

CIS-REGULATORY EVOLUTION IN *HELICONIUS* BUTTERFLIES

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
James Joseph Lewis
January 2017

© 2017 James Joseph Lewis

CIS-REGULATORY EVOLUTION IN *HELICONIUS* BUTTERFLIES

James Joseph Lewis, PhD

Cornell University 2017

cis-Regulatory element evolution is a key mechanism of biological diversification. Surprisingly little is known, however, about patterns of gene regulatory evolution across a range of divergence times, and the extent to which such variation drives local genomic adaptation. In chapter 1, we introduce the functional genomic methods used in this dissertation, and briefly discuss the current state and future prospects for the study of gene regulatory evolution. In chapter 2, we characterize the evolution of regulatory loci in butterflies and moths using ChIP-seq annotation of regulatory elements across three stages of *Heliconius* head development. In the process we provide a high quality, functionally annotated genome assembly for the butterfly *Heliconius erato*. Comparing cis-regulatory element conservation across six lepidopteran genomes, we find that regulatory sequences evolve at a pace similar to that of protein-coding regions. we also observe that elements active at multiple developmental stages are markedly more conserved than elements with stage-specific activity. Surprisingly, we also find that stage-specific proximal and distal regulatory elements evolve at nearly identical rates. This study provides a benchmark for genome-wide patterns of regulatory element evolution in insects, and shows that developmental timing of activity strongly predicts patterns of regulatory sequence evolution. In chapter 3, we use functional assays for chromatin accessibility and histone modifications to test the hypothesis that intraspecific genomic divergence is linked to regulatory variation between distinct populations of *Heliconius* butterflies. We show that population-level variability in both chromatin accessibility and regulatory activity are abundant within the *Heliconius* genome. We further show that

differences in regulatory activity between populations do not require associated differences in chromatin accessibility, illustrating that different modes of regulatory variation can be evolutionarily decoupled. Importantly, patterns of regulatory variation depart from neutral expectations, suggesting that selection underlies much of the observed regulatory divergence. Supporting this, genomic regions with high F_{st} are highly enriched for variable regulatory elements, and half of all differentially expressed genes have variable promoter-associated regulatory elements. Our work shows that regulatory elements vary between populations at different functional levels, and that selection on variable elements is a major force underlying genomic divergence within specie

BIOGRAPHICAL SKETCH

James Lewis received a BA in Philosophy from the University of Evansville in 2006. He studied the Philosophy of Biology with Dr. Richard Burian at Virginia Tech from 2007-2010, where he received his MA in Philosophy. Working with Dr. Robert D. Reed, James received his PhD in Ecology and Evolutionary Biology from Cornell University in 2017.

To Vic and Karin, for your support along the way.
And to my Dad, for getting me started in Biology.

ACKNOWLEDGEMENTS

We thank Andy Clark, Amy McCune, Charles Danko, Philipp Messer, and members of the Reed Lab for their helpful comments. This work was supported by NSF grant DEB-1354318.

TABLE OF CONTENTS

Biographical Sketch	v
Dedication	vi
Acknowledgements	vii
1. Cis-Regulatory evolution and the role of functional genomics in the study of evolutionary biology	1
2. ChIP-seq annotated <i>Heliconius erato</i> genome highlights patterns of cis-regulatory evolution in Lepidoptera	12
3. Chromatin profiling reveals that genome-wide regulatory adaptation shapes population-level genomic landscapes in <i>Heliconius</i> butterflies	36
Appendix I: Supplement to Chapter 2	65
Appendix II: Supplement to Chapter 3	86
References	98

Chapter 1:
Cis-Regulatory evolution and the role of functional genomics in the study of evolutionary
biology

James J. Lewis

Department of Ecology and Evolutionary Biology, Cornell University.

Introduction

The core foci of contemporary evolutionary biology—the genetic basis of phenotypic variation, mechanisms of speciation, and the patterns of diversification—have seen a recent renaissance due largely to the availability of massively parallel DNA sequencing. Over the past decade whole genome sequences have shifted from a privilege for those studying a few model organisms, whose genomes were laboriously pieced together by large consortiums, to a broadly obtainable resource of even individual research groups [1]. Whole genome resequencing data has subsequently inundated the evolutionary community, leading to rapid advances in how we think about genomic loci driving evolutionary change. Our focus as a whole has shifted from a gene-centric perspective to a more holistic network-based model where transcriptional regulation, coding variation, and post-transcriptional processes are all illuminated via signatures of selection as key elements in understanding the mechanisms of phenotypic variation [2, 3].

This influx of genomic data has led to the inevitable dilemma of knowing what is important, but not how these loci influence and are influenced by evolutionary dynamics. That is, we are often left wanting as to the precise mechanisms of speciation, phenotypic variation, and diversification and how these mechanisms are acting to induce evolutionary change. Yet we are now at the cusp of a second revolution in evolutionary biology, where we can see the emergence of evolutionary functional genomics as a discipline tying genomic and transcriptomic data to mechanisms of genomic, transcriptional, and post-transcriptional regulation [4, 5]. Once the purview of highly funded research groups working solely on model organisms, we are now seeing the introduction of a wide variety of functional

genomics methods into the mainstream world of evolutionary biology, with even more to arrive as methods are refined and research questions in non-model systems adapt to the availability of novel resources. Here I synthesize a diverse range of methods for assaying gene regulatory mechanisms, and review recent efforts at incorporating these functional assays into evolutionary biology. Furthermore, I aim to illuminate how a transition to functional association of genomic data can establish the developmental relevance of sequence variation, and finally to identify novel research questions we can only now begin to consider asking.

What is functional genomics?

Broadly construed, functional genomics is the generation of DNA sequence data as a proxy for functional, biochemical processes occurring throughout the genome at the time of assay. This definition somewhat arbitrarily excludes protein and protein sequencing based assays from functional genomics, which differ significantly in their methodology and maturity, and differs from traditional genomics in that we are using genomic sequences to identify specific biochemical interactions within the cell or tissue under investigation. Furthermore, biochemical interactions are assayed at the time of investigation, rather than identifying static markers of genetic variation independent of their biological or developmental context. From the perspective of evolutionary biology, this allows us to target loci under evolutionary pressure in association with the developmental context of their significance, providing deeper insight into the mechanistic basis of evolutionary change. Though methods in functional genomics are rapidly changing as sequencing technology, molecular biology, and computational

infrastructures mature, here I will briefly review some of the more significant recent functional genomic approaches to assaying biochemical interactions.

Assays of regulatory element activity and function

Publication of whole genome analyses of regulatory elements and transcriptional activity in human and other model systems by the ENCODE [6] and modENCODE [7] projects introduced functional genomics, and assays of regulatory function, to a broad audience. For the first time, a holistic picture of whole genome regulatory activity was starting to form—at least for a few model organisms. We began to see at large scale resolution the ubiquity and significance of promoter elements, which recruit transcriptional machinery to the transcription initiation site of gene bodies, distant enhancers elements, that assist with recruiting RNA Polymerase II via transcription factor induction of long-range chromatin interaction, and insulator elements, or binding sites for a small set of transcription factors that act as gatekeepers of regulatory function by blocking enhancer-promoter interactions across insulator boundaries. Prior to these publications, many functional genomic methods were pursued by only a few labs at the forefront of molecular biology, while the majority of regulatory assays required tedious and difficult dissection of specific functional elements using transgenic techniques [8]. Subsequent method refinement, sequencing technology improvements, and advances in analysis pipelines have ushered in a wave of novel research into regulatory element function in molecular biology. To a lesser extent, this activity has carried over to evolutionary biology, in part due to technical and methodological limitations of many evolutionary models, though it is easy to foresee a future where regulatory assays become common in the evolutionary literature.

The earliest methods for detecting regulatory elements have their roots in pre-genomic molecular biology. DNase footprinting, developed in 1977 [9], leveraged the DNase-blocking behavior of DNA-bound protein complexes to identify DNA-protein binding sites. As most DNA-binding proteins associated with enhancer, promoter, and insulator activity require that genomic DNA be accessible, or lacking histone octamers around the binding locus, this process could be modified to perform a whole genome screen of regulatory element activity (DNase-seq) [10]. Briefly, live nuclei with intact chromatin are extracted from a cell or tissue type of interest and exposed to a low concentration of DNase I enzyme for a brief period of time, producing preferential cleavage of genomic DNA in functional loci lacking histone octamers. These small DNA fragments are then isolated, sequenced, and aligned back to a reference genome. Significant fragment pileup from the sequencing data is interpreted as a DNase I hypersensitive locus, indicating the presence of a regulatory element. Alternate protocols have since been developed to address some of the technical roadblocks of DNase-seq. Formaldehyde-Assisted Isolation of Regulatory Elements, or FAIRE-seq [11], utilizes formaldehyde crosslinking to reduce the dependency on live nuclei at the time of assay (though nuclei must be intact at the time of crosslinking). More recently, development of an Assay for Transposase-Accessible Chromatin and sequencing (ATAC-seq) [12] uses the preferential transposition of a Tn5 transposase in accessible genomic loci to assay regulatory elements, and requires fewer than 50,000 cells for a successful screen. A 2011 comparison by Song et al. [13] of complimentary FAIRE-seq and DNase-seq experiments found that approximately 64% and 72% of the 25,000 most significant FAIRE and DNase

detected elements, respectively, were contained in the 50,000 most significant elements of the complimentary data set. Similarly, Buenrostro et al. [12] found approximately 74% of functional loci intersected between ATAC and DNase-derived regulatory assays, suggesting that all three methods provide robust screens of regulatory activity.

The most widely adopted assay of regulatory element activity and function has been chromatin immunoprecipitation and sequencing (ChIP-seq) [14]. Used to detect protein-DNA binding sites, ChIP-seq has been adapted for use with multiple classes of DNA-binding associated proteins, including site specific and general transcription factors, histone protein variants, and transcription co-factors. During ChIP-seq, protein-DNA interactions in cells or a tissue of interest are crosslinked with formaldehyde, after which the crosslinked DNA is sheared or digested to produce 100-500bp long fragments. An antibody to the epitope (protein of interest) is used to capture, or immunoprecipitate, any protein-bound DNA fragments. These fragments are then purified from associated proteins and sequenced using short-read high throughput sequencing. Alignment of the ChIP sequencing product results in loci of significant fragment pileup, or ChIP “peaks”, indicating protein binding at that locus. ChIP-seq has been used extensively in model organisms to identify genome wide binding patterns for numerous generalist and site specific transcription factor proteins [6, 7, 15].

Our knowledge of regulatory element biochemistry and function has been vastly improved by the use of ChIP-seq against antibodies to histone protein variants [16]. Regulatory

element function requires that genomic DNA be made accessible to transcription factor binding for subsequent enhancer-promoter interaction (in the case of enhancers) or recruitment of the core transcriptional machinery (in the case of promoters). This process begins with binding of specialized pioneer factor proteins to regions of closed or inaccessible chromatin—DNA that is tightly wound around histone octamers composed of dual copies of each of the core histone proteins H2A, H2B, H3, and H4, collectively known as the nucleosome [17]. Pioneer factors then signal to nucleosome repositioning proteins, which shift neighboring nucleosomes such that the internucleosomal space containing the regulatory locus becomes accessible to transcription factor binding and proper regulatory activity [18]. Whole genome assays of histone variants, which are modified or alternate forms of the core histone proteins, has identified a number of distinct functional states associated with nucleosomes flanking regulatory elements.

While more than 65 histone variants are known, a few well studied modifications have provided a core set of markers for regulatory activity [19]. Distal regulatory loci, such as enhancer elements, are often marked first by mono-methylation of lysine residue 4 of histone H3 (H3K4me1). Similarly, proximal regulatory elements, most often promoters, are frequently recognized by tri-methylation of the same lysine (H3K4me3), though both H3K4me1 and H3K4me3 can also mark proximal and distal regulatory elements, respectively, to a lesser degree. Lysine 27 acetylation on histone H3 has been shown to be a strong signal of active regulatory function for both distal and the most highly active proximal regulatory loci. Additional markers, such as H3K9ac, H3K27me3, and H3K14ac have been used to provide additional evidence of regulatory function and activity [6, 7,

19]. For our purposes here, such assays of regulatory function have proven most useful in identifying different types of regulatory loci for downstream evolutionary analysis.

cis-Regulatory evolution and organismal divergence

Much of the early evidence tying gene regulatory mechanisms to species divergence was specific to a select few model organisms. *Drosophila melanogaster* was particularly influential in this regard, with multiple studies demonstrating that regulatory variants acting on melanin-associated genes induce divergent wing and abdominal pigmentation patterns across *Drosophila* species [20, 21]. More recent examples of regulatory evolution include loss-of-function nucleotide variants at the *pitx1* locus in stickleback fish [22] and ZRS enhancer in snakes [23] driving adaptive morphological change via pelvis and limb reduction, respectively. Gain-of-function gene regulatory evolution has been less prevalent to date, though several studies have implicated gain of enhancer elements in lepidopteran wing pattern evolution, such as red and black color patterns in *Heliconius* butterflies [24].

Adoption of high-throughput assays of regulatory elements has provided increasing evidence in support of rapid cis-regulatory turnover as a primary driving force behind organismal diversification. Whole-genome ChIP-seq analysis of active enhancer and promoter conservation across 180MY of mammalian evolution determined that only ~1% of enhancer and 16% of promoter loci were evolutionarily conserved [25], a major departure from rates of gene body turnover. A similar study of *Drosophila* enhancer conservation using transgenic reporter assays in S2 cells found that 22% of enhancer

elements have lost their function over only 11MY of divergence between *D. melanogaster* and *D. yakuba* [26]. While evidence is more limited for population-level regulatory divergence, several studies in human cell lines have provided initial perspectives on intra- and inter-population variability of gene regulatory activity. Intra-population assays of regulatory element accessibility by DNase-I hypersensitivity assays found that .03% of hypersensitive loci were variable and associated with changes in nearby gene expression within 70 LCL samples of Yoruban descent [27]. Expanding observation of regulatory variability to inter-population divergence, ChIP-seq assay of multiple histone modifications in LCLs from four human ethnic lineages found that up to 5-6% of active regulatory loci varied by ethnic lineage [28].

Looking Forward: future directions for the study of cis-regulatory evolution and divergence

Despite initial benchmarks of macro- and micro-evolutionary turnover of cis-regulatory elements, we still know little about whole-genome patterns of regulatory evolution. Two points of departure from prior observations of cis-regulatory evolution stand out in particular: First, we lack a developmental context for much of what we currently know about regulatory element turnover. Previous studies have primarily focused on adult vertebrate tissues and various assays of in vitro cell lines, lacking developmental information or even contextual endocrine and paracrine signaling in the latter. Moving forward, it will become important to identify patterns of regulatory evolution associated with periodic developmental processes in primary tissue types. As second point departure, we know little about how population- and species-level adaptation drives

regulatory variation in natural systems. Much of our current knowledge hinges on studies of domesticated species (though not entirely) with human regulatory loci used as a common reference for evolutionary inference, or relies entirely on human cell cultures. We have much to learn about how populations and species have diverged under natural evolutionary conditions. This is especially important for the study of evolutionary biology, as borderline cases of speciation and population divergence where limited introgression occurs are often the most informative for understanding the developmental and evolutionary processes underlying adaptive evolution. Moreover, study of natural cis-regulatory associated population and species divergence will, in part, address much of the concern over incorporating in vivo signaling processes into our understanding of gene regulatory evolution.

Conclusion

In the next two chapters of this dissertation, I aim to begin to address some of the outstanding questions raised above using whole-genome assays of chromatin accessibility and regulatory element activity. In Chapter 2 I use ChIP-seq for two histone modifications marking active enhancer and promoter elements across three stages of lepidopteran head development followed by evolutionary analysis of regulatory element sequence conservation to elucidate the developmental and mechanistic bases of regulatory element turnover in butterflies. In Chapter 3, I use ChIP-seq and ATAC-seq to show that regulatory element evolution is both rampant and multi-layered between populations of *Heliconius* butterflies. Moreover, I demonstrate that variability in cis-regulatory architecture is tied to signatures of high nucleotide divergence between

populations, suggesting that much of the regulatory variation between populations is adaptive.

Chapter 2:
ChIP-seq annotated *Heliconius erato* genome highlights patterns of cis-regulatory
evolution in Lepidoptera

James J. Lewis^{*}, Karin R. L. van der Burg, Anyi Mazo-Vargas, and Robert D. Reed
Department of Ecology and Evolutionary Biology, Cornell University.

Introduction

One of the paradigm-defining discoveries emerging from efforts to functionally annotate genomes is the degree to which regulatory elements dominate the genomic landscape. Indeed, assays of chromatin accessibility, a general signature of most regulatory loci, identified over 2 million regulatory elements across 125 human cell lines [6]. This discovery, coupled with the many case studies implicating cis-regulatory activity as a driving force of morphological evolution [29, 30], clearly points to the importance of regulatory elements in shaping not only organisms, but genome structure itself. Unfortunately, despite the centrality of regulatory sequences to organismal development, function, and evolution, we still lack a general understanding of genome-wide patterns of regulatory element evolution, especially outside of major vertebrate lineages. One of the challenges of doing large-scale comparative work on regulatory sequences has been the difficulty of annotating regulatory elements on a genomic scale. Efforts to predict and compare putative regulatory elements based on purely computational approaches (e.g. sequence conservation, binding motif predictions, etc.) have produced important results [31] but also have limitations [32, 33]. More recent efforts to incorporate functional regulatory element annotations have made use of chromatin immunoprecipitation and sequencing (ChIP-seq), where antibodies targeting DNA-binding proteins of interest are used to isolate genomic sequences with regulatory activity [5, 25]. As yet, however, this approach has seen limited use outside of a few model organisms, despite holding exceptional potential for applications in emerging model systems and comparative studies. A broader sampling of stage- and tissue-specific genome-wide functional

annotations across a diverse set of lineages will be essential for gaining an understanding of general patterns of regulatory evolution in eukaryotes.

To date, relatively few published studies have used functional annotation data to examine whole-genome trends in regulatory sequence evolution. Of significant interest here are two comparative studies that used whole-genome ChIP annotations of mature vertebrate liver tissue. In one study Schmidt et al. used CEBPA ChIP assays to study conservation of transcription factor binding in livers of five vertebrate species [5]. Human CEBPA binding sites displayed between 15% and 2% conservation across 300 My of evolution in five vertebrate species. Another investigation of active regulatory elements in livers of 20 mammalian species, this time using histone tail modifications associated with active regulatory loci (H3K27ac and/or H3K4me3), found similar results [25]. Comparing all active regulatory loci, Villar et al. found only 1% of presumptive enhancers and 16% of presumptive promoters were conserved between all 20 species over 180 My of divergence. Slight incongruences between the two ChIP-based studies are likely the result of targeting a conserved transcription factor in the former study, combined with a different taxon sampling scheme in the latter. The results of both studies, however, support the view of rapid regulatory element turnover with somewhat greater conservation of promoter elements relative to more distal transcription factor binding sites (e.g. enhancers). These studies are important landmarks for understanding the functional evolution of genome structure in animals. Surprisingly, however, we are unaware of similar investigations outside of amniotes. We thus lack even a preliminary benchmark of

genome-scale trends in regulatory sequence evolution for most of the major lineages of life.

The increasing availability of genome assemblies for emerging model organisms has precipitated a heightened interest in broad taxonomic patterns of genome-scale regulatory architecture outside of vertebrate systems [34, 35], though as yet there has been little work on large-scale patterns of regulatory evolution in non-vertebrate lineages. Compounding this problem, we also lack a fundamental understanding of the degree to which developmental context and utility governs the evolutionary trajectory of regulatory loci. Genome-wide studies of regulatory activity in invertebrate species have thus far, with a few notable exceptions, focused primarily on *ex vivo* assays of cell culture activity or whole organism tissue samples [7, 36]. Even the few exceptions [37-39] have rarely focused on more than one developmental time point, and to the best of our knowledge, none have assayed regulatory activity over multiple periods of major developmental reorganization from tissue patterning to maturation. Despite this, several common features of regulatory activity have become apparent. One important observation is that regulatory elements are frequently reutilized between tissue-specific developmental programs. Of the 155,000 transcription factor binding sites annotated by the modENCODE consortium, assayed over a broad spectrum of developmental stages in whole *D. melanogaster*, only 35,000 binding sites were unique, stage-specific genomic loci [7]. Even allowing for multiple factor binding events at most regulatory elements, this indicates a high degree of developmental reutilization of regulatory sequence loci. Importantly, this trend appears to be conserved broadly amongst eukaryotes. Observation

of regulatory element accessibility in a diverse array of human cell lines found that 66% of observed regulatory loci were accessible in two or more cell lines [6]. Interestingly, however, only 0.1% of elements were accessible in all 125 assayed cell types, suggesting that study of a single cell type or tissue is unlikely to be universally representative. The general tendency towards complex regulatory reutilization—i.e. when a regulatory element is active in multiple developmental stages or tissue types—raises an interesting question regarding the relationship between stage-specific regulatory landscapes and evolutionary conservation of regulatory loci, and highlights a deep need for additional comparative study of *in vivo* regulatory activity across multiple developmental stages.

