports 7 (2017).
2. Chaturvedi, Samridhi, Lauren K. Lucas, Chris C. Nice, James A. Fordyce, Matthew L. Forister, and Zachariah Gompert. "The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies." Molecular ecology 27, no. 12 (2018): 2651-2666.

Signed: 

Date: 1 April 19

I hereby give permission to Samridhi Chaturvedi to include the following articles in her thesis/dissertation, of which I am a coauthor.

Chaturvedi, Samridhi, Lauren K. Lucas, Chris C. Nice, James A. Fordyce, Matthew L. Forister, and Zachariah Gompert. "The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies." Molecular ecology 27, no. 12 (2018): 2651-2666.

Signed:

Date: 2 April 2019

I hereby give permission to Samridhi Chaturvedi to include the following articles in her thesis/dissertation, of which I am a coauthor.

Chaturvedi, Samridhi, Lauren K. Lucas, Chris C. Nice, James A. Fordyce, Matthew L. Forister, and Zachariah Gompert. "The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies." Molecular ecology 27, no. 12 (2018): 2651-2666.

Signed: 

Date: April 1, 2019

I hereby give permission to Samridhi Chaturvedi to include the following articles in her thesis/dissertation, of which I am a coauthor.

Chaturvedi, Samridhi, Lauren K. Lucas, Chris C. Nice, James A. Fordyce, Matthew L. Forister, and Zachariah Gompert. "The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies." Molecular ecology 27, no. 12 (2018): 2651-2666.

Signed: 

Date: 1 April 2019

I hereby give permission to Samridhi Chaturvedi to include the following articles in her thesis/dissertation, of which I am a coauthor.

Chaturvedi, Samridhi, Lauren K. Lucas, Chris C. Nice, James A. Fordyce, Matthew L. Forister, and Zachariah Gompert. "The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies." Molecular ecology 27, no. 12 (2018): 2651-2666.

Signed: 

Date: 1 April 2019

APPENDIX B

Copyright Letters

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Mar 31, 2019

This Agreement between 5305 Old Main Hill ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4559700502627
License date	Mar 31, 2019
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Molecular Ecology
Licensed Content Title	The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies
Licensed Content Author	Samridhi Chaturvedi, Lauren K. Lucas, Chris C. Nice, et al
Licensed Content Date	May 3, 2018
Licensed Content Volume	27
Licensed Content Issue	12
Licensed Content Pages	16
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	QUANTIFYING THE PREDICTABILITY OF EVOLUTION AT THE GENOMIC LEVEL IN LYCAEIDES BUTTERFLIES
Expected completion date	May 2019
Expected size (number of pages)	200
Requestor Location	5305 Old Main Hill 5305 Old Main Hill LOGAN, UT 84322 United States Attn: 5305 Old Main Hill
Publisher Tax ID	EU826007151
Total	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright

Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS

OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes

all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library

http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

CURRICULUM VITAE

Samridhi Chaturvedi

EDUCATION

Utah State University, Logan, UT *August 2014 - Present*

Ph.D. Candidate in Biology/Ecology, *Expected graduation: April- May 2019*

Dissertation: "Genomic basis of local adaptation in *Lycaeides* butterflies."

Advisor: Dr. Zach Gompert

Committee members: Dr. Matthew Forister, Dr. Karen Kapheim, Dr. Susannah French and Dr. Karen Mock

Vellore Institute of Technology, India *2008-2010*

M.Sc. in Applied Microbiology

Dissertation: "Molecular Genetics and Pharmacogenomics of Asthma."

Advisor: Dr. Uros Potochnik, University of Maribor, Slovenia

Christ College, India *2005-2008*

B.Sc. in Chemistry, Botany, Zoology

PEER-REVIEWED PUBLICATIONS

1. **Chaturvedi, Samridhi**, Lauren K. Lucas, Chris C. Nice, James A. Fordyce, Matthew L. Forister, and Zachariah Gompert. "The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies." *Molecular ecology*, 27:2651-2666 (2018). **Covered in the news and views section with a perspective article titled "Predicting evolutionary predictability" by Dr. Catherine Linnen*

2. **Chaturvedi, Samridhi**, Alexandre Rego, Lauren K. Lucas, and Zachariah Gompert. “Sources of Variation in the Gut Microbial Community of Lycaeides melissa Caterpillars." *Scientific reports* 7, no. 1 (2017): 11335.

3. Pandey RN, Nag AK, **Chaturvedi S**. “Synthesis, spectral and antimicrobial activity of Organotin (IV) compounds ligated by anthranilic acid." Journal of Ultra Chemistry, Volume 9, Issue 3, Page number 332-337, 2016

MANUSCRIPT IN PREPARATION

1. **Chaturvedi, Samridhi**, Lauren K. Lucas, and Zachariah Gompert. “Does historical admixture predict patterns of introgression in a contemporary hybrid zone? Insights from *Lycaeides* butterflies."(*Target journal: Science Advances*)

OTHER PUBLICATIONS

1. **Chaturvedi, Samridhi**. Bioinformatics skills âĂŞ How to get them and not get scared. American Society of Naturalist, Graduate Student Blog, January 16, 2018

2. **Chaturvedi, Samridhi**, Uros Potochnik. Molecular Genetics and Pharmacogenomics of Asthma. MS Thesis, Vellore Institute of Technology, India and University of Maribor, Slovenia (2010).