Here we generate the first portrait of genome-wide patterns of regulatory element evolution in an insect lineage, the Lepidoptera, and ask if the genomic position of elements and/or the developmental timing of regulatory activity is predictive of regulatory sequence conservation. We provide a high quality draft genome assembly for the butterfly *Heliconius erato* (race *lativitta*), a model organism for research on wing pattern mimicry and speciation. Using antibodies targeting histone modifications, we annotated a time series of active regulatory elements during three key stages of *H. erato* head development—a data set which should be useful for future studies of behavior and vision in this species and other Lepidoptera. We identified a core set of regulatory elements active across three stages of head development, as well as sets of regulatory loci with stage-specific activity. To determine broad trends of regulatory sequence evolution, we investigated sequence conservation of *H. erato* regulatory elements across genomes from five additional lepidopteran species spanning 116 million years of evolution. We

provide evidence of regulatory evolution at both transcription start site (TSS)-proximal and TSS-distal loci, and show that regulatory element loci with limited, stage-specific activity have diverged more rapidly than elements active across multiple stages of development. Moreover, we show that developmental timing of activity is a stronger predictor of regulatory sequence than TSS proximity alone.

Results

***H. erato* genome assembly and annotation**

Illumina short read (~220bp) and mate pair (3kb, 8kb, 12kb) libraries made from a single, outbred female *H. erato lativitta* (*Hel*) pupa was assembled to produce an initial assembly of 12,985 scaffolds, with scaffold and contig N50 values of 362kb and 13.2kb, respectively. The total assembly length, including scaffold gaps, was ~670Mb. As previously reported, flow cytometry estimated a genome size of approximately 400Mb for *Heliconius erato petiverana* [40], suggesting a significant percentage of our initial *Hel* assembly consisted of dual haplotypes. Haplotype scaffolds from the initial Illumina assembly were merged together and rescaffolded using HaploMerger [41], producing an assembly with a total length of ~385Mb and considerably improving the scaffold and contig N50 values to 4.3Mb and 15.3kb, respectively. This assembly was further improved by gap-filling and additional scaffolding with Pacific Biosciences long read sequences, improving the scaffold and contig N50 values to 5.5Mb and 123kb, respectively.

Previous linkage mapping demonstrated 21 linkage groups in both *H. erato* and the close relative and co-mimic butterfly *Heliconius melpomene*, which are separated by only ten million years [40], and comparison of two assembled BAC sequences for both species shows highly similar gene order [42]. Given the observed similarity in chromosome number and local gene order, we used synteny to manually map our assembled *Hel* scaffolds to each of the 21 *H. melpomene* chromosomes, correcting 19 presumed misassembly errors in our prior *H. erato* assembly in the process [43]. Comparisons with *Eueides isabella*, which split from the *Heliconius* genus ~18 Ma, found that *Heliconius* possessed all 31 *E. isabella* chromosomes largely intact, though they subsequently fused into the 21 chromosomes found in *H. melpomene* and *H. erato* [43]. Davey et. al also identified 21 as the ancestral chromosome number for *Heliconius* species, suggesting highly conserved chromosome content between the two species for which genomes have now been assembled, further justifying our use of synteny mapping. None of the initial *Hel* scaffolds mapped to separate chromosome ends, providing additional indication that no additional chromosome fusion events had occurred, and that high-level chromosome composition is likely conserved between the *H. erato* and *H. melpomene*. Because we had no evidence to support or reject minor chromosomal mutations (e.g., small inversions, deletions, etc.), we retained low level scaffold sequence composition produced during the prior assembly. A syntenous, chromosome-level assembly was generated from previously assembled and gap-filled scaffolds to produce a final genome of 418Mb, with a scaffold N50 of 5.48Mb and a contig N50 of 129.8kb. All further analyses were performed on this final genome assembly.

A total of 14,613 genes were predicted based on three iterations of MAKER [44] incorporating a combination of mRNA sequence data, *H. melpomene* protein sequences, and SNAP and Augustus gene predictions. Orthologs of 9,741 genes were identified in *D. melanogaster* using protein BLAST (E-value threshold of 1e-5), and 9,439 genes had domains that were annotated by either the Pfam or the SUPERFAMILY analysis, where Pfam identified 14,407 protein families and SUPERFAMILY resulted in 12,750 annotations. Blast2Go annotated 5,730 GO terms for 8,086 genes [45]. Analysis of genome completeness identified 95% of the 248 core CEGMA [46] genes. Our genome assemblies and annotated gene set are available for download and browsing at butterflygenome.org

Functional annotation of head tissue cis-regulatory elements

Antibodies for two histone modifications indicative of active regulatory loci, H3K4me3 and H3K27ac, were used to identify presumptive regulatory elements in three developmental stages of *Hel* head tissue via chromatin immunoprecipitation and sequencing (ChIP-seq) (Figure 2.1, see also Figure S2.1).

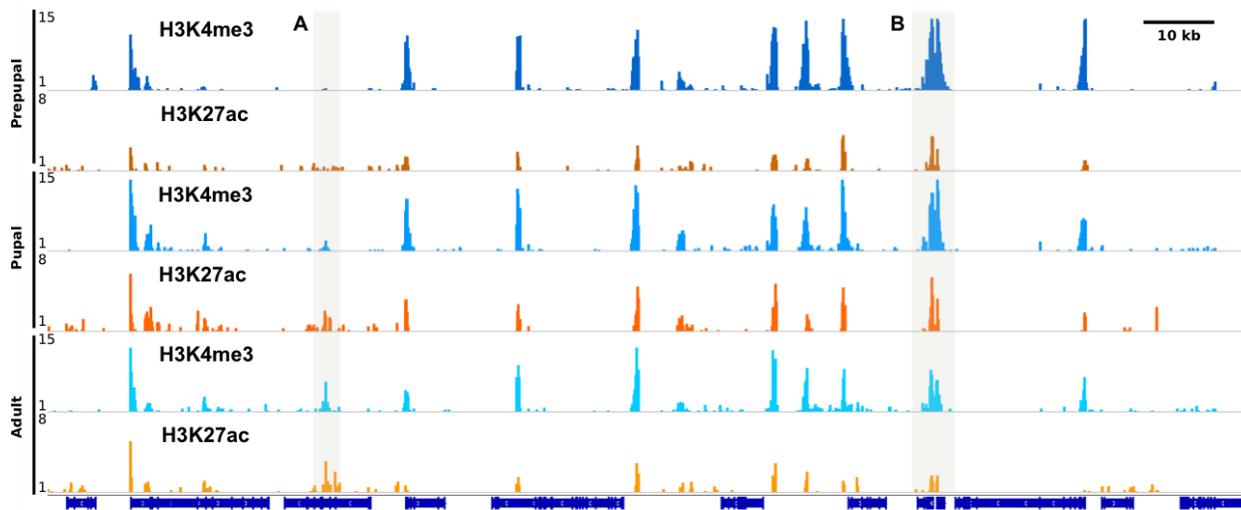


Fig. 2.1: ChIP-seq signal shows change in regulatory activity during tissue maturation. Input normalized fold enrichment profiles for H3K4me3 and H3K27ac ChIP-seq at prepupal (top), pupal (middle), and adult (bottom) developmental stages on scaffold 'chr3_5'. Representative examples of increasing (**A**) and decreasing (**B**) regulatory activity during head maturation are highlighted. See also Fig. S1.

Despite the occasional use of H3K27ac and H3K4me3 marks to distinguish between enhancer and promoter activity, respectively, multiple reports have shown that these modifications co-occur at a very high frequency in both enhancer and promoter elements [7, 47, 48], and are thus not absolutely diagnostic of promoter versus enhancer identity. In support of this view, we observed significant overlap between regulatory loci marked by the two histone modifications, though H3K4me3:H3K27ac signal intensity ratios appear to vary along with TSS proximity (Figure S2.2).

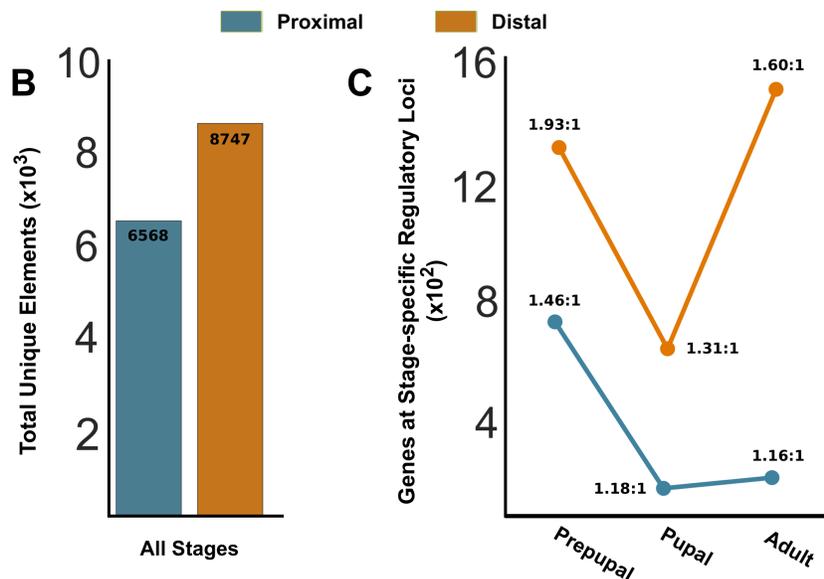
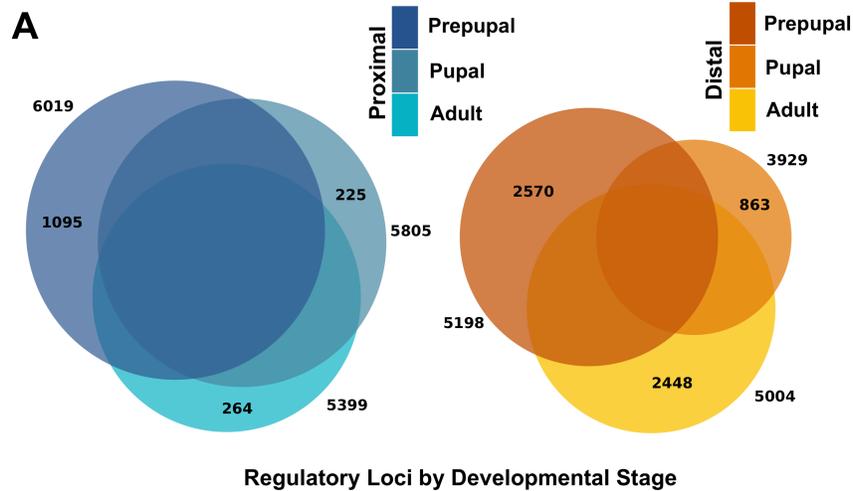


Fig. 2.2: *cis*-Regulatory architecture of *Heliconius* head tissue highlights key transitional stages. **A**) Representative overlap of proximal (blue) and distal (orange) regulatory elements, by stage. Total (outer numbers) and stage-specific (inner numbers) proximal and distal elements at each stage are numbered. Proximal elements show increased overlap relative to distal elements, and a decrease in number during tissue maturation. Distal element counts display greater variation between stages, and show more stage-specific activity across all stages. **B**) Total counts of stage-specific proximal and distal regulatory loci across all assayed developmental stages. **C**) The number of genes near stage-specific regulatory elements, by stage, with ratio of stage-specific regulatory loci to genes labeled. Genes were identified via proximity, with each point representing the count of non-repeating genes from the same scaffolds, closest to the regulatory elements. Proximal (blue) and distal (orange) elements show noticeably different gene set distributions. See also Fig. S2.

Because of this, we did not follow some previous studies in distinguishing between promoters and enhancers based on relative composition of H3K27ac and H3K4me3 marks. Instead we opted to categorize presumptive regulatory elements as either “proximal” (within 2kb of nearest TSS) or “distal” (>2kb to nearest TSS) to annotated genes, reasoning that proximal sites include promoters, while most or all distal sites are enhancers or other noncoding regulatory elements. ChIP-seq datasets are available for download and browsing at butterflygenome.org.

Proximal vs. distal elements show different patterns of stage-specific activity

In total we annotated 11,217, 9,734, and 10,403 cis-regulatory elements for prepupal, ommochrome stage pupal (approximately 6-7 days post-pupation at 30C, hereafter “pupal”), and two-day old adult (hereafter “adult”) head tissues, respectively, with our data following a trend of decreased regulatory activity over the course of tissue maturation. We observed 6,019 proximal and 5,198 distal prepupal stage regulatory loci, 5,805 proximal and 3,929 distal pupal stage regulatory loci, and 5,399 proximal and 5,004 distal adult regulatory loci (Figure 2.2a) for a total of 6,568 and 8,747 unique proximal and distal loci across all stages (Figure 2.2b). Of the annotated proximal elements, 18% of prepupal elements had stage-specific activity, while only 4% and 5% of regulatory loci in pupal and adult stage tissue were stage-specific (Figure 2.2a), suggesting that the majority of novel adult head regulatory elements become active during the transition from larval to pupal development. Unexpectedly, this was not true for annotated distal elements. While distal regulatory elements were overall more often active only in a single stage, 49% of distal

regulatory loci in both prepupal and adult stages were stage-specific, but only 22% of pupal stage elements were specific to that stage (Fig 2.2a). Therefore, our data clearly show that proximal and distal regulatory elements display very different patterns of stage-specific activity, and that the transition from larva to pupa marks the greatest period of stage-specific proximal regulatory activity. Distal stage-specific regulatory element activity appears to be most common at prepupal and adult stages, with less apparent activity during pupal head maturation.

To determine whether spatial composition of stage-specific regulatory elements could reveal patterns of gene regulatory activity during head maturation, we identified the number of genes nearest to stage-specific proximal and distal regulatory elements (Figure 2.2c). For every developmental stage, genes identified this way were, on average, closest to multiple stage-specific regulatory elements. While some number of distal stage-specific elements may be proximal to currently unannotated genes, our annotations are similar to those of *H. melpomene* and other lepidopteran species [43, 49-51], therefore suggesting that this is unlikely to be a major complication. Using the ratio of stage-specific regulatory elements to nearby genes (equal to the average number of stage-specific elements per neighboring gene) as a proxy for regulatory complexity at each stage, we observed several noticeable patterns during the process of head maturation (Figure 2.2c). As expected, the ratio of distal elements to nearby genes was in general higher than observed for proximal elements. We found a decreasing trend in proximal stage-specific loci during development, with prepupal, pupal, and adult ratios of approximately 1.5:1, 1.2:1 and 1.2:1, respectively. Distal stage-specific regulatory elements showed a more

variable trend, with prepupal, pupal, and adult ratios of approximately 1.9:1, 1.3:1, and 1.6:1. We postulate that these trends are likely indicative of an increased role of complex developmental pre patterning during early transitional periods in adult head development, while fewer regulatory interactions are required in pupal and adult head tissue. GO enrichment analysis of the nearest gene for combined proximal and distal regulatory loci at each stage supported these divergent trends in head development, with cellular communication, localization, and transport biological processes dominating early stage enriched GO categories, while later stage categories were primarily metabolic and biosynthetic (see also Table S2.1). Thus, we infer a stage-specific regulatory landscape for *H. erato* head development composed of highly complex regulatory patterning during the larval to pupal transition period, followed by a more modest regulatory landscape likely driving structural and metabolic pathways in pupal and adult head tissues.

Evolutionary divergence of regulatory elements in Lepidoptera

We used multiple recent genome assemblies across a broad phylogenetic range of Lepidoptera to investigate the degree to which functionally annotated regulatory sequences in *Hel* head tissue have been conserved during lepidopteran evolution. Nucleotide sequences at *Hel* proximal and distal regulatory loci for all three stages of head development were compared to whole genome assemblies of *Heliconius melpomene* (*Hm*), *Melitaea cinxia* (*Mc*), *Danaus plexippus* (*Dp*), *Papilio xuthus* (*Px*), and *Bombyx mori* (*Bm*) [43, 49-52]. These species were chosen as representative members of major macrolepidopteran lineages including the families Nymphalidae, Papilionidae, and Bombycidae, with the nymphalid subfamilies Danainae, Heliconiinae, and

Nymphalinae represented as well. Divergence time estimates for these six species range from recent (10Ma) for the two *Heliconius* species to the early Cretaceous (116Ma) for divergence between *Heliconius* and *Bombyx* lineages [53, 54].

We used pairwise comparisons of *Hel* regulatory elements with each of the lepidopteran species to discern patterns of regulatory sequence divergence across a range of time scales. *Hel* regulatory element sequences were considered conserved in a corresponding genome assembly if they passed a reciprocal best-hit BLAST query with a conservative threshold for acceptance (acceptance threshold had little effect on conservation counts, see Table S2.2). This approach provides a measure of the maximum possible conservation in pairwise comparisons between species, although it is important to note that it does not guarantee functional conservation. Previous work on sequence and functional conservation at regulatory loci in mammals indicates that sequence conservation alone likely overestimates functional conservation, yet nonetheless provides an important “ceiling” estimate of regulatory element conservation that can serve a benchmark for subsequent comparative and functional work [55].

As expected, proximal regulatory loci were more conserved on average than distal loci, and we observed decreasing conservation of regulatory sequences as divergence time increased (Figure 2.3a, see also Table S2.3).

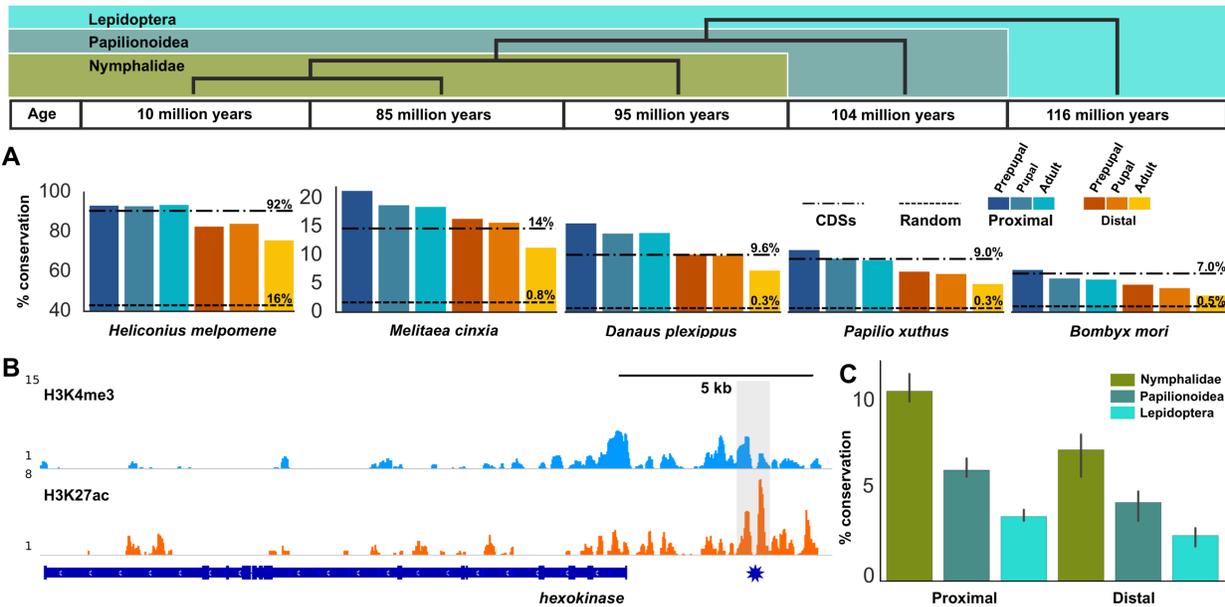


Fig. 2.3: Evolutionary trends in annotated *cis*-regulatory elements. A) Pair-wise conservation of proximal (blue) and distal (orange) regulatory elements, by stage, across five lepidopteran genomes. Short dashed lines show null expectation of conservation (*H. melpomene* null conservation not to scale) as determined by pair-wise comparison of randomly selected of genomic sequences. Long dashed lines show conservation of all annotated gene CDSs. Phylogenetic scale and taxonomic groups highlighted above. **B)** Example of a conserved distal regulatory element highly enriched for the H3K27ac histone mark, present in all lepidopteran species studied. Star indicates conserved locus upstream of hexokinase, an important constituent of the glucose (a primary component of butterfly nectar) metabolic pathway. **C)** Conservation of lepidopteran regulatory loci over increasingly broad taxonomic groups, covering approximately 116 million years of evolution. Black bars indicate conservation scores across developmental stages. See also Figure S3 and Tables S2-4.

A noticeable conservation threshold at the transition from the genus *Heliconius* to more distantly related lepidopteran species was observed. Within the genus *Heliconius*, approximately 93% of proximal and 80% of distal regulatory loci were conserved, leading us to speculate the presence of a highly conserved genus-specific developmental program associated with similar life-history traits for the two mimetic species. Moving outside of the genus *Heliconius*, conservation of regulatory loci decreased greatly with increased divergence time. Average conservation frequency of regulatory loci in these species were 19%, 14%, 9%, and 6% for proximal loci and 14%, 9%, 6%, and 4% of distal

loci, for *Mc*, *Dp*, *Px*, and *Bm* respectively. Divergence times for these lineages have been estimated at approximately 78Ma (*Mc*), 90Ma (*Dp*), 104Ma (*Px*), and 116Ma (*Bm*) [53, 54].

Importantly, the observed degree of sequence conservation in both proximal and distal regulatory loci suggested a significant departure from the null expectation of sequence conservation due to phylogenetic relatedness alone. We analyzed 10,000 sequences randomly sampled from the *Hel* genome assembly (including both coding and non-coding loci), matching the estimated size distribution of our annotated regulatory element datasets, to test whether the observed degree of conservation differed significantly from expectation under a random sampling model. Of these, 50 random loci were sampled from unfilled gaps with significant 'N' content and were subsequently discarded. Repeating the analysis with the remaining 9,950 randomly sampled sequences indicated a highly significant degree of conservation of *Hel* regulatory element sequences relative to our random model in all species comparisons (Chi-square test, $p < 0.001$) (Figure 2.3a). Performing a similar analysis with all annotated transcripts showed both proximal and distal regulatory loci diverging at similar, or often lower, rates than annotated gene CDSs (Fig 2.3a). Together our data show that *Hel* regulatory elements show significant conservation across Lepidoptera, and are subject to a degree of stabilizing selection similar to that affecting protein-coding sequences.

We applied clade-level analysis of regulatory sequence conservation to identify conservation patterns across increasingly inclusive phylogenetic groups within the order

Lepidoptera (Figure 2.3b,c). Rather than pairwise comparison of *Hel* regulatory elements between individual species as above, we instead identified all elements shared by monophyletic groups at each taxonomic level. In general, we observed results similar to those described in vertebrate studies, with proximal regulatory elements displaying increased conservation relative to distal elements. Mean conservation of regulatory sequences for all three developmental stages was 10% of proximal and 7% of distal element sequences across nymphalids (*Hm*, *Mc*, and *Dp*), 6% and 4% for all butterflies (superfamily Papilionoidea, incorporating *Px*), and 3% and 2% for all lepidopterans studied (i.e. incorporating *Bm*). Thus our analysis shows a similar degree of conservation of distal regulatory elements as previously observed in vertebrate evolution [25]. Contrary to prior observations of highly reduced turnover in TSS-proximal regulatory elements [5, 25], we found that proximal and distal regulatory loci evolve at very similar rates across lepidopteran lineages.

Multiple reports have shown that small numbers of orthologous regulatory loci can retain their function despite sequence divergence sufficient to prevent detectable pairwise alignment. A recent comprehensive analysis of regulatory sequence conservation in vertebrates found that between 0.71% and 7.1% of conserved sequences in a pairwise species comparison were likely functional, but undetectable by sequence alignment, in a more distantly related species [56]. These values are dwarfed by a prior study showing that 33% of conserved, alignable regulatory elements studied were no longer functional, suggesting that our analysis is likely to be overly conservative [55]. Nonetheless, adjusting our conservation counts according to the most significant results by Taher et al.

(7.1%) produced negligible change in our observed evolutionary trends (Table S2.4). For example, adjusted conservation counts for prepupal proximal and distal loci compared with *B. mori* were increased from 7.5% and 4.9% to 7.8% and 5.1%, respectively. Thus, while we acknowledge that some small number of loci could retain their function in distantly related species, our observed trends in regulatory sequence evolution are robust to such concerns and are more likely to be an overestimate of true functional conservation.

Stage-specific activity is associated with extremely rapid sequence divergence

Making use of our developmental time series, we classified our regulatory elements as either stage-specific (active at only a single developmental stage) or shared (active at two or more developmental stages). Differences in conservation between stage-specific and shared regulatory sequences were quite extreme (Figure 2.4, see also Table S2.3).

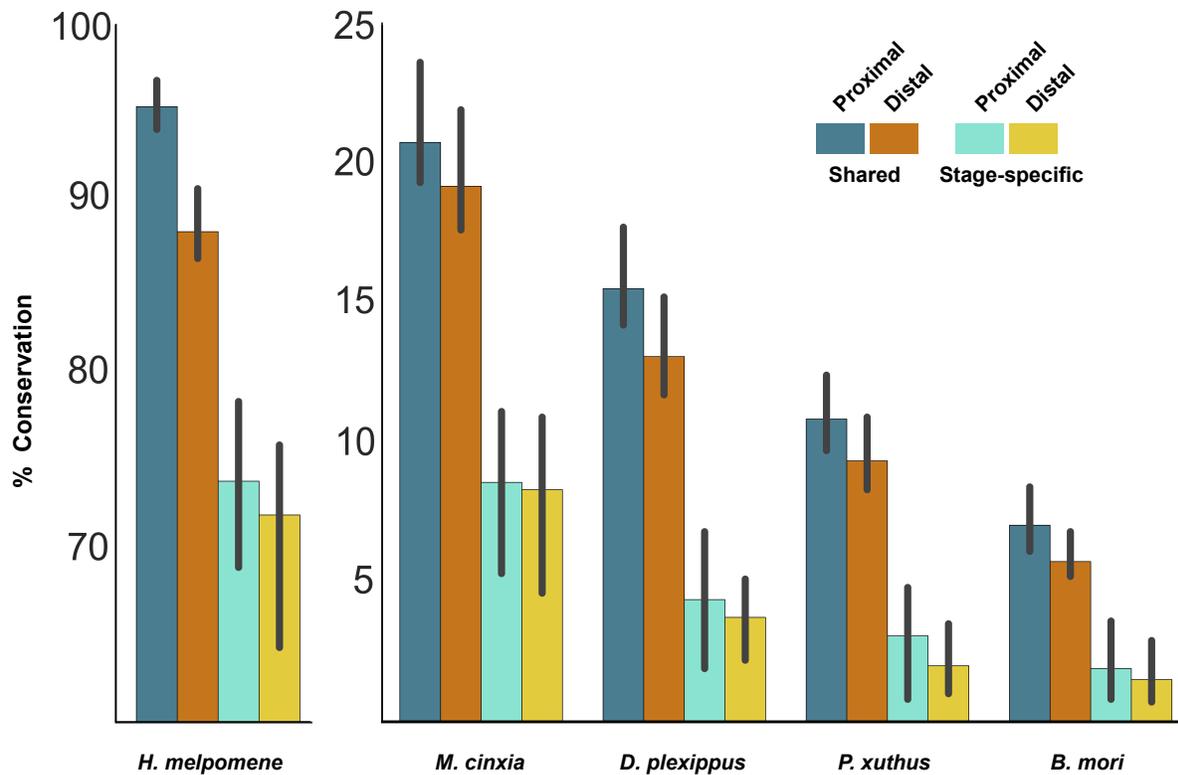


Fig. 2.4: Stage-specific and shared regulatory elements display highly dissimilar evolutionary patterns. Conservation of shared (dark) and stage-specific (light) regulatory sequences for proximal (blue) and distal (orange) regulatory elements. Shared regulatory elements show disparity in conservation between proximal and distal elements, while stage-specific loci are evolving rapidly, independent of stage and proximity to the nearest TSS. Black bars indicate conservation scores across developmental stages. See also Tables S3.

Mean conservation of shared proximal elements across all three stages was 95%, 20%, 15%, 10% and 7% for *Hm*, *Mc*, *Dp*, *Px*, and *Bm* respectively. For shared distal elements, mean conservation scores were slightly less, at 88%, 19%, 13%, 9%, and 5% for *Hm*, *Mc*, *Dp*, *Px*, and *Bm*. When considering stage-specific regulatory sequences only, conservation values between proximal and distal elements were very similar, and all were less than observed for shared elements. Mean sequence conservation of stage-specific proximal regulatory elements was 73%, 8%, 4%, 3%, and 1%, while mean conservation of stage-specific distal regulatory loci was 67%, 8%, 3%, 2%, and 1%, for *Hm*, *Mc*, *Dp*,

Px, and *Bm*. When developmental stages were considered separately, we found that prepupal stage elements were the most conserved when observing either shared or stage-specific regulatory loci, while adult loci diverged to the greatest extent. These patterns of sequence conservation suggest that stage-specific regulatory loci evolve at a rapid rate relative to shared regulatory elements, and both do so mostly independent of proximity to the nearest TSS. Moreover, the degree to which regulatory elements are shared or stage-specific in a given tissue dominates observed evolutionary patterns. For example, 95% of proximal elements from adult head tissue are shared with at least one other developmental stage, effectively driving the overall observed conservation rate of 93% in the close relative, *H. melpomene*. In contrast, only 51% of distal elements from adult head tissue are shared, with a corresponding conservation rate of 75% in *H. melpomene*. In sum, our data show that the duration of time over which an element is active during development is a strong predictor of evolutionary conservation.

Discussion

Here we present a high quality draft assembly of the *H. erato* genome and provide the first example of ChIP-based regulatory annotations for a butterfly, and one of the few such functional annotations outside a model system. By analyzing more than 15,000 unique regulatory loci over three key stages of head development we were able to identify both developmental and phylogenetic patterns of regulatory activity. While further study would be required to ascertain the functional significance of individual loci, aggregating over thousands of regulatory sequences across time paints a clear picture of regulatory activity

trends during the process of head development. Specifically, our results suggest that the transitional period from last-instar larva to pupa is marked by a large, genome-wide shift in active regulatory elements, with prepupal head tissue showing an especially high ratio of regulatory loci to genes. Interestingly, adult head tissue showed the greatest number of genes around stage-specific distal regulatory elements. The lower overall ratio of genes to regulatory elements at this stage suggests a relatively simpler regulatory landscape, presumably maintaining a large cohort of metabolic and structurally important proteins.

Here we provide evidence of invertebrate regulatory sequence conservation across a developmental time series and identify core sets of conserved regulatory sequences at multiple phylogenetic levels. Overall we found that genome wide trends in lepidopteran regulatory element conservation are similar to what has been seen in vertebrates over similar time scales (Schmidt et al., 2010; Villar et al., 2015). This is perhaps unsurprising as per generation mutation rates are similar in mammals and *Heliconius* butterflies [57, 58]. Interestingly, however, we found that lepidopteran proximal regulatory element sequences evolve almost as rapidly as those of distal elements, leading us to speculate whether this pattern may be related to developmental genetic and/or life history features particular to insects. Whatever the case, the wealth of natural history, ecological, and evolutionary data on numerous butterfly species, combined with their amenability to functional genomic work [59, 60] and the availability of multiple genome assemblies (Davey et al., 2016, Van Belleghem et al., submitted), suggests that *Heliconius* and other lepidopterans could become useful models for understanding the ecological and adaptive processes that underlie cis-regulatory evolution.

Previous studies of cis-regulatory sequence conservation have primarily emphasized regulatory elements from single, adult tissue types or computational prediction of elements isolated from their biological context [5, 25, 61]. Thus, sorting our annotated regulatory elements by stage specificity produced novel insights into regulatory sequence evolution. Conditioning our evaluation of regulatory sequence conservation on stage specificity—that is, classifying elements as active only at a single stage or active at two or more developmental stages—identified strong patterns of sequence conservation. We found that sequences of stage-specific regulatory elements evolved rapidly relative to regulatory loci active in multiple stages, and appeared to do so regardless of classification as TSS proximal or distal. These shared element sequences also demonstrate a much higher degree of conservation than expected relative to overall element sequence conservation. The trend of greater conservation of proximal loci was only noticeable in analyses of shared regulatory loci, thus suggesting that prior studies highlighting the relative stability of promoter sequences may have been impacted by the increased reutilization of promoter elements. Our observation of shared regulatory elements across three developmental stages supports this view, with proximal elements showing a high degree of reutilization across all three stages, and with reutilization being greatest at later developmental stages. In fact, combining our results for both proximal and distal elements at different developmental stages suggests that the choice of developmental stage plays a significant role in observed evolutionary trends. In conclusion, our data demonstrate the importance of tissue specific, multi-stage analyses of regulatory element evolution, and provide an important benchmark for future investigations across all eukaryotic taxa.

Furthermore, these results have profound implications for the often stated proposition that rapid enhancer evolution is a driving force behind morphological change [29]. Our results suggest that such statements must be qualified, as it appears that developmental utility of regulatory loci plays an important role in cis-regulatory turnover.

Experimental Procedures

Short insert, mate pair, and SMRT libraries were constructed using high molecular weight DNA from a single, female *Heliconius erato lativitta* pupa. An initial assembly was produced using Allpaths-LG [62], resulting scaffolds were merged using HaploMerger [41], and additional scaffolding and gap filling was performed with long read sequences using PBJelly [63]. A Satsuma [64] derived synteny map was used to produce a mostly ordered and oriented assembly of the 21 *H. erato* chromosomes (Table S2.5). Tophat [65] and Cufflinks [66] were used to assemble mRNA-seq data from head and wing tissue at multiple stages into a reference transcriptome. This reference transcriptome and *H. melpomene* protein annotations were used to perform gene annotation on the final *H. erato* genome assembly using three iterations of MAKER [44].

Chromatin immunoprecipitation of prepupal, pupal, and adult head tissue was performed using a SimpleChIP Enzymatic Chromatin IP Kit (Cell Signaling Technology) with modifications, using antibodies to H3K4me3 (Abcam ab8580) and H3K27ac (Abcam ab4729). Sequencing reads were aligned to the reference genome with Bowtie2 [67] and enriched loci, “peaks”, were called using MACS2 [68] (Table S2.6). Final peak sets for

each histone mark and tissue were called from overlapping replicate peak sets using bedtools [69]. Final peak sets for each stage were merged and classified as “proximal” or “distal” using custom python scripts. Comparison of developmental stages was performed using bedtools and bedops [70]. GO enrichment of neighboring genes to stage-specific regulatory elements was determined using the PANTHER database [71].

A reciprocal best-hit BLAST algorithm was used to perform conservation analysis of *H. erato* regulatory loci in five other lepidopteran genomes. A null model of expected sequence conservation was produced using a custom python script. Analysis of null model loci was performed identically to that of annotated regulatory elements. A custom python script was used to identify conserved elements across various taxonomic clades. Adjusted conservation scores were determined following a process similar to that used by Taher et al. [56] to identify non-aligning, functionally conserved elements. See also Supplemental Experimental Procedures.

Custom scripts used for assembly and data analyses are available at butterflygenome.org.

Chapter 3:
Chromatin profiling reveals that genome-wide regulatory adaptation shapes population-
level genomic landscapes in *Heliconius* butterflies

James J. Lewis* & Robert D. Reed

Department of Ecology and Evolutionary Biology, Cornell University.

Introduction

cis-Regulatory variants have been widely implicated as causal elements in numerous ecological, morphological, behavioral, and sexual adaptations in a wide range of eukaryotes [30, 72, 73]. This is consistent with evidence that allelic variation in gene regulatory loci is a major mechanism for variation in transcription factor binding, histone modification, and gene transcription [27, 28, 74, 75]. Unfortunately, we still have a poor understanding of how local selection pressures and introgression can tailor population-specific regulatory landscapes, thus resulting in genomic and transcriptomic divergence that allow populations to adapt to local conditions. Moreover, much of our limited knowledge of population-level regulatory variation fails to incorporate both chromatin accessibility—a necessary but insufficient biochemical state for most regulatory element utilization where a locus becomes accessible for transcription factor binding—and regulatory element activity, the functional, transcription factor-bound state of an accessible regulatory locus [76]. Studies of regulatory variability in human lymphoblastoid cell lines suggests that some regulatory variation is associated with ancestry [28], raising the possibility that reproductively isolated populations could undergo rapid regulatory diversification as a mechanism of localized adaptation. Here we use *Heliconius erato* clade butterflies, which have a well-understood population structure [77], as a model to study the relationship between variation in chromatin organization and regulatory activity, and to test the hypothesis that regulatory variation can shape the population-level genomic landscape of a species.

Heliconius butterflies are well known for their Mullerian mimicry rings, where multiple species converge on locally adapted wing phenotypes as a shared warning signal to predators, producing sharp population boundaries driven by wing phenotype [78]. *H. erato* radiated throughout South and Central America approximately 2-4M years ago, rapidly diversifying into dozens of named morphs [77, 79, 80]. We study three populations of neotropical *H. erato* clade butterflies—*H. erato petiverana*, *H. erato lativitta*, and *H. himera*—to determine how population boundaries and limited introgression have driven localized regulatory adaptation between allopatric and parapatric populations of *Heliconius* butterflies. *H. e. petiverana* inhabits much of Central America, while *H. e. lativitta* is found in the western Amazonian Basin. *H. himera*, an incipient species nested within the *H. erato* radiation, inhabits more arid, high elevation locales in the western Andean regions of Ecuador and Peru [78, 81]. The *H. erato* clade radiation appears to have proceeded along geographical boundaries, with phylogenetic analysis of showing an early separation of Central American (including *H. e. petiverana*), east Amazonian, and west Amazonian races (including *H. e. lativitta*), with the incipient species *H. himera* nested early within the west Amazonian phylogeny [80, 82]. Direct introgression occurs between *H. e. lativitta* and *H. himera* in three narrow hybrid zones in Andean valleys [81], while indirect introgression likely occurs between *H. e. petiverana* and the two South American populations via hybridization chains incorporating geographically interposed races of *H. erato*.

We provide evidence of widespread, population-level variation in both chromatin accessibility and chromatin activity between parapatric and allopatric populations of the

H. erato clade derived from ATAC-seq [12] and ChIP-seq assays for H3K4me3 and H3K27ac histone modifications, the latter of which tend to associate with active promoter and enhancer loci, respectively [15]. We provide tissue-specific assays of the *H. erato* clade chromatin landscape for both hindwings and forewings at mid- and late- pupal development. We find that regulatory variation between populations has evolved separately between wing tissues and developmental stages, with distinct evolutionary patterns associated with each regulatory assay, and provide strong evidence for regional variation in regulatory loci as a major force driving genomic population divergence and local adaptation.

Population level regulatory accessibility and activity

Our first aim was to identify the degree of regulatory divergence between distinct, reproductively isolated *H. erato* populations. We determined population-level variability in chromatin accessibility and regulatory activity profiles for three populations of *H. erato* clade butterflies (Figure 3.1) using ATAC-seq and ChIP-seq for H3K4me3 and H3K27ac, respectively, for mid- and late-pupal stage forewings and hindwings.

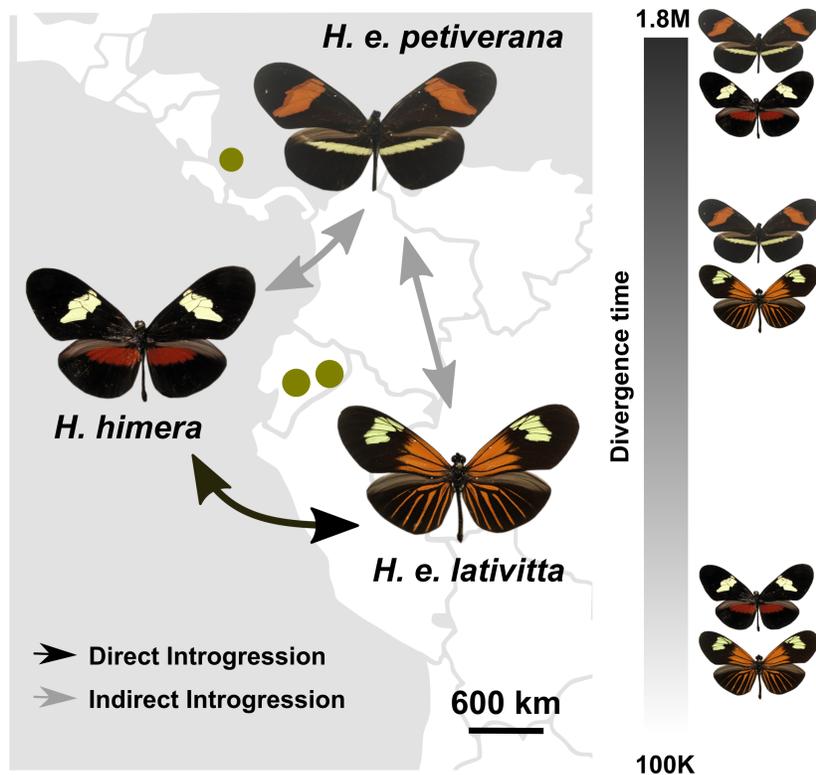


Figure 3.1. Population locales and introgression patterns of *Heliconius erato* clade butterflies. Three populations were used in this study: *H. e. petiverana* (Costa Rica), *H. e. lativitta* (Ecuador), and incipient species *H. himera* (Ecuador). Black arrow indicates direct introgression between *H. e. lativitta* and *H. himera*, gray arrows indicate indirect introgression between *H. e. petiverana* and the Ecuadorian populations. Approximate divergence time for each population pair derived from [82].

We performed two biological replicates of each assay, including ChIP-seq input controls, for a combined set of 24 ATAC-seq, and 72 ChIP-seq and control experiments (Figure S3.1-S3.3). In total, we called 258,698 unique loci with open chromatin, or transposase accessible sites (TASs), as well as 89,343 and 141,958 unique H3K4me3 and H3K27ac peaks, respectively. Mean peak call counts for each tissue and developmental stage for TASs, H3K4me3 and H3K27ac marks were 212,189, 42,046, 58,380. We observed a high degree of regulatory element reutilization between tissues and stages, with an average of 147,514 (69.5%), 19,710 (46.9%), and 22,890 (39.2%) peaks for TASs, H3K4me3, and H3K27ac marks shared between one or more tissues or developmental

stages. TASs, and to a lesser degree, H3K4me3 sites, which are often associated with transcription start site (TSS-) proximal promoter elements [15, 83], were notably more likely to be shared with one or more additional tissues or stages. This highlights previously observed patterns of molecular divergence between accessibility and activity of regulatory loci [76] and a tendency towards tissue and developmental stage specificity of H3K27ac marks, which are often associated with TSS-distal enhancer activity [15, 84].