POSTERS AND TALKS

1. **Samridhi Chaturvedi**, Lauren Lucas, Zach Gompert. “Does historical admixture predict patterns of introgression in a contemporary hybrid zone? Insights from *Lycaeides* butterflies." International Conference on the Biology of Butterflies, Bangalore, India, June 2018. (Talk)

2. **Samridhi Chaturvedi**, Lauren Lucas, Matt Forister, Zach Gompert. “Genomic analyses uncover parallel and idiosyncratic evolutionary changes following the colonization of a novel host plant." Evolution, Portland, OR, June 2017. (Talk)

3. **Samridhi Chaturvedi**, Lauren Lucas, Matt Forister, Zach Gompert. "Genomic analyses uncover parallel and idiosyncratic evolutionary changes following the colonization of a novel host plant." Entomological Society of America, Pacific Branch Meeting, Portland, OR, April 2017. (Invited symposium talk)

4. Tyler Ayers*, **Samridhi Chaturved**i, Zach Gompert. "Effect of photoperiod on diapause behavior in *Lycaeides melissa* butterflies." Undergraduate Student Research Symposium, Utah State University, Fall 2016. (Poster) **Undergraduate assistant*

5. **Samridhi Chaturvedi**, Lauren Lucas, Matt Forister, Zach Gompert. "Genomic analyses uncover parallel and idiosyncratic evolutionary changes following the colonization of a novel host plant." Evolution Meeting, Austin, TX, 2016. (Poster)

6. **Samridhi Chaturvedi**, Zach Gompert. "Genomic insights on the recent evolution of novel host use in the Melissa blue butterfly." Inaugural Meeting, Pan American Society for Evolutionary Developmental Biology, Clark Kerr Campus, University of California Berkeley, 2015. (Poster)

7. **Samridhi Chaturvedi**, Zach Gompert. "Genomic insights on the recent evolution of novel host use in the Melissa blue butterfly." Student Research Symposium, Utah State University, Logan, 2015. (Poster)

8. **Samridhi Chaturvedi**, Zach Gompert. "Genomic Insights on the Recent Evolution of Novel Host Use in the Melissa Blue Butterfly (*Lycaeides melissa*)." Plant and Animal Genome XXIII Conference. Plant and Animal Genome, 2015. (Poster)

RESEARCH GRANTS

USU RGS, Dissertation Fellowship (Full Tuition + $5000)	*Fall 2018 - Spring 2019*
American Society for Naturalist Travel Award for Evolution 2017 ($500)	*May 2017*
USU Ivan J. Palmblad Graduate Research Award ($1000)	*April 2017*
USU Ecology Center Grant Award ($4000)	*2015 - 2016*
USU Department of Biology Travel Awards ($1800)	*2015, 2016, 2017*

University of Slovenia, Maribor - Semester Award Fellowship (Euro 800) *Spring 2010*

TEACHING EXPERIENCE

Evolutionary Biology, *Teaching Assistant* *Fall 2018*

Assisted students to draft essays and graded the essays (This is a communication intensive course).

Principles of Genetics, *Instructor* *Spring 2018*

Designed syllabus, delivered lectures, prepared teaching schedule, teaching material and exams.

Introduction to Biology (BIOL 1610), *Teaching Assistant* *2015, 2016, 2017*

Delivered lectures before labs, helped students in conducting experiments, graded assignments and provided feedback.

Biology and Citizen Science (BIOL 1010), *Teaching Assistant* *Fall 2016*

Delivered lectures before labs, helped students in conducting experiments, graded assignments and provided feedback.

FIELD WORK EXPERIENCE

Cache Valley, Logan, UT (Collecting butterflies and sweep netting) *2015-2018*

Grand Tetons and Yellowstone NPR, USA (Collecting butterflies) *July-August 2016*

Western Ghats Regin, India (Collecting frogs) *December 2011*

WORK EXPERIENCE

Project research assistant, NCBS, India *September 2013 - January 2014*

Project research assistant, IISc, India *November 2011 - March 2013*

Content writer, Banyon Tree Inc., India *May 2013 - August 2013*

OUTREACH

Facilitator/Coordinator, Native American STEM Mentoring Program *May - June 2018*

Graduate Student Member, American Society for Naturalists Diversity Committee *2017 -*

Student Volunteer, USU Department of Biology, SACNAS student visit *October 2017*

Student Helper, Software Carpentry Workshop, Utah State University *September 2017*

Intern, Community Health Cell, Bangalore *Oct - Dec 2010*

VOLUNTEER AND OTHER ACTIVITIES

Reviewer, American Society for Naturalists Student Research Award *2019*

Student Representative, Graduate Program Committee, Utah State University *2018-2019*

Judge, Undergraduate Research Symposium, Utah State University *April 2018*

Volunteer, Science Unwrapped, Nabokov's Butterflies, Utah State University *Spring 2018*

Judge, Undergraduate Research Symposium, Utah State University *December 2017*

Academic Chair of Biology Graduate Student Association, Utah State University *2017 - 2018*

Student Representative, Departmental Seminar Committee, Utah State University *2017 - 2018*

Student Volunteer, Evolution, Austin, TX *2016*

Student Volunteer, Student Conference for Conservation Science, India *2013*

Student Representative, Christ College, India *2006 - 2008*

SOCIETY MEMBERSHIPS

Society of Evolutionary Biology

Entomological Society of America

American Society for Naturalists

Genetics Society of America

MANUSCRIPT REVIEW

Molecular Ecology (3 papers), Scientific Reports (1 paper)

REFERENCES

1. Dr. Zach Gompert (PhD advisor)

   Assistant Professor

   Department of Biology

   5303 Old Main Hill

   Utah State University, Logan, Utah 84322

   Email: zach.gompert.usu.edu

2. Dr. Matt Forister (Committee member and collaborator)

   McMinn Professor of Biology

   Dept. of Biology / MS 314

   1664 N. Virginia St.

   University of Nevada, Reno, Nevada 89557

   Email: forister@gmail.com

3. Dr. Alan Savitzky

   Professor

   Department of Biology

   5303 Old Main Hill

   Utah State University, Logan, Utah 84322

   Email: savitzky@usu.edu