Population-level variability of TASs, and H3K4me3 and H3K27ac marks was both abundant and widespread throughout the genome (Figure 3.2a-c).

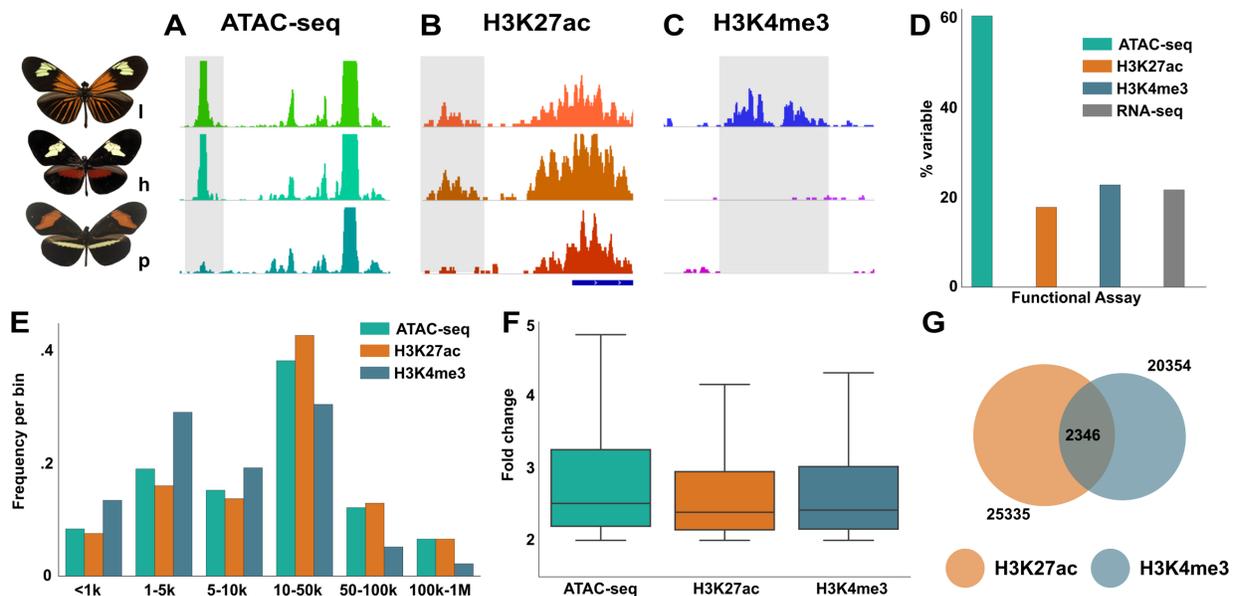


Figure 3.2. Variation in regulatory loci between three populations of *H. erato*. **A-C**, Examples of regulatory loci displaying population level variability for ATAC-seq (**A**), H3K27ac ChIP-seq (**B**), and H3K4me3 ChIP-seq (**C**). For all panels: p = *H. e. petiverana*, l = *H. e. lativitta*, h = *H. himera*. (**D**), Percent of all unique regulatory elements displaying variability between one or more populations by assay. (**E**), Distance of variable regulatory loci from the nearest TSS for ATAC-seq, H3K27ac, and H3K4me3 marked loci. (**F**), Boxplots showing fold change between mean population signals for all variable sites in ATAC-seq, H3K27ac, and H3K4me3 ChIP-seq comparisons. Whiskers removed due to very high fold change in the top 1% of datasets representing presence/absence of peaks. (**G**), Overlap between all unique variable H3K27ac and H3K4me3 peaks.

Using a threshold of 2-fold or greater mean signal difference between populations, we identified 156,582 (60.5%), 20,354 (22.8%), and 25,335 (17.8%) TAS, H3K4me3, and H3K27ac variable loci across all tissues and stages, respectively (Figure 3.2d, Figure S3.1-S3.4). This places variable TAS, H3K4me3, and H3K27ac loci every 2.65kb, 20.39kb, and 16.37kb in the *H. erato* genome, on average. Fold change distributions for variable regulatory loci were heavily skewed for all three signal types, with 25% of all variable sites displaying very high signal variability between population (greater than 3-fold change, Figure 3.2e). In general, both forewings and hindwings showed a similar degree of regulatory differentiation by population for all three assays, and variable regulatory loci for all data types appear to be drawn non-randomly from the pool of all regulatory elements of the same type (Figure S3.5-S3.6). Regulatory variability between *H. erato* clade butterflies was substantially greater than previously observed between human ethnic groups, and appears on par with estimates of regulatory divergence between species isolated by more than 10 million years. TAS variability, for which there is currently no comparable dataset, was much higher over both intra-specific and inter-specific comparisons of regulatory loci (Figure S3.7). Interestingly, we noticed a significant increase in TAS variability in both forewings and hindwings during the late-pupal stage, while there was not a substantial change in histone mark variability (Figure S3.6). To provide a reference for our observed population level variability, we performed three biological replicates of mRNA-seq for mid-pupal forewings and hindwings for each population. Of the 14,617 annotated genes in the *H. erato* reference assembly [85], 3,165 (21.7%) genes were differentially expressed in the same tissue between one or more

populations (Figure 3.2d). This compares favorably with the combined observations of histone mark variability between populations, which are indicative of population specific variation in gene regulation [86], and suggests that much of the observed population structure in regulatory activity is functional.

Observations of regulatory variability were spatially distributed roughly as expected with variable H3K4me3 marks showing a larger TSS-proximal frequency relative to variable H3K27ac marks, and with variable TSSs being distributed approximately equal to the combined distributions of both active histone marks (Figure 3.2f). H3K4me3, which has been shown to mark highly active TSS-distal enhancer loci in addition to TSS-proximal promoter regions [48], was also enriched for variable sites in the same distal categories as H3K27ac, leading us to posit that many variable TSS-distal regulatory loci are strong enhancer elements. In support of this view, 2,346 (11.5%) variable H3K4me3 loci overlapped a variable H3K27ac locus (Figure 2g), with 1,054 (44.9%) of these being >10kb from the nearest annotated TSS, a notable increase over the 37.9% of all variable H3K4me3 marks falling into this range.

Regulatory activity is evolutionarily decoupled from chromatin accessibility

Chromatin accessibility is necessary for all but a few transcription factor binding events, with the exceptions being limited to pioneer factor binding associated with induction of an accessible state [17, 76]. Thus we next aimed to identify the relationship between variation in chromatin accessibility and regulatory activity. To discern whether variable

histone mark loci were dependent on variability in chromatin accessibility, we assessed the co-occurrence of variable TASs and histone marks for each tissue and developmental stage. Overall, TAS and histone marked peaks were highly similar (Figure 3.3a-b), with 91% and 95% of H3K27ac and K3K4me3 peaks overlapping TAS (ATAC-seq) peaks, respectively (Figure 3.3c).

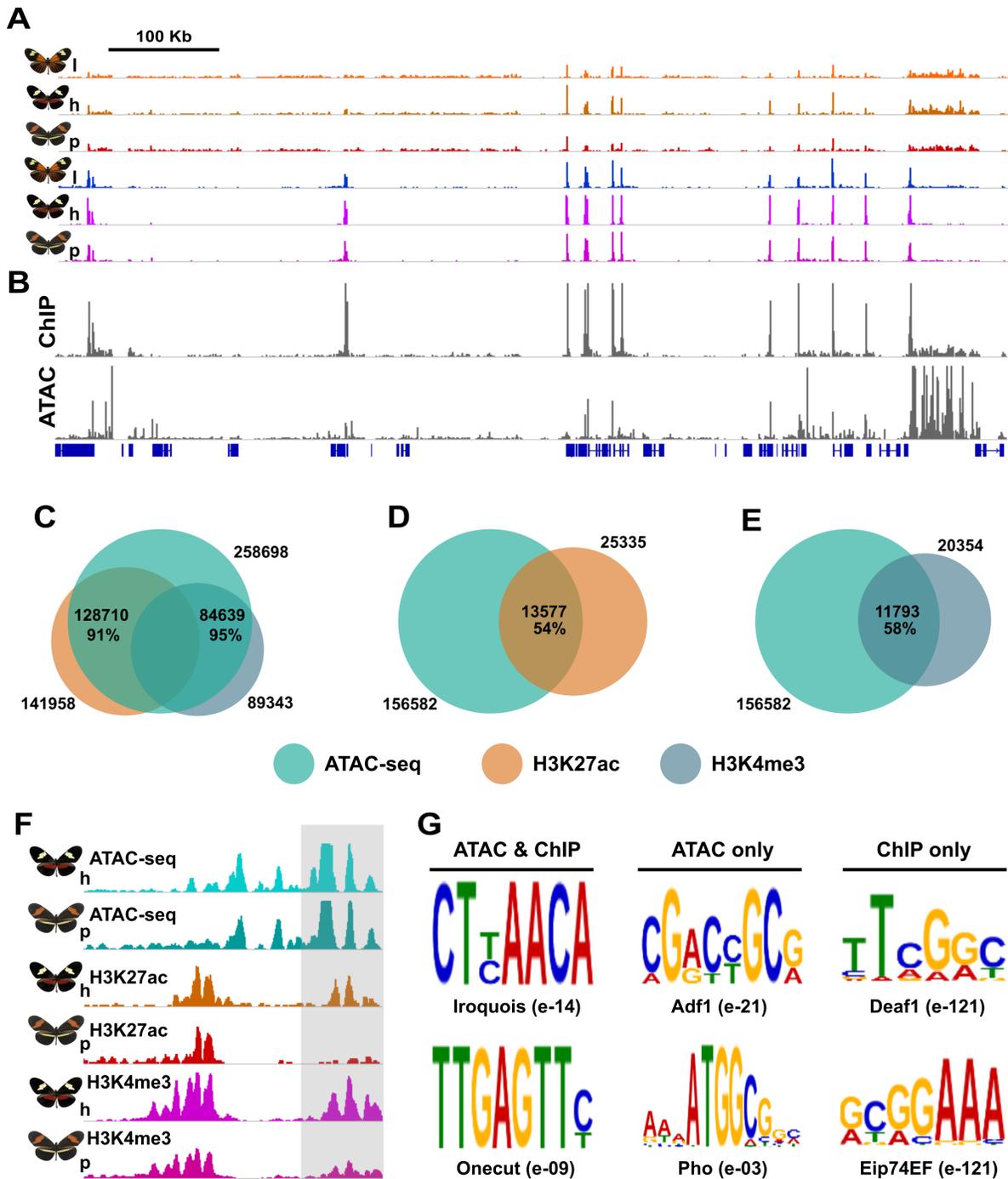


Figure 3.3. ChIP-seq signal recapitulates chromatin accessibility and indicates decoupled evolution of chromatin accessibility and regulatory activity. $p = H. e. petiverana$, $l = H. e. lativitta$, $h = H. e. himera$. **(A)**, ChIP-seq tracks for mid-pupal forewings showing H3K27ac (top, oranges) and H3K4me3 (bottom, purples) across approximately 900kb of chromosome 19. **(B)**, Combined enrichment profiles for all mid-pupal histone ChIP-seq (top) and ATAC-seq (bottom) tracks shows a high degree of similarity between ATAC-seq and ChIP-seq signals across all loci. This is confirmed in **(C)**, with 91% of all H3K27ac and 95% of all H3K4me3 ChIP-seq peaks overlapping an ATAC-seq peak by

at least 1bp. Limited overlap of variable H3K27ac (**D**) and H3K4me3 (**E**) ChIP-seq peaks with variable ATAC-seq peaks indicates decoupled evolution of chromatin accessibility and regulatory activity at cis-regulatory loci in the *H. erato* clade. (**F**) Representative locus (gray) demonstrating variability in H3K27ac and H3K4me3 signal without change in chromatin accessibility. (**G**) Enriched transcription factor binding site motifs in discriminative analyses of regulatory loci variable for both ATAC-seq and ChIP-seq, ATAC-seq only, and ChIP-seq only signals.

Co-occurrence of variable histone marks and TASs were quite similar for both active marks (Figure 3.3d-e), with an average of 1,519 (24.8%) and 1,510 (22.1%) variable H3K4me3 and H3K27ac marks overlapping a variable TAS locus, respectively, in each tissue and developmental stage studied. Overlapping variable histone marks were often specific to a single stage or tissue, as the overlap of all variable histone loci and TASs was 11,793 (57.9%) and 13,577 (53.6%) for H3K4me3 and H3K27ac, respectively. Interestingly, variable forewing histone mark loci showed consistently less co-occurrence with variable TAS loci than did hindwing loci in both mid- and late-pupal stages, with variable late-pupal histone marks displaying the greatest disparity between tissues (mean decreases for forewings and hindwings were 4.9% and 9.2% in mid- and late-pupae, respectively). Comparison between developmental stages showed that variable late-pupal histone marks overlapped variable TAS loci to a much greater degree than mid-pupal elements (mean increase was 11.4% in late-pupal loci).

While a moderate number of variable histone mark loci co-occur with variable TASs, a large fraction of variable H3K4me3 and H3K27ac loci occur independent of variation in chromatin accessibility (Figure 3.3f), giving strong evidence for divergence at both levels of gene regulation. Moreover, the degree of divergence between variability in accessibility and activity is dependent on both tissue and stage, leading us to posit that evolutionary

divergence of chromatin accessibility and transcription factor binding at accessible sites has been decoupled. We performed a discriminative transcription factor binding site (TFBS) enrichment analysis to determine sequence-specific factors associated with decoupled variability in ChIP-seq and ATAC-seq signals (Figure 3.3g). Regulatory loci showing both ChIP-seq and ATAC-seq variability were enriched for onecut and Iroquois complex transcription factors, which have been shown to be differentially expressed between some populations [87]. Loci with only variable ChIP-seq signal and no detectable difference in accessibility were enriched for TFBSs associated with known wing-related phenotypes such as Deaf1 and Eip74EF, a factor responding to ecdysone signaling. Loci displaying only variability in ATAC-seq signal were enriched for Adf1 and Pho TFBSs, regulators known to bind at Polycomb response elements [88], indicating that variability in chromatin accessibility includes a number of repressed loci in addition to active enhancer and promoter regions.

Regulatory variants underlie local adaptation

We next directly tested the expectation that regulatory divergence is associated with population divergence. Under a model of neutral divergence, we would expect to see the same patterns of regulatory variability in each of our three functional assays. To determine whether local adaptation and introgression could selectively shape population-level regulatory landscapes, we took the variable TAS and histone mark loci from each population and identified those shared between population pairs for all three pairwise population combinations (Figure 3.4a-c, Figure S3.8).

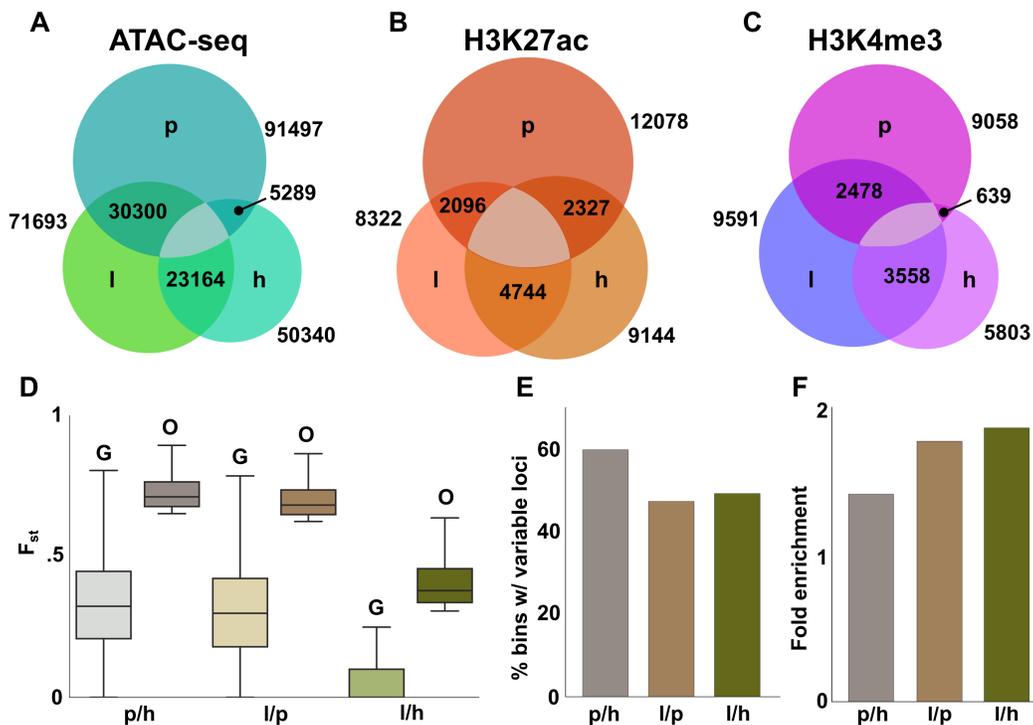


Figure 3.4. Population structure and evidence of selection around variable regulatory loci. Variable ATAC-seq (A), and H3K27ac (B) and H3K4me3 (C) ChIP-seq loci show different patterns of divergence between populations, contrary to the neutral expectation. Gray areas indicate that the union set does not exist. (D), Genome wide F_{st} for all 5kb bins (marked “G”) and for the top 5% outlier bins (marked “O”) for all three population comparisons shows a large increase in divergence in outlier F_{st} bins. (E), Percent of outlier bins with variable regulatory loci for all three population comparisons. (F), Fold enrichment of variable regulatory elements in outlier bins relative to the average enrichment in all bins for all three population comparisons. For all panels: p = *H. e. petiverana*, l = *H. e. lativitta*, h = *H. himera*.

Combining tissues and stages, 71,693 variable TASs were present in *H. e. lativitta*, with 91,497 and 50,340 variable TASs in *H. e. petiverana* and *H. himera*. Of these, 23,164 (32.3% *lativitta* total / 46.0% *himera* total) were shared between *H. e. lativitta* and *H. himera*, 30,300 (42.3% *lativitta* total / 33.1% *petiverana* total) were shared between *H. e. lativitta* and *H. e. petiverana*, and 5,289 (10.5% *himera* total / 5.8% *petiverana* total) were shared between *H. himera* and *H. e. petiverana* (Figure 3.4a), reflecting the greatest evolutionary divergence between *H. himera* and *H. e. petiverana* relative to the reference

assembly used (*H. e. lativitta*). These results also suggest a significant degree of local adaptation of the regulatory landscape, with even the most distantly related pair (*H. e. petiverana* and *H. himera*) displaying thousands of shared regulatory loci absent in the reference. Similar pairwise population analysis of variable H3K4me3 and H3K27ac loci showed different patterns. 3,558 (37.1% *lativitta* total / 61.3% *himera* total), 2,478 (25.8% *lativitta* total / 27.4% *petiverana* total), and 639 (11.0% *himera* total / 7.1% *petiverana* total) variable H3K4me3 loci (Figure 3.4b) were shared between *H. e. lativitta* and *H. himera*, *H. e. lativitta* and *H. e. petiverana*, and *H. himera* and *H. e. petiverana*, respectively. In turn, 4,744 (57.0% *lativitta* total / 51.9% *himera* total), 2,096 (25.2% *lativitta* total / 17.4% *petiverana* total), and 2,327 (25.4% *himera* total / 19.3% *petiverana* total) variable H3K27ac loci (Figure 3.4c) were shared between the same populations. In sum, patterns of shared regulatory variability were distinct for each signal type assayed (i.e. ATAC-seq, H3K27ac ChIP-seq, and H3K4me3 ChIP-seq) despite identical population demographics. Under a neutral model, where regulatory variants are sampled randomly from each population, we would expect similar distributions of shared regulatory variants for each data type. Thus we hold that the dissimilar patterns of regulatory divergence between signals indicates that at least two (allowing for the possibility of one neutrally diverging signal), but likely all three, regulatory signals provide evidence of non-neutral variation in regulatory activity between populations.

To further test whether population level regulatory variants were associated with population divergence via natural selection and local adaptation, we used F_{st} plots from whole-genome sequencing data to identify genomic loci displaying elevated levels of

nucleotide divergence for four *H. e. lativitta*, six *H. himera*, and five *H. e. petiverana* individuals. In total, 7,749,900 variants were called across all races, and F_{st} was calculated with quality filtered nucleotide variants for pairwise population comparisons in bins of 5kb across the genome. Mean F_{st} values for pairwise comparisons between *H. e. lativitta* and *H. himera*, *H. e. lativitta* and *H. e. petiverana*, and *H. himera* and *H. e. petiverana* were .013, .299, and .325 (Figure 3.4d, marked “G”). Estimated population divergence was consistent with previously observed divergence between parapatric and allopatric races in both *H. erato* and *H. melpomene* clade butterflies [78, 89]. We selected the top 5% (3,971 bins) of 5kb bins ranked by F_{st} as “outlier bins” (Figure 3.4d, marked “O”). Outlier bins were distributed across 138 of the 142 scaffolds, including previously identified wing color pattern associated loci. To identify potential regulatory variability associated with high levels of population divergence, we tested the number of outlier bins containing variable regulatory loci and their frequency within these bins using the combined set of variable regulatory loci from all developmental stages and tissues from pairwise population comparisons corresponding to the pairwise F_{st} scans. 1,953 (49.2%), 1,878 (47.3%), and 2,375 (59.8%) outlier bins contained variable regulatory loci in *H. e. lativitta/H. himera*, *H. e. lativitta/H. e. petiverana*, and *H. himera/H. e. petiverana* comparisons (Figure 3.4e). Mean F_{st} values for outlier bins containing variable regulatory loci were .417, .708, and .737. These outlier bins were enriched in variable regulatory elements relative to the genome wide mean, with 2.36, 2.45, and 3.29 variable TAS, H3K4me3, or H3K27ac loci per bin, on average—1.86, 1.77, and 1.41 fold increases over genome wide averages of 1.27, 1.38, and 2.33 variable sites per bin, respectively (Figure 3.4f). Thus, we find that approximately half of genomic loci showing the greatest

population level divergence are enriched for variable TAS, H3K4me3, or H3K27ac loci in wing tissue, providing strong evidence for natural selection on variable regulatory elements as a major force shaping the genomic landscape.

Regulatory variants are associated with gene network evolution

To investigate whether regulatory variants were associated with divergence in any well-studied developmental pathways we mapped genes associated with variable regulatory loci to the KEGG pathway map [90]. While linking regulatory loci to gene expression on a genome-wide scale remains a challenge for molecular and computational biology, presence of regulatory histone marks, such as H3K4me3 and H3K27ac, in a gene's promoter region has been strongly correlated with regulation of that gene (Figure 3.5a).

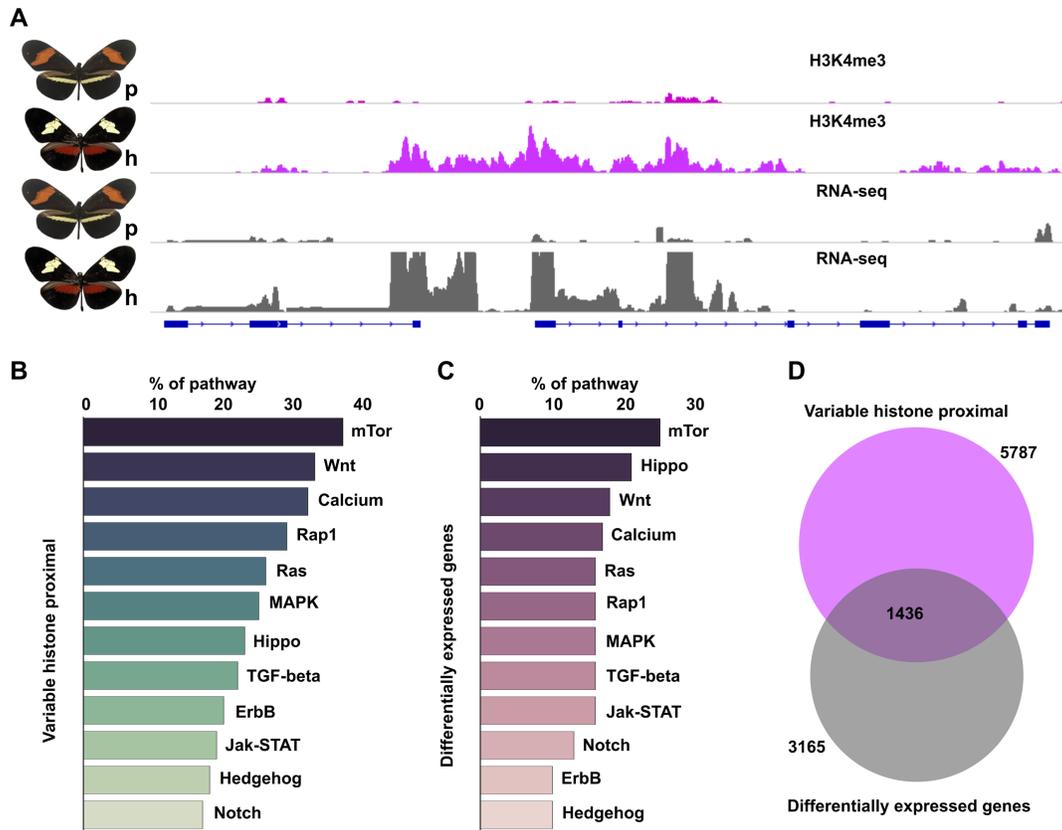


Figure 3.5. Population-level regulatory variation associated with variation in gene activity. (A), Evidence of variable regulatory activity driving change in gene expression between *H. erato petiverana* and *H. himera* on chromosome 15. p = *H. e. petiverana*, h = *H. himera*. (B-C), KEGG pathway enrichment for genes with variable histone peaks less than 2kb from the TSS (B) and differentially expressed genes determined via RNA-seq (C) shows very similar divergence between the two gene sets. (D), In support of the similarity between (B) and (C), approximately half of all differentially expressed genes have a variable histone peak within 2kb of the TSS.

We identified 5,787 genes with variable histone marks within 2kb of the TSS and used BlastKOALA [91] to attain KEGG pathway annotations (Figure 3.5b). We then performed the same BlastKOALA analysis for our 3,165 differentially expressed genes (Figure 3.5c), and identified major developmental pathways associated with both analyses. Divergence in KEGG signaling pathways was highly similar between both gene sets (Figure 3.5b-c, Figure S3.9). In support of a likely causal relationship between gene expression divergence and histone variability around gene TSSs, 1,436 genes with TSS-proximal

variable histone peaks were also differentially expressed genes in our mRNA-seq analysis (45.4% of differentially expressed genes). In both gene sets, multiple components of the Ras, Rap1, MAPK, Wnt, Notch, Hedgehog, TGF-beta, and Hippo signaling pathways, were identified as being notably associated with both variable regulatory loci and gene expression variation. This suggests that many fundamental developmental processes implicated in wing and scale cell development, such as cell proliferation and differentiation, wing morphogenesis, and determination of cell polarity [92], have undergone local adaptation in *H. erato* populations (Figure S3.9).

To further assess these findings, we used MEME-ChIP [93] to characterize enriched motifs within 300bp on either side of the center of variable TAS loci. We discovered 116 enriched motifs, of which 106 have been previously reported and 10 are potentially novel binding motifs. Unsurprisingly, Trithorax-like, or GAGA factor, a known trithorax-group regulator of chromatin accessibility via association with the NURF ATP-dependent chromatin remodeling complex, was highly enriched in variable TAS elements. Other enriched motifs include known binding motifs for Mothers against dpp (Mad), Hairless (H), Brinker (Brk), Dorsal (dl), and Cubitus interruptus (Ci), which play important roles in TGF-beta (Mad, Brk, Dorsal), Notch (H), and Hedgehog (Ci) signaling during development. Thus we see strong independent evidence for selection on many of the same major signaling pathways identified in the KEGG analysis above.

Discussion

Recent population genetic work in *H. erato* has largely focused on three major mimicry-related color pattern loci and has portrayed these regions as “hotspots” of genomic adaptation set in a mostly free-flowing genomic landscape [78, 94]. This view has been supported by genome-wide sequence comparisons in racial hybrid zones that have identified only a handful of divergent genomic regions, including the previously identified color pattern mimicry loci [van Belleghem 2017, in revision]. Association mapping of neighboring *H. erato* hybrid zones in Ecuador and Peru found ten novel genomic intervals on three chromosomes diverging between allopatric populations. Of these ten intervals, only two appeared unlinked to previously identified color pattern loci [95]. Leveraging the power of functional assays for chromatin accessibility and chemical modifications of histone H3 tails—biochemical indicators of regulatory elements presence and activity—we show that functional genomic divergence is actually ubiquitous and widespread throughout the *H. erato* genome, even between hybridizing populations of *H. erato* clade butterflies. Moreover, we tie variability at regulatory loci to regions of elevated population divergence between both allopatric and parapatric populations of the *H. erato* group despite evidence of both direct and indirect introgression between these populations. Our results suggest that local ecological conditions and purifying selection on mimicry-related wing color patterns between populations has been a driving force shaping the genomic landscapes of these three *H. erato* populations. Furthermore, we suggest that the *H. erato* genome is likely under much greater local selection throughout than would have been previously predicted by hybrid zone-focused studies, and that we must begin to reassess current models of the genomic mechanisms of adaptation and population divergence.

Understanding the origins of diversification via adaptation to local selective pressures is one of the primary goals of evolutionary and population biology. Interestingly, our results suggest that gene expression, chromatin accessibility, TSS-proximal, and TSS-distal active regulatory elements are all under genome-wide divergent selection between regional morphs of the *H. erato* clade. We found evidence of population-, tissue-, and stage-specific regulatory variation, indicating a highly complex adaptive landscape. Surprisingly, at many loci chromatin accessibility and regulatory activity appear to be decoupled from one another, thus showing that there are higher order mechanisms of gene regulatory evolution beyond the simple gain and loss of transcription factor binding sites. Our findings highlight the need for further comparative functional genomic work at the population level to refine our understanding of how selection and adaptation intertwine with complex regulatory architectures to determine the genomic landscape of species.

Materials and Methods

***Heliconius* stocks and tissue sampling**

All samples of *Heliconius erato lativitta*, *Heliconius erato petiverana*, and *Heliconius himera* were taken from laboratory colonies at Cornell University derived from individuals collected from Ecuador (*H. e. lativitta* and *H. himera*) and Costa Rica (*H. e. petiverana*). All colonies were pure for relevant color pattern elements. Mid-pupal samples were collected from individuals reared for 72 hours (3 days) at approximately 30 C, and were phenotyped for the emergence of early wing scale buds. Late-pupal samples were

collected as “ommochrome” stage pupae, approximately 7 days post-pupation at 30C, and were phenotyped for ommochrome pigment deposition in the wing and eyes without any signs of melanin pigmentation.

ChIP-seq, input control, and ATAC-seq sample preparation

ChIP-seq was performed as previously described [85], with minor modifications. For each population, developmental stage, tissue, and biological replicate, wing pairs (left and right) from 4-6 individuals were fixed for 5 minutes with 1% freshly prepared formaldehyde, quenched for 5 minutes with 1M glycine solution to a final concentration of .125M, rinsed with 2 washes of cold PBS, then combined prior to tissue homogenization. Post-extraction nuclear samples were incubated for approximately 12 and 13 minutes with .5ul micrococcal nuclease at 37C before adding EDTA to quench digestion. Digested nuclear preps were then split into 3 aliquots for immunoprecipitation and input control prep. Chromatin immunoprecipitation was performed with antibodies to H3K27ac and H3K4me3 (Abcam: ab4729 and ab8580), using 3-5ug of digested chromatin per immunoprecipitation. Libraries were prepared with the NEB DNA Ultra library prep kit using approximately 40ng of input, and amplified for 14 cycles prior to agarose gel size selection.

ATAC-seq was performed as described in [12], with minor modifications. Tissue dissection was performed as previously described [85]. For all samples, nuclear extractions were performed on freshly dissected (<15 minutes) wing pairs from a single

individual, which were dounce homogenized in sucrose buffer using pestle B (mid-pupal samples) or pestles A and B (late-pupal samples). Nuclei were counted using a hemocytometer, and approximately 400,000 nuclei were isolated for each ATAC-seq library prep. Libraries were amplified for 10 cycles, and size selected on an agarose gel for fragments between 35-1000bp.

Sequencing for ChIP-seq, input control, and ATAC-seq libraries was done on a NextSeq 500 at Cornell University using 2x37bp PE reads. ChIP-seq, input control, and ATAC-seq libraries were sequenced to a minimum depth of 20M, 30M, and 50M paired reads, respectively (See SD1).

Read alignment and peak calling

Read alignment and filtering for ChIP-seq, input control, and ATAC-seq samples was performed as previously described [85]. Briefly, raw sequence reads were aligned using Bowtie2 and filtered for uniquely mapping pairs with a custom python script. ChIP-seq peaks were called using MACS2 [96] for each biological replicate using combined input from both input control replicates to avoid variation in enrichment profiles between replicates due to minor differences in MNase digestion and library size selection. ATAC-seq peaks were called using F-seq [97].

Analysis of population variation in ChIP-seq and ATAC-seq peaks

Population specific data sets were determined as follows: peak calls from biological replicates for each tissue, developmental stage sample, and data type were combined and merged using bedtools with duplicate peaks removed if they overlapped by 147bp, or 1 nucleosome, (ChIP-seq) or 50bp (ATAC-seq). Population level peak sets for each data type tissue, and developmental stage, were merged as described above to produce species wide tissue and developmental stage specific peak sets for subsequent comparison between populations.

To account for peak calling algorithms using strict FDR thresholds that often fail to call shared peaks between even replicate datasets, we performed population-level comparisons of all regulatory loci by individually testing every species-level peak call in each population comparison. Thus, peaks were identified using a FDR thresholded peak calling process, then populations were compared using a conservative approach comparing signal tracks for each population as follows: Alignment files for each dataset were converted to bedgraph format, then normalized by read depth (RPM normalization) to produce normalized raw signal tracks for each sample. To test for population-level variability at regulatory loci, we performed pairwise comparisons of peak signals for individual regulatory loci between populations. To determine whether a regulatory locus was variable, we first took the maximum read depth normalized (RPM) raw signal score, or the argmax value from the vector representing read depth at each position in the peak, from each biological replicate. We then identified loci where the RPM normalized maximum signal (argmax value) between biological replicates was less than 50% of the largest value within a population, and the mean signal difference between populations

was greater than 200% of the smallest population mean value. Peak loci positively identified according to this process were considered variable. Thus, we sampled loci with less than 2-fold signal variation within populations and greater than 2-fold mean signal variation between populations. This process was chosen over alternate depth comparison analyses, such as DESeq2 used below for analysis of mRNA-seq data, to explicitly account for the expected read distribution within peaks (e.g., a “peak” with tails) while allowing for slight variation in local peak maxima across populations. Moreover, this analysis was designed to minimize batch effects inherent in all high throughput sequencing assays. Finally, this approach was shown to be conservative relative to an alternative statistical approach (See below). Subsequent analysis of population level variable peak sets was performed using bedtools and linux command line utilities.

Validation of our regulatory variability metric

We aimed to provide a robust, conservative metric for population variability in regulatory loci that was easily interpreted. Thus, we adopted a 2-fold difference metric based on related ENCODE standards [98]. To verify the stringency of our metric, we tested all loci from our mid-pupal forewing ATAC-seq datasets from *H. e. lativitta* and *H. himera* (the two most related populations) using a Student’s t-test. As expected from an assay capturing the product of multiple random variables, ATAC-seq signal at all loci tested was lognormally distributed (See SD3.2). We log transformed the signal values at each locus to meet the assumption of normality, and performed a simple t-test for significant deviation between mean signals between populations. This approach has been used previously to

detect statistical differences between ChIP-seq and ATAC-seq datasets, e.g. [99], and is analogous to tests used in [6]. The results of this test were then FDR corrected using the Benjamini-Hochberg method, leading to 61,189 (32.6%) variable loci between the two populations at an FDR of 0.1. As the results of our 2-fold difference metric produced only 16,917 variable loci in the same comparison, our test was deemed more conservative than this alternative statistical approach.

To further validate the quality of our results, we used a low-powered non-parametric two-sample Kolmogorov-Smirnov test (K-S test) to test only the loci passing the filters built into our 2-fold difference metric. While high throughput datasets are often modeled with a discrete negative binomial or poisson distribution, we believe that once RPM normalized, high throughput count data can be properly modeled with a continuous probability distribution. If this assumption is not met, however, the K-S test has been shown to be more conservative rather than less so in identifying significant deviation between distributions [100]. This property again makes the K-S test ideal as a critical test of our variability metric. Results from the K-S test were highly concordant with those from our own 2-fold change metric. Application of the K-S test ($\alpha = 0.1$) identified 16,514 variable loci, representing a 97.6% overlap between the two variability metrics.

RNA-seq sample preparation

RNA-seq was performed on mid-pupal forewings and hindwings as previously described [85], with 3 biological replicates for each tissue from each population (for a total of 18

samples). Samples were sequenced on a Nextseq 500 to a minimum depth of 18M paired reads (See SD3.1).

Differential gene expression analysis

RNA-seq data was aligned to the reference assembly using Tophat2 [101]. Read counts for each annotated gene were determined using HTSeq [102], and differentially expressed genes were identified using DESeq2 [103] as prescribed, with an adjusted p-value cutoff of “0.01”.

Whole genome resequencing sample preparation

DNA was extracted from 4 *H. e. lativitta*, 6 *H. himera*, and 5 *H. e. petiverana* samples with a DNeasy kit following the manufacturer’s guidelines. Sequencing libraries were prepared using the Nextera DNA Library Prep kit following manufacturer’s guidelines and sequence on a NextSeq 500 at 2x37bp PE reads. Each sample was sequenced to a minimum of 2x coverage (See SD3.1).

SNP calling and F_{st} Analysis

Whole genome sequencing samples were aligned to the reference assembly using Bowtie2 [104]. Aligned read files were sorted using samtools, followed by duplicate read marking and read group addition using picardtools 2.1.1 “MarkDuplicates” and

“AddOrReplaceReadGroups” functions. Raw variant VCF files were produced for each sample using GATK [105] “HaplotypeCaller”. Joint genotyping was performed using GATK “GenotypeGVCFs” with “-stand-emit-conf” set to “30”. To remove variants with hypercoverage, low coverage, low quality, and strand biased variant calls, the joint genotype variant file was filtered using GATK “VariantFiltration” with the following setting: “--filterExpression "DP<5||DP>500||QD<2.0||FS>60||MQ<20.0||MQRankSum < -12.5 || ReadPosRankSum < -8.0". This process removed approximately 250,000 variants.

Fst analysis was performed for pairwise population comparisons using 5kb bins across the genome with VCFtools [106] “--weir-fst-pop” function with “--fst-window-size” and “--fst-window-step” set to “5000”. All subsequent Fst analysis was performed using the unweighted “Mean Fst” column. Identification of top 5% outlier bins and analysis of regulatory variants in outlier bins was performed using bedtools and linux command line utilities.

Discriminative motif analysis

A custom script was used to extract 300bp around the center of each ATAC-seq peak for annotated regulatory loci sets: Loci with both variable ATAC-seq and ChIP-seq signal, loci with only variable ATAC-seq signal, and loci with only variable ChIP-seq signal overlapping an invariant ATAC-seq peak. To determine motif enrichment for each class, MEME-ChIP [93] was run in discriminative mode. The set of loci with variability in both ATAC-seq and ChIP-seq signal were run against regions displaying ATAC-seq variability

only as a background model. Enrichment analyses of both ATAC-seq only and ChIP-seq only variable sites were run against the set of loci showing both ATAC-seq and ChIP-seq variability. Unidentified consensus motifs in the MEME-ChIP output were then curated using the JASPAR insect core database [107].

Gene set and motif analysis

The set of genes with variable regulatory loci within 2kb of the TSS were identified using bedtools and linux command line tools. The differentially expressed gene set was identified as described above. A custom script was used to extract protein fasta files for both gene sets from the complete set of reference assembly peptides. BlastKOALA [91] at www.kegg.jp/blastkoala was used to map both protein sequence files to annotated KEGG pathway genes, followed by manual curation of all listed signal transduction pathways for pathways associated with genes in both data sets. For motif enrichment analysis, the center of each peak was calculated as the point halfway between the start and stop positions. A custom script was used to extract 300bp around each peak center, removing any loci where this region extended beyond the scaffold end (approximately 30 loci). MEME-ChIP [93] was run using this sequence set to identify enriched motifs, after which we manually curated the enriched motif set for overlap with previously identified pathways.

Data availability

Custom scripts and processed CHIP-seq, ATAC-seq, RNA-seq, and WGS data files are available for download at *dryad.org* and download and interactive searching and browsing at *butterflygenome.org*

APPENDIX I:
SUPPLEMENT TO CHAPTER 2

Supplemental Experimental Procedures

Genomic DNA extraction, library preparation, and sequencing

High molecular weight DNA was extracted from a single, ommochrome stage (here defined as the point at which red pigmentation appears on the pupal wing, approximately 6-7 days post-pupation at 30C) female *Hel* pupae (matching adult phenotype shown below) from Ecuador using a Qiagen Genomic Tips Kit with minor modifications. The individual we sequenced was taken from a laboratory stock that originated with individuals collected along a ~150km transect between Puyo and Coca, Ecuador. Tissue was excavated from the pupal case, and highly chitinous leg tissue was removed from the specimen. The specimen was homogenized in the provided lysis buffer using a dounce homogenizer, after which all subsequent steps followed the provided protocol. 220bp, 3kb, 8kb, and 12kb short fragment and mate-pair Illumina sequencing libraries were prepared by the Epigenomics Core Facility at Weill Cornell Medical College. Libraries were combined and sequenced to produce 2 x 150bp paired-end reads on a NextSeq 500 at the Cornell Genomics Facility. Pacific Biosciences SMRTbell libraries were prepared and sequenced by the Genome Sequencing Core Facility at Duke University.



Reference phenotype for *Heliconius erato lativitta* genome assembly.

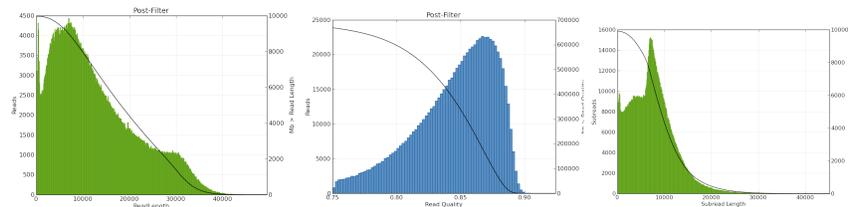
Genome assembly

Initial genome assembly was performed using Allpaths-LG [62] version 49148 as recommended in the provided manual, though with greater coverage of long-jump (8kb & 12kb) mate pair reads and with Haploidify=True and CLOSE_UNIPATH_GAPS=False. The following read depths were used for assembly: 63M (~45x coverage) 220bp short insert & 3kb mate pair reads, 23M (~16.5x coverage) 8kb & 12kb mate pair reads (see table below). As we expected to later regain short genomic sequences via gap-filling with Pacific Biosciences long read sequences, short haplotype scaffolds (less than 7kb, approximately equal to the pre-filter Pac Bio read length, see below) were discarded from the assembly, and the remaining long scaffolds were merged and rescaffolded using HaploMerger [41] version 20120810 with default settings. Prior to running HaploMerger, repetitive elements were masked using RepeatMasker [108] and the *H. melpomene* repetitive element database [109]. The resulting merged and scaffolded assembly roughly met our expectations from previous genome size estimation. We next used 792,825 post-filter Pacific Biosciences reads containing approximately 1Gb of sequences (~24x coverage, mean post-filter read length: 12,621bp, see figure below) for additional scaffolding and gap filling with three rounds of PBJelly following the provided pipeline [63] (PBSuite version 15.8.24). To quality check our assembly, 15x coverage of the short read data used to produce the Allpaths-LG assembly was mapped back to the resulting scaffolds using Bowtie2 [67] to verify assembly quality. Scaffold coverage was then manually observed to verify coverage quality and depth across the assembly.

Library Type	Library Size	Library Stats	Number of Reads	% of Reads Used	Sequence Coverage	Pairs Assembled	Physical Coverage
frag	200bp	-96 +/- 35	126,000,000	83.4	30	51,039,086	24.5
jump	12kbp	12302 +/- 1679	46,000,000	29.5	4	1,881,085	56.2
jump	3kbp	2915 +/- 150	126,000,000	38	14	11,237,382	85.8
jump	8kbp	7963 +/- 438	45,999,996	41.9	5.5	3,069,371	61.2
jump	total		217,999,996	37.3	23.1	16,187,838	203.2

Libraries used in Allpaths-LG genome assembly.

Run Summary			
Number of SMRT Cells	10		
Analysis Configuration	<ul style="list-style-type: none"> ◦ Minimum Polymerase Read Quality= 75 ◦ Minimum Read Length (bp)= 50 ◦ Minimum SubRead Length (bp)= 50 		
Pre-Filter Polymerase Read Bases	Pre-Filter Polymerase Read Length (bp)	Pre-Filter Polymerase Read Quality	Pre-Filter Polymerase Reads
10,860,250,963	7,226	0.462387670002	1,502,920
Post-Filter Polymerase Read Bases	Post-Filter Polymerase Read Length (bp)	Post-Filter Polymerase Read Quality	Post-Filter Polymerase Reads
10,006,819,564	12,621	0.843286599187	792,825



Pacific Biosciences sequence data used for gap filling and scaffolding with PBJelly.

After the initial scaffold assembly, Satsuma [64] version 3.1 was used for synteny analysis (performed as suggested in the software guidelines) between our assembled *Hel* scaffolds and the *H. melpomene* v. 2 genome for which additional linkage mapping had produced a mostly anchored, ordered, and oriented assembly [43]. Syntenous scaffold segments mapping to each *H. melpomene* chromosome were manually curated to produce a syntenous *Hel* assembly map with each scaffold segment anchored to a corresponding *H. melpomene* chromosome. Scaffold segments were then ordered and oriented along chromosomes according to overall consensus with order and orientation of *H. melpomene*. In many cases, where a single large *Hel* scaffold was syntenous with many small unordered or unoriented scaffolds from the *H. melpomene* assembly (loci where the precise assembly path in *H. melpomene* was uncertain), we were able to provide the order and orientation along the chromosome in our *H. erato* assembly. Where appropriate during this process, our *H. erato* scaffolds were broken at presumed misassemblies based on synteny observations, all of which occurred where a single scaffold was mistakenly joined from two or more internal chromosome segments at highly

repetitive loci. The Z chromosome (chromosome 21) was extracted from the initial assembly and gap-filled separately with a single run of PBJelly, and assembled via synteny mapping to the *H. melpomene* Z chromosome as described above. A custom python script was used to extract, order, orient, and rename scaffold segments from the original scaffolded assembly. Scaffold segments for which no synteny information was available were placed into a separate file.

Gene annotation

Total RNA was extracted from 2 day old adult head tissue and prepupal (the larval stage approximately 6-12 hours pre-pupation at 30C, characterized by hanging larvae), day 3 pupal, and ommochrome pupal stage forewings, hindwings, and head tissue samples (see Supplemental Experimental Procedures) preserved in RNA-later at -20C using Trizol Reagent and a Purelink RNA mini kit (purchased separately) as described in the Trizol Plus RNA Purification kit protocol. mRNA-seq libraries were prepared using NEB Ultra RNA-seq library preparation kit, including two rounds of mRNA purification with NEB Poly(A) mRNA Magnetic Isolation Module, following the provided protocol. Multiplexed libraries were pooled and sequenced at 2 x 37bp paired end reads on a NextSeq 500 (see table below).

<u>mRNA-seq Tissue & Read Depth</u>	
<u>Tissue</u>	<u>Read Depth</u>
Prepupal Head	13,660,665
Pupal Head	15,000,698
Adult Head	16,938,828
Prepupal Wings	16,415,098
Pupal Forewings, Day 3	17,828,637
Pupal Hindwings, Day 3	12,467,577
Pupal Forewings, Day 6, Proximal	16,061,072
Pupal Forewings, Day 6, Medial	17,217,417

Pupal Forewings, Day 6, Distal	15,588,866
Pupal Hindwings, Day 6	18,131,864

Tissue description and read depth for mRNA-seq samples used in genome annotation.

mRNA-seq reads were aligned against the reference genome and splice junctions mapped using Tophat version 2.1.1 [65]. Aligned read locations from all libraries were merged into one transcriptome with “cuffmerge” in Cufflinks version 2.2.1 [66]. This merged transcriptome was then used for gene annotation with MAKER [44]. Three iterations of MAKER version 2.32 were used to annotate genes. We first ran MAKER with the reference transcriptome, a *Heliconius*-derived repeat library [109], and protein predictions from the *H. melpomene* genome [43]. We then used the resulting annotations to train the ab initio gene predictor SNAP, and ran MAKER again incorporating both SNAP and Augustus (using *H. melpomene* Augustus training model) gene predictors. We then retrained SNAP and ran MAKER as before to produce the final annotation set.

The resulting MAKER protein file was BLASTed against the *D. melanogaster* protein database downloaded from the Swissprot database to predict putative gene functions [110]. We ran the domain finder InterProScan, utilizing Pfam and SUPERFAMILY analyses with GOterms to annotate protein families and their putative domain functions. BLAST2GO basic was used to identify associated GO terms for each protein [45]. CEGMA [46] was ran with default settings, identifying 219 core genes. This was followed by manual TBLASTN analysis of the remaining 29 core genes, of which we identified an additional 16 as being present as either complete or partial fragments.

Chromatin immunoprecipitation

Head tissue (mixed sex) from 5 prepupae, 5 ommochrome-stage pupae, and 6 2 day old adults (all reared at approximately 30C, from the same laboratory stock used for genome assembly) were dissected in cold PBS buffer, then fixed on a nutator at room temperature for 7 minutes in a 1% fresh formaldehyde and cold PBS solution. 1M glycine was added to a final concentration of 125mM followed by incubation on a nutator at room temperature for an additional 5 minutes. Samples were washed twice with cold PBS for two minutes on a nutator, flash frozen in liquid nitrogen, and stored at -80C.

Preparation of nuclei and chromatin immunoprecipitation were performed using the SimpleChIP Enzymatic Chromatin IP Kit with modifications. All SimpleChIP buffers were prepared as directed and adding 20ul/mL of Roche Protease Inhibitor Cocktail tablets dissolved to 50x concentration in water. Head tissue for each developmental stage was pooled and gently disrupted in sucrose buffer [111] using a dounce homogenizer with pestle A. Cold buffer A (lysis buffer) from Cell Signaling Technologies was added to the disrupted tissue, then the sample was inverted 5 times, and incubated on ice for 30 seconds. Nuclei were washed once with cold buffer B, then resuspended in 500uL buffer B for Micrococcal nuclease (Mnase) digestion. 0.5uL of Mnase was added to the lysed nuclei, then the sample was inverted 5 times and incubated at 37C for 10 minutes. Chromatin digestion was halted with 0.5M EDTA, followed by chromatin extraction as described in the provided protocol using the optional dounce homogenization process. Digested chromatin for each stage-specific tissue sample was then aliquoted into separate Eppendorf tubes for immunoprecipitation.

For each immunoprecipitation reaction, 2.5ug of either H3K27ac (Abcam ab4729) or H3K4me3 (Abcam ab8580) antibody was added to ~3-4ug of digested chromatin. Samples were incubated in solution overnight at 4C on a nutator. Magnetic isolation and purification of antibody bound chromatin fragments was performed as described in the SimpleChIP protocol, with the exception that elution of chromatin from magnetic beads was performed for 2 hours at 65C with 10-15 seconds of vortexing every 15 minutes. Two experimental replicates were performed for each ChIP reaction, with the exception of ChIP targeting H3K27ac in adult head tissue. Due to limited tissue sample at this stage, there was sufficient chromatin for only one immunoprecipitation.

A chromatin input control sample was prepared for each tissue type at each developmental stage. Prior to aliquoting, raw digested chromatin from each sample was set aside for input library preparation. Input chromatin was reverse crosslinked and purified following the same procedure as used for immunoprecipitated chromatin.

Library preparation and sequencing

Illumina sequencing libraries were prepared for all ChIP and input control samples following the NEB Ultra DNA library preparation protocol using ~40-50ng of sample as input. Multiplexing was performed with NEBNext multiplex oligos for Illumina. Adaptor ligated samples were PCR amplified for 14 cycles, followed by gel extraction and purification of DNA representing 100-600bp of unligated fragment (1-4 histones). Libraries were pooled and sequenced on a NextSeq 500 at Cornell University's Genomics Facility.

ChIP-seq peak calling

Raw sequencing reads were aligned to version 1 of the *Hel* genome (described above) using Bowtie2. A custom script was used to remove non-uniquely mapping reads and to filter low quality read alignments. Peaks were called for each dataset and the appropriate input control file using the MACS2 [68] version 2.1.1 “callpeak” function with `-g` set to “3.6e8” and default parameter settings. ChIP-seq peaks for each histone mark and stage were considered confirmed if they were present in any replicate dataset from the same developmental stage, requiring a minimum 33% overlap in peaks when compared with the bedtools function “bedtools intersect -u -f .33 -F .33 -e” [69]. For adult H3K27ac peaks, we used the X peaks with the greatest fold change over input control, where X peaks is the average percent of confirmed H3K27ac peaks from the previous two stages.

Comparison of regulatory elements at different developmental stages

A combined set of unique ChIP-seq peaks from H3K4me3 and H3K27ac confirmed peak sets were called using bedtools for each developmental stage. Custom scripts were used to classify proximal (≤ 2 kb) and distal (> 2 kb) regulatory elements based on distance from the nearest annotated TSS. We chose a 2kb proximal cutoff following ENCODE and modENCODE precedents of 1kb from the nearest TSS (ENCODE Consortium 2012), then moderately increasing the proximal range to account for a known 3' bias in poly-A precipitation mRNA sequencing. Essentially, by increasing the proximal cutoff to 2kb we aimed to compensate for some low quality TSS annotations due to the incomplete assembly of some 5' UTRs. Novel, stage-specific regulatory elements were determined using the bedtools “intersect” command as described above. Additional custom scripts

were used to identify the unique set of genes closest to each annotated regulatory element and within the same scaffold. Unique proximal and distal regulatory element sets across all developmental stages were determined bedtools 'merge' to identify unique loci, requiring a minimal overlap of 1bp to merge two or more elements. GO enrichment was performed on stage-specific regulatory elements for each developmental stage as follows: Proximal and distal stage-specific regulatory elements were combined for each developmental stage, and the above scripts were used to identify the unique set of genes closest to each regulatory element. Stage-specific gene sets were then BLASTed against drosophila, and enriched GO terms were identified using the PANTHER database [71] available at geneontology.org.

Conservation of regulatory element loci

We identified evolutionarily conserved proximal and distal regulatory elements for each stage using a reciprocal best-hit BLAST algorithm. Sequences for each annotated regulatory element were extracted from the *Hel* genome assembly. We used the command "blastn -db database -query peak_sequences.fa -outfmt "6 qseqid sseqid slen sstart send evalue pident sseq" -evalue E-20 -max_target_seqs 1 -num_threads 16 -out blast_output_file" to identify conserved regulatory loci in each of the five selected lepidopteran genomes. To identify the optimal similarity threshold for this process, results from a single dataset using multiple e-value thresholds were compared, which produced little difference in observed counts of conserved regulatory sequences. An e-value minimum of "E-20" was adopted as a conservative threshold for all additional analyses. Sequences were then reciprocally BLASTed to the original regulatory element sequence

set to ensure only uniquely BLASTing loci were used in downstream analysis. A custom python script was used to identify conserved regulatory sequences for various phylogenetic groups. Adjusted conservation scores were determined following the process similar to that used by Taher et al. (2011) to identify non-aligning, functionally conserved elements: Conservation scores for a given species X were increased by the value of 7.1% multiplied by the difference between the number of conserved loci in X and the number of conserved loci in the next most recently diverged species. We performed additional manual curation of several loci for detailed presentation.

A null model was produced using custom python scripts to compare observed conservation counts with expected counts if sequences had been sampled randomly from the *Hel* genome assembly. Because conservation scores in BLAST results are partially dependent upon sequence length, we used the kernel density estimation function in the scikit-learn python library, with a Gaussian kernel and a bandwidth of “10”, to estimate the length distribution of our annotated regulatory sequences. A set of 10,000 sequence lengths were drawn from this distribution to generate our random sequence seq. A random sequence for each sequence length was then drawn from the *Hel* genome assembly, excluding 1kb on each scaffold end. Assuming approximately uniform coverage of the *Hel* scaffolds, we assigned the probability of drawing any sequence from a given scaffold as equal to the percent of the genome contained within that scaffold. All subsequent analyses of randomly generated sequences were performed as described above for annotated regulatory elements. Comparison of annotated gene CDSs was

performed as described for annotated regulatory elements using the complete CDS set from the gene annotation process.

Supplemental Data

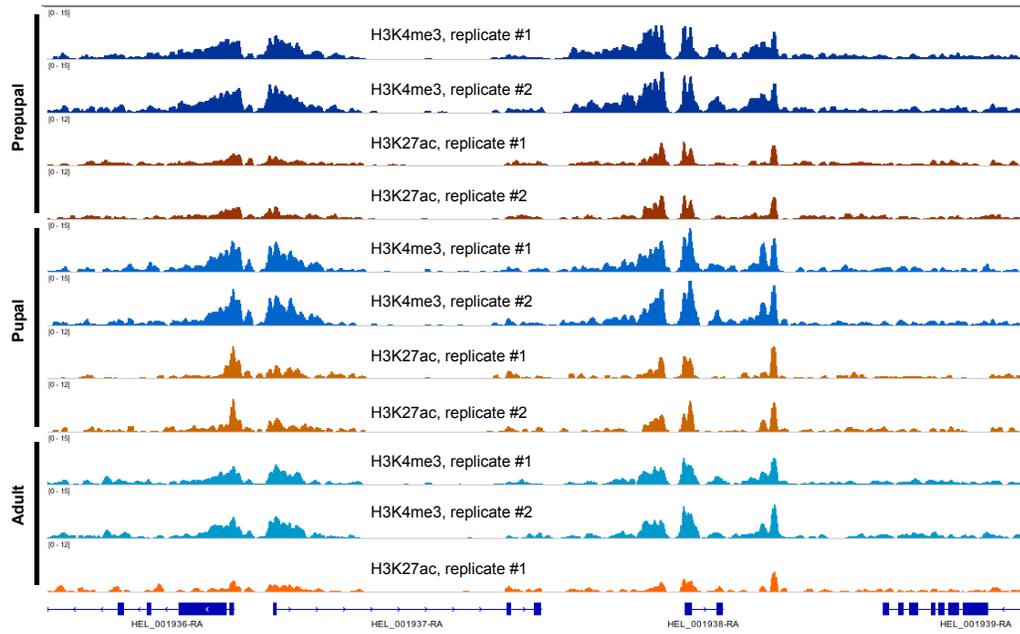


Figure S2.1, related to Figure 2.1. Representative browser tracks of ChIP-seq replicates. Input normalized fold-enrichment tracks for all ChIP-seq replicates show a high degree of similarity between replicates.

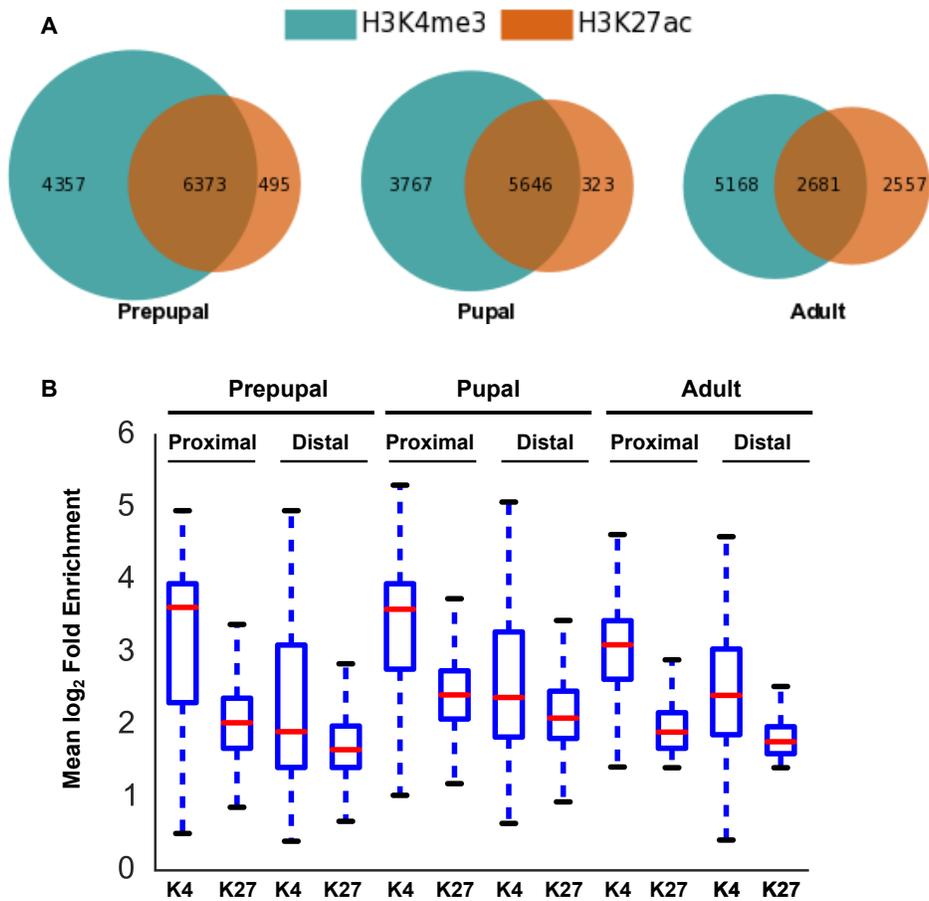


Figure S2.2, related to Figure 2.2. Overlap and enrichment profiles of H3K4me3 and H3K27ac peaks. A) H3K27ac enrichment displays a high degree of overlap with H3K4me3 enriched loci. B) Mean \log_2 normalized fold enrichment scores for proximal and distal regulatory elements at all stages. Proximal and distal elements at each stage show distinct enrichment profiles of H3K4me3 and H3K27ac, with a low H3K4me3:H3K27ac ratio in distal elements.

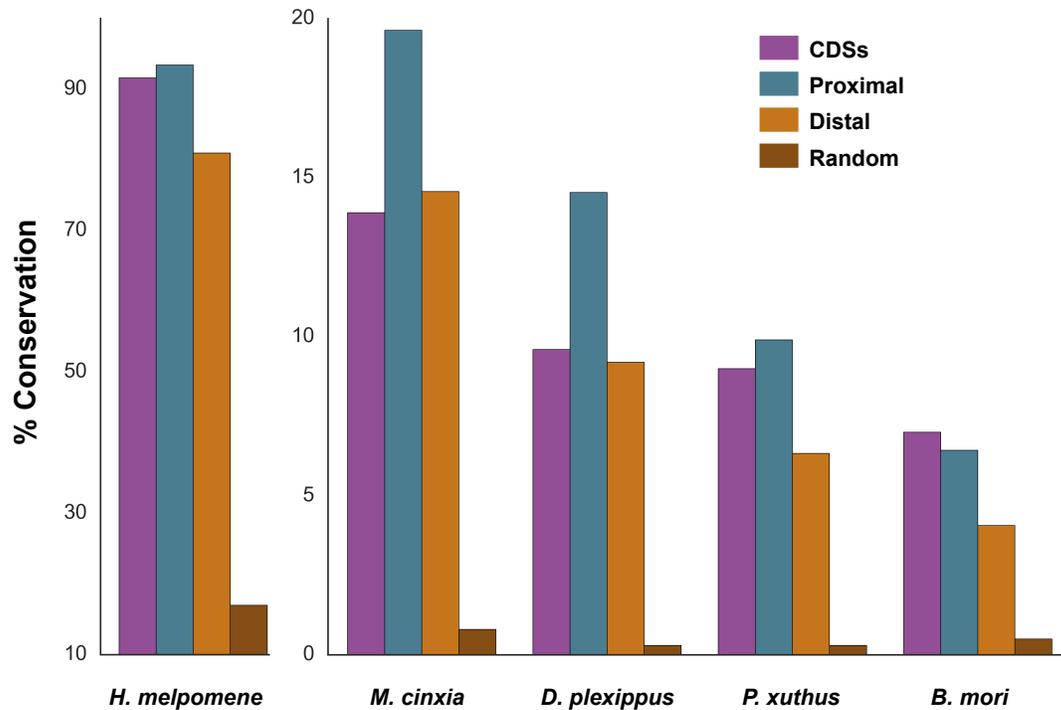


Figure S2.3, related to Figure 2.3. Conservation of genomic sequences shows similar degree of selection between regulatory elements and coding sequences. Conservation of proximal (blue) and distal (orange) regulatory elements is approximately the same or greater than conservation of gene CDSs (purple), and much greater than observed conservation of random sequences chosen independent of genomic features (brown). In most cases, proximal loci are more conserved than CDS sequences. Mean scores are shown for proximal and distal bars.

GO Enrichment of Stage-Specific Regulatory Loci

<u>Stage</u>	<u>GO #</u>	<u>Biological Processes</u>	<u>P-value</u>
Prepupal	1902578	single-organism localization	3.62E-03
	0007154	cell communication	1.23E-02
	0044765	single-organism transport	4.19E-02
	0051179	localization	3.24E-02
	0044763	single-organism cellular process	5.14E-08
	0044699	single-organism process	5.36E-10
	0008152	metabolic process	3.12E-03
	0071704	organic substance metabolic process	2.46E-02
	0009987	cellular process	4.99E-07
Pupal	0044699	single-organism process	1.14E-06
Adult	0006629	lipid metabolic process	4.20E-03
	0019752	carboxylic acid metabolic process	4.81E-02
	0055085	transmembrane transport	2.12E-02
	0044249	cellular biosynthetic process	1.11E-02
	1901576	organic substance biosynthetic process	2.88E-02
	0009058	biosynthetic process	2.07E-02
	0034641	cellular nitrogen compound metabolic process	2.27E-02
	0044237	cellular metabolic process	3.15E-05
	0044763	single-organism cellular process	1.49E-08
	0044707	single-multicellular organism process	1.33E-02
	0044767	single-organism developmental process	5.03E-03
	0007275	multicellular organism development	4.27E-02
	0032502	developmental process	7.97E-03
	0044699	single-organism process	2.31E-11
	0044238	primary metabolic process	9.69E-03
	0008152	metabolic process	3.35E-04
	0009987	cellular process	7.80E-10
0071704	organic substance metabolic process	6.32E-03	

Table S2.1, related to Figures 2.3-2.4. GO enrichment categories of nearest genes to combined proximal and distal stage-specific regulatory loci by developmental stage.

BLAST e-value Comparison (Adult Proximal Loci)

<u>BLAST e-value</u>	<u><i>H. melpomene</i></u>		<u><i>M. cinxia</i></u>		<u><i>D. plexippus</i></u>		<u><i>P. xuthus</i></u>	
	<u>Count</u>	<u>%</u>	<u>Count</u>	<u>%</u>	<u>Count</u>	<u>%</u>	<u>Count</u>	<u>%</u>
1.00E-05	5707	94.7	1467	24.4	1091	18.1	808	13.4
1.00E-10	5678	94.3	1406	23.4	1061	17.6	768	12.8
1.00E-20	5621	93.4	1286	21.4	947	15.7	664	11.0

Table S2.2, related to Figure 2.3. Representative comparison of BLAST e-value thresholds

Pairwise conservation of regulatory elements in six lepidopteran genomes*

<u>Stage</u>	<u><i>H. erato</i></u>	<u><i>H. melpomene</i></u>	<u><i>M. cinxia</i></u>	<u><i>D. plexippus</i></u>	<u><i>P. xuthus</i></u>	<u><i>B. mori</i></u>
Proximal						
Prepupal	6019 (100%)	5621 (93.4%)	1286 (21.4%)	947 (15.7%)	664 (11.0%)	453 (7.5%)
Pupal	5805 (100%)	5409 (93.1%)	1097 (18.9%)	805 (13.9%)	554 (9.5%)	351 (6.0%)
Adult	5399 (100%)	5066 (93.8%)	1004 (18.6%)	754 (14.0%)	498 (9.2%)	315 (5.8%)
Distal						
Prepupal	5198(100%)	4305 (82.8%)	856 (16.5%)	530 (10.2%)	376 (7.2%)	255 (4.9%)
Pupal	3929 (100%)	3307 (84.2%)	621 (15.8%)	392 (10.0%)	269 (6.8%)	168 (4.3%)
Adult	5004 (100%)	3796 (75.9%)	572 (11.4%)	368 (7.4%)	249 (5.0%)	150 (3.0%)
Null						
Null	9950 (100%)	1645 (17%)	78 (0.8%)	35 (0.3%)	27 (0.3%)	47 (0.5%)

Conservation of regulatory elements by taxonomic group*

<u>Stage</u>	<u>Heliconius</u>	<u>Nymphalidae</u>	<u>Papilionoidea</u>	<u>Lepidoptera</u>
Proximal				
Prepupal	5621 (93.4%)	679 (11.3%)	413 (6.9%)	234 (3.9%)
Pupal	5409 (93.1%)	570 (9.8%)	332 (5.7%)	194 (3.3%)
Adult	5066 (93.8%)	541 (10.0%)	314 (5.8%)	184 (3.4%)
Distal				
Prepupal	4305 (82.8%)	418 (8.0%)	253 (4.9%)	152 (2.9%)
Pupal	3307 (84.2%)	305 (7.8%)	183 (4.7%)	107 (2.7%)
Adult	3796 (75.9%)	285 (5.7%)	163 (3.3%)	95 (1.9%)

Pairwise conservation of shared and stage-specific regulatory elements*

<u>Stage</u>	<u><i>H. erato</i></u>	<u><i>H. melpomene</i></u>	<u><i>M. cinxia</i></u>	<u><i>D. plexippus</i></u>	<u><i>P. xuthus</i></u>	<u><i>B. mori</i></u>
Shared						
Proximal						
Prepupal	7026 (100%)	4772 (96.8%)	1165 (23.6%)	872 (17.7%)	611 (12.4%)	414 (8.4%)
Pupal	8782 (100%)	5245 (94.0%)	1076 (19.3%)	795 (14.2%)	546 (10.4%)	348 (6.6%)
Adult	7045 (100%)	4885 (95.1%)	990 (19.3%)	749 (14.6%)	496 (9.7%)	313 (6.1%)
Distal						
Prepupal	3539 (100%)	2381 (90.6%)	575 (21.9%)	399 (15.2%)	287 (10.9%)	180 (6.8%)
Pupal	4270 (100%)	2654 (86.6%)	540 (17.6%)	358 (11.7%)	256 (8.3%)	160 (5.2%)
Adult	3119 (100%)	2229 (87.2%)	460 (18.0%)	315 (12.3%)	224 (8.8%)	134 (5.2%)
Stage-specific						
Proximal						
Prepupal	1095 (100%)	858 (78.4%)	121 (11.1%)	75 (6.8%)	53 (4.8%)	39 (3.6%)
Pupal	225 (100%)	167 (74.2%)	21 (9.3%)	10 (4.4%)	8 (3.6%)	3 (1.3%)
Adult	264 (100%)	182 (68.9%)	14 (5.3%)	5 (1.9%)	2 (0.8%)	2 (0.8%)
Distal						
Prepupal	2570 (100%)	1941 (75.5%)	281 (10.9%)	131 (5.1%)	89 (3.5%)	75 (2.9%)
Pupal	863 (100%)	655 (75.9%)	81 (9.4%)	34 (3.9%)	13 (1.5%)	8 (0.9%)
Adult	2448 (100%)	1573 (64.3%)	112 (9.6%)	53 (2.2%)	25 (1.0%)	16 (0.7%)

*percent conservation rounded to nearest 0.1%

Table S2.3, related to Figures 2.3-2.4. Complete analysis of regulatory sequence conservation in six lepidopteran genomes.

Adjusted Conservation Counts												
<u>Stage</u>	<u>H. erato</u>		<u>H. melpomene</u>		<u>M. cinxia</u>		<u>D. plexippus</u>		<u>P. xuthus</u>		<u>B. mori</u>	
	Old	New	Old	New	Old	New	Old	New	Old	New	Old	New
Proximal												
Prepupal	6019	NA	5621	5649	1286	1593	947	971	664	684	453	467
Distal												
Prepupal	5198	NA	4305	4368	856	1100	530	553	376	386	255	263

Table S2.4, related to Figures 2.3-2.4. Representative adjusted conservation counts.

Allpaths-LG Assembly			HaploMerger Assembly		
Scaffolds	withGaps	withoutGaps	Scaffolds	withGaps	withoutGaps
#Seqs	12,985		#Seqs	183	
Min	875	875	Min	7,364	3,506
1st Qu.	1,426	1,424	1st Qu.	232,920	170,270
Median	2,491	2,359	Median	1,073,441	856,873
Mean	52,022	41,118	Mean	2,107,322	1,698,945
3rd Qu.	15,390	9,276	3rd Qu.	3,170,965	2,612,802
Max	4,501,536	4,265,285	Max	16,843,803	13,447,983
Total	675,515,114	533,921,454	Total	385,640,091	310,907,109
n50	362,517	298,548	n50	4,347,866	3,429,711
n90	62,710	49,822	n90	1,200,952	965,890
n95	21,154	12,791	n95	760,836	624,051
Contigs	withNs	withoutNs	Contigs	withNs	withoutNs
#Seqs	75,794		#Seqs	35,440	
Min	77	77	Min	3	3
1st Qu.	1,770	1,770	1st Qu.	2,253	2,251
Median	3,671	3,671	Median	5,267	5,258
Mean	7,044	7,044	Mean	8,772	8,762
3rd Qu.	8,623	8,622	3rd Qu.	11,188	11,168
Max	251,500	251,480	Max	251,500	251,463
Total	533,921,454	533,901,557	Total	310,907,109	310,549,193
n50	13,209	13,209	n50	15,349	15,332
n90	2,760	2,760	n90	3,994	3,988
n95	1,825	1,825	n95	2,409	2,407
Gaps			Gaps		
#Seqs	62,809		#Seqs	35,257	
Min	25		Min	25	
1st Qu.	690		1st Qu.	655	
Median	1,426		Median	1,331	
Mean	2,254		Mean	2,119	
3rd Qu.	3,040		3rd Qu.	2,715	
Max	16,520		Max	16,520	
Total	141,593,660		Total	74,732,982	
n50	4,055		n50	3,830	
n90	1,106		n90	1,040	
n95	770		n95	725	

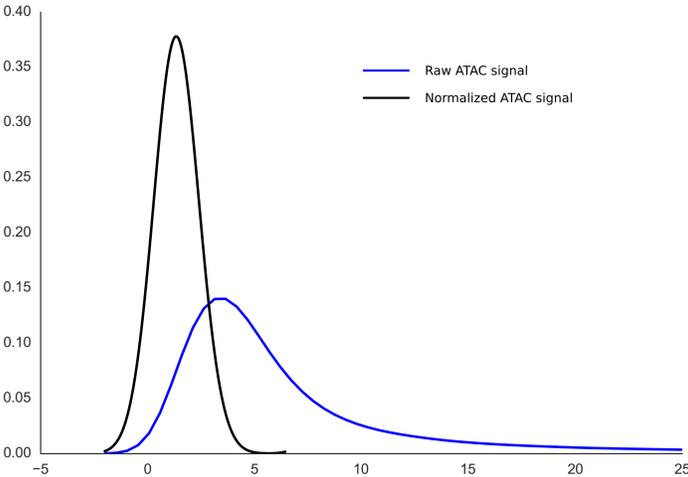
PBJelly Assembly			Final Assembly		
Scaffolds	withGaps	withoutGaps	Scaffolds	withGaps	withoutGaps
#Seqs	161		#Seqs	142	
Min	8,604	8,604	Min	8,886	8,861
1st,Qu.	267,417	264,274	1st Qu.	787,422	782,757
Median	1,513,428	1,498,657	Median	1,772,275	1,765,952
Mean	2,631,368	2,595,548	Mean	2,946,279	2,909,153
3rd,Qu.	3,816,056	3,794,770	3rd Qu.	3,958,606	3,901,366
Max	18,507,051	18,192,080	Max	18,493,827	18,178,881
Total	423,650,317	417,883,299	Total	418,371,739	413,099,771
n50	5,653,572	5,574,988	n50	5,483,780	5,466,767
n90	1,614,376	1,586,803	n90	1,432,179	1,426,368
n95	985,882	970,776	n95	951,373	936,409
Contigs	withNs	withoutNs	Contigs	withNs	withoutNs
#Seqs	5,679		#Seqs	5,326	
Min	3	3	Min	248	248
1st Qu.	20,727	20,690	1st Qu.	21,796	21,763
Median	46,498	46,474	Median	48,288	48,277

Mean	73,583	73,521	Mean	77,562	77,499
3rd Qu.	94,825	94,688	3rd Qu.	98,266	98,230
Max	1,206,344	1,206,239	Max	1,344,250	1,344,142
Total	417,883,299	417,526,057	Total	413,099,771	412,759,846
n50	123,404	123,214	n50	129,862	129,789
n90	35,917	35,875	n90	37,479	37,461
n95	23,693	23,689	n95	25,008	24,951
Gaps			Gaps		
#Seqs	5,518		#Seqs	5,184	
Min	25		Min	25	
1st Qu.	25		1st Qu.	25	
Median	25		Median	25	

Table S2.5, related to experimental procedures. Complete statistics for each stage of the *H. erato* assembly.

APPENDIX II:
SUPPLEMENT TO CHAPTER 3

SD3.2. Distribution of raw and normalized ATAC-seq signal



Supplemental Figures

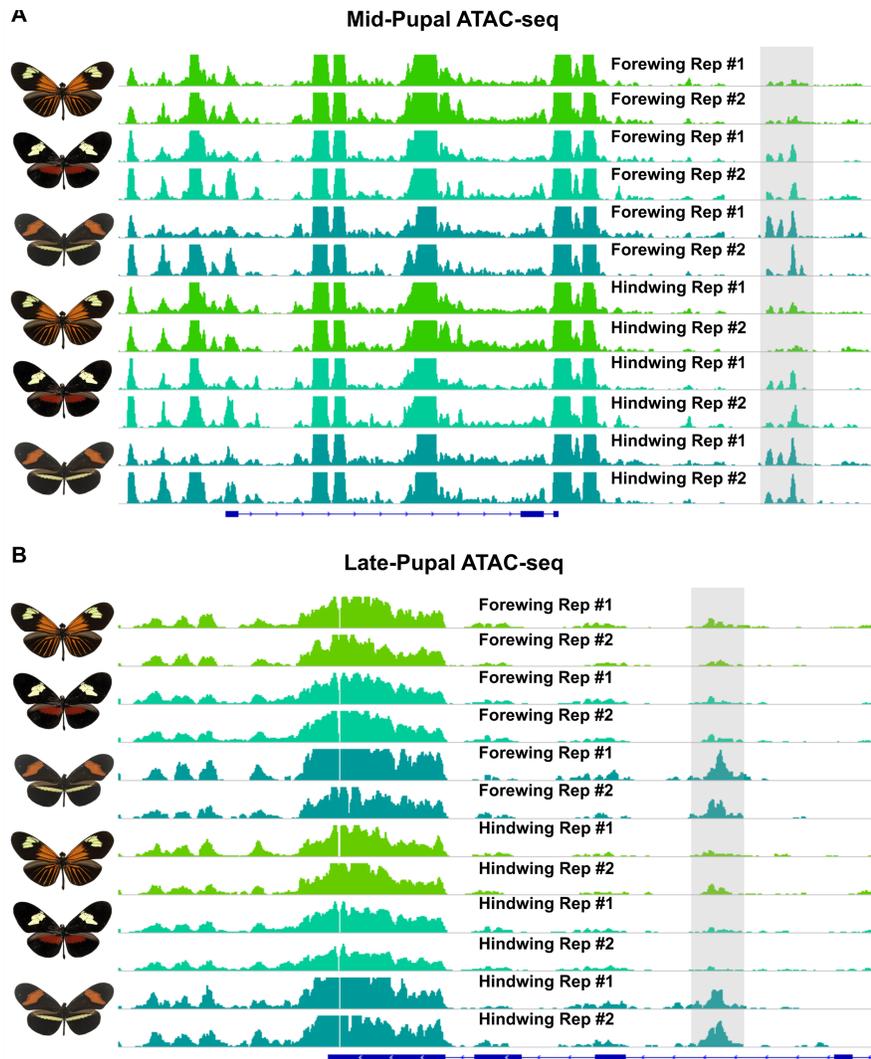


Figure S3.1. Replicate tracks for ATAC-seq data showing variable and non-variable loci. Example loci showing variable (shaded regions) and non-variable RPM normalized ATAC-seq signal for mid-pupal (**A**) and late-pupal (**B**) forewings and hindwings.

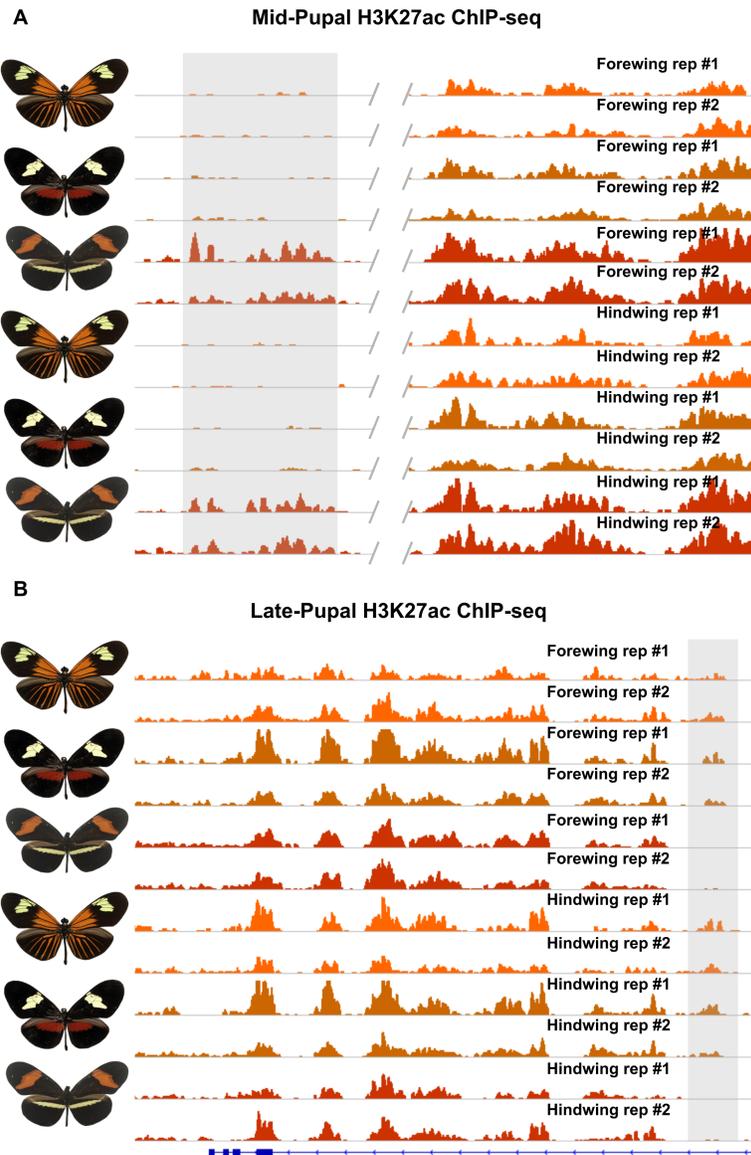


Figure S3.2. Replicate tracks for H3K27ac ChIP-seq data showing variable and non-variable loci. Example loci showing variable (shaded regions) and non-variable RPM normalized H3K27ac ChIP-seq signal for mid-pupal (**A**) and late-pupal (**B**) forewings and hindwings.

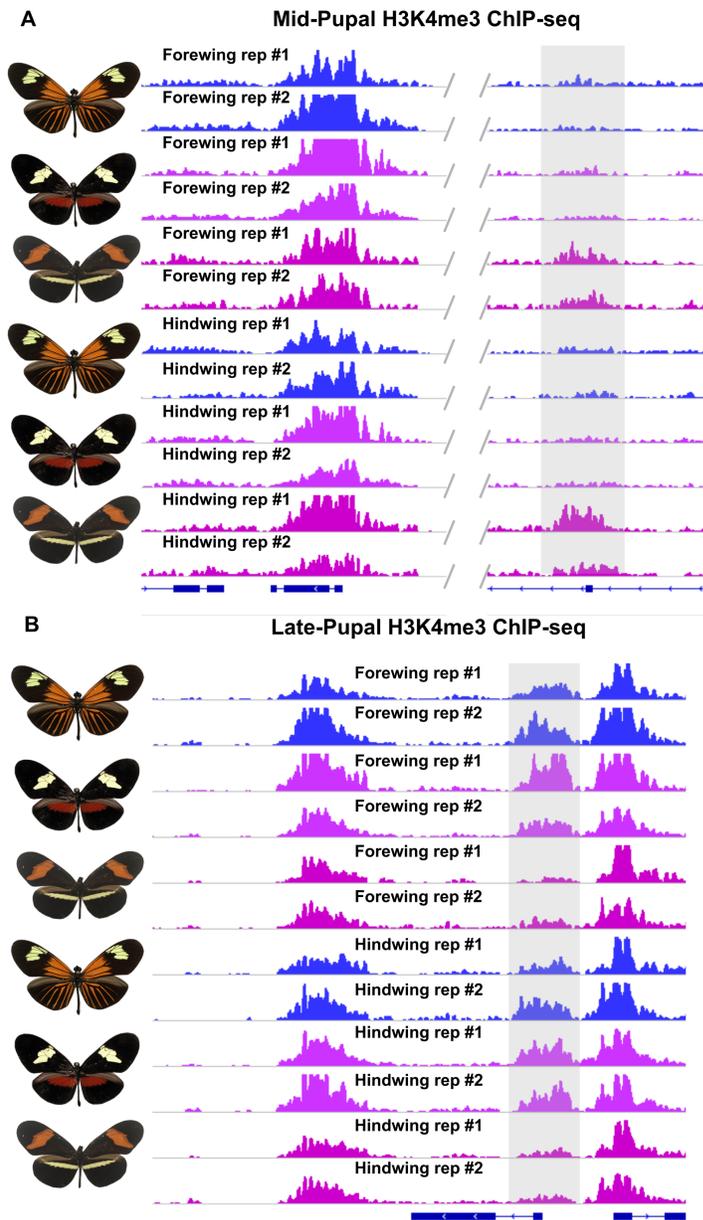


Figure S3.3. Replicate tracks for H3K4me3 ChIP-seq data showing variable and non-variable loci. Example loci showing variable (shaded regions) and non-variable RPM normalized H3K4me3 ChIP-seq signal for mid-pupal (**A**) and late-pupal (**B**) forewings and hindwings.

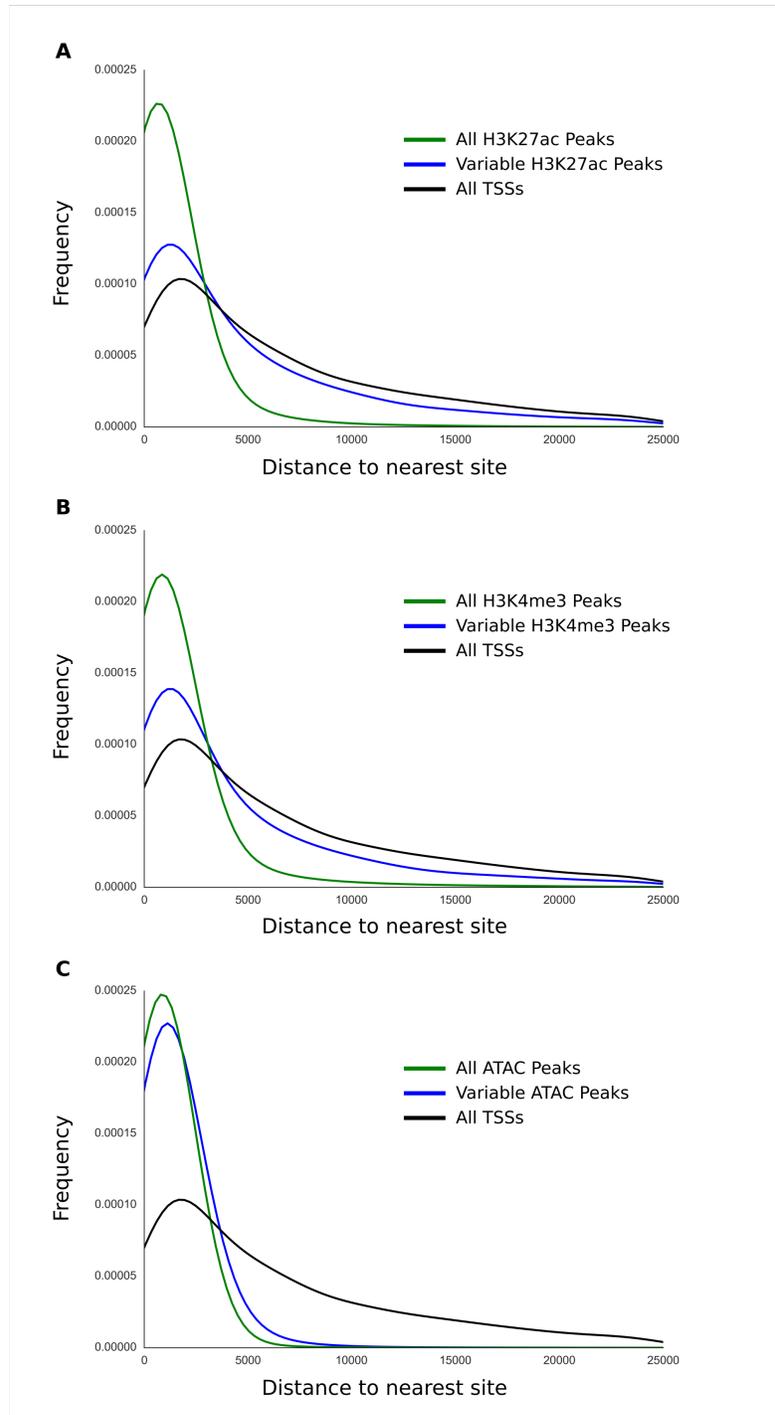


Figure S3.5. Distribution of spatial relationships for variable and non-variable regulatory loci indicates non-random sampling of active variable regulatory loci. Frequency distributions for the distance between variable (blue line) non-variable (green line) peak calls for H3K27ac (A), H3K4me3 (B), and ATAC-seq (C) peaks. In each panel, black line shows the same analysis for all annotated TSSs as a reference.

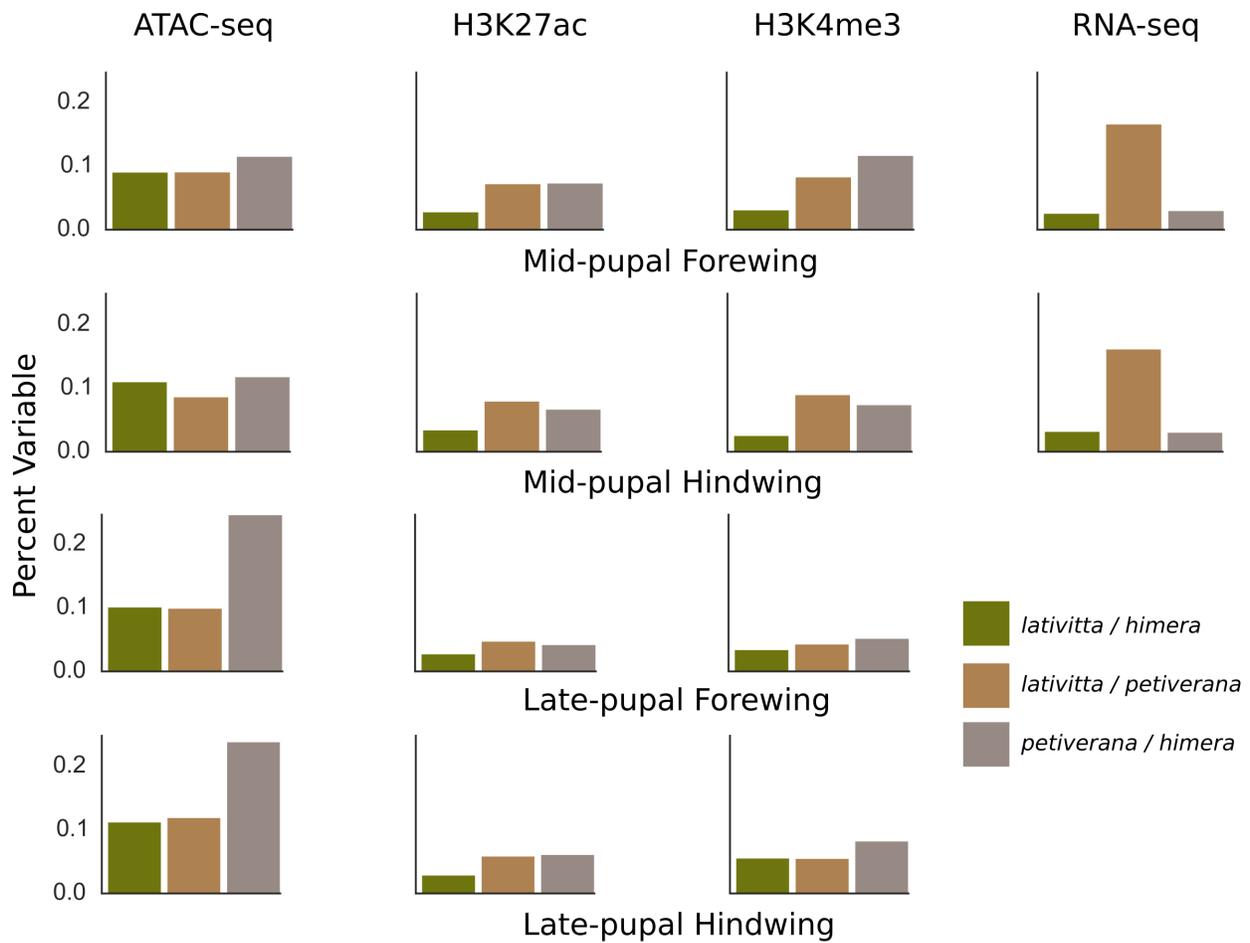


Figure S3.6. Frequency of each regulatory data type that is variable. Percent variability of each signal type by pairwise population comparison, tissue, and developmental stage.

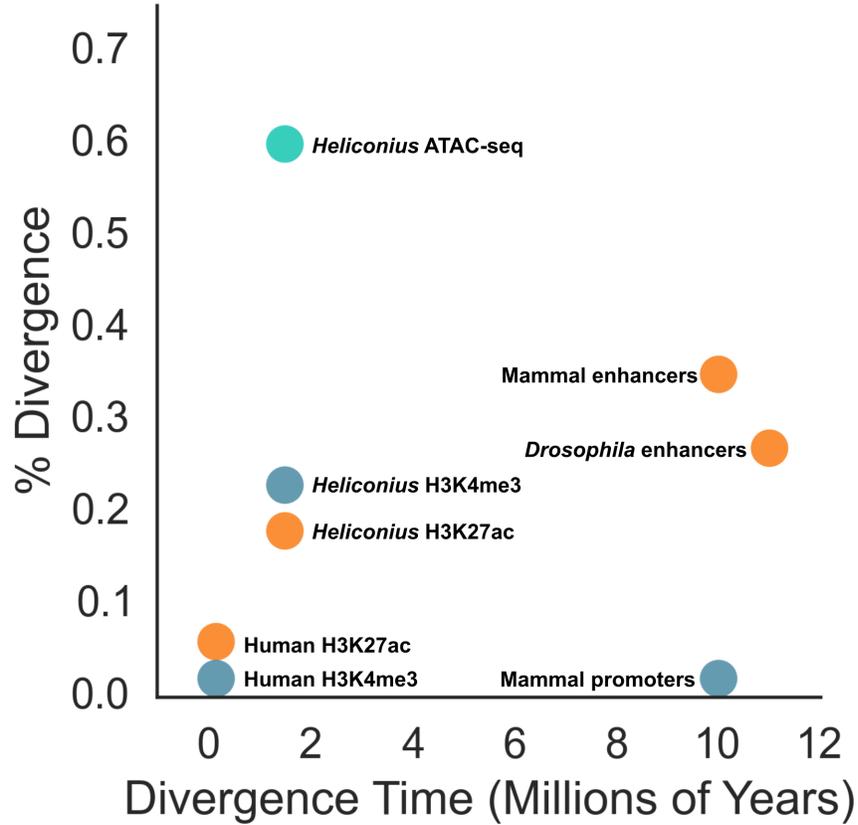


Figure S3.7. Regulatory variability in *Heliconius* populations, Human populations, inter-specific mammalian and drosophila comparisons. Population specific regulatory locus variability in *Heliconius* is significantly higher than observed in Human cell lines[28], and appears similar to inter-specific comparisons in mammalian[25] and *Drosophila* species[26]. ATAC-seq variability has no direct comparative equivalent, but is much greater than observed in prior population-level or species-level studies. Data points for mammalian and *Drosophila* studies represent the youngest species pair.

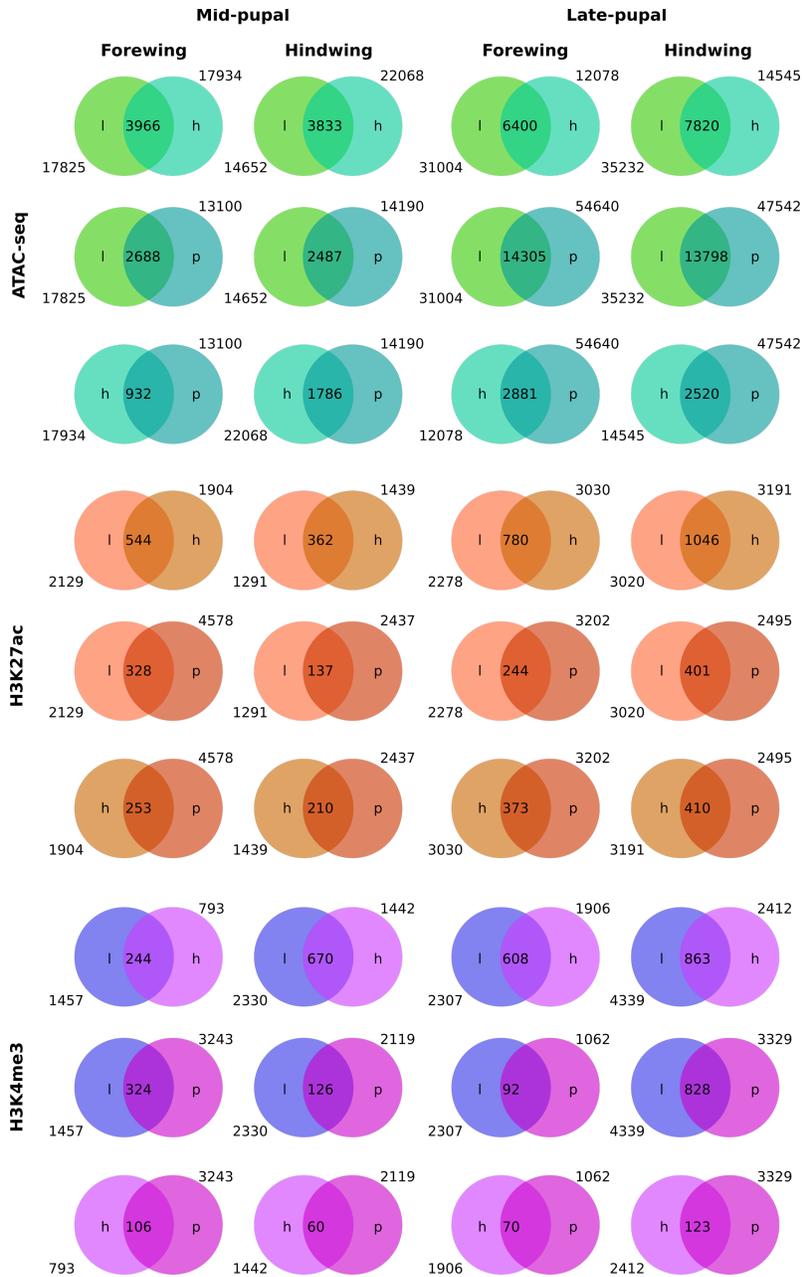


Figure S3.8. Shared variable regulatory loci for each regulatory data type. Shared variable regulatory loci for each data type, by developmental stage, tissue, and pairwise population comparison.

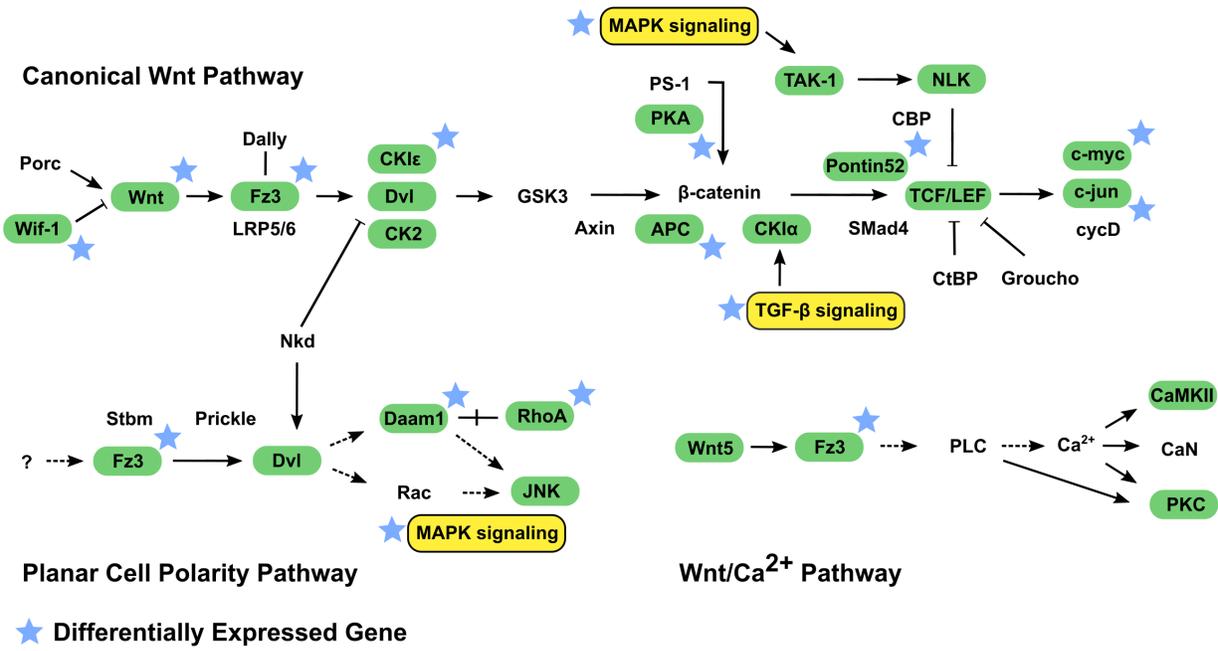


Figure S3.9. Canonical and non-canonical Wnt signaling pathways highlight divergence between populations of *H. erato*. Example of regulatory and gene expression divergence in signaling pathways. Canonical and non-canonical Wnt signaling pathways shown. Green indicates genes that are differentially expressed genes and/or have a variable histone mark within 2kb of the TSS, blue stars specifically indicate differentially expressed genes. Yellow indicates an interacting pathway for which additional genes were variable in either nearby regulatory activity or gene expression.

REFERENCES

1. Baker, M., *De novo genome assembly: what every biologist should know*. Nat Meth, 2012. **9**(4): p. 333-337.
2. Proulx, S.R., D.E.L. Promislow, and P.C. Phillips, *Network thinking in ecology and evolution*. Trends in Ecology & Evolution, 2005. **20**(6): p. 345-353.
3. Davidson, E.H. and D.H. Erwin, *Gene Regulatory Networks and the Evolution of Animal Body Plans*. Science, 2006. **311**(5762): p. 796.
4. Feder, M.E. and T. Mitchell-Olds, *Evolutionary and ecological functional genomics*. Nat Rev Genet, 2003. **4**(8): p. 649-655.
5. Schmidt, D., et al., *Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding*. Science, 2010. **328**(5981): p. 1036-1040.
6. Consortium, T.E.P., *An integrated encyclopedia of DNA elements in the human genome*. Nature, 2012. **489**(7414): p. 57-74.
7. Negre, N., et al., *A cis-regulatory map of the Drosophila genome*. Nature, 2011. **471**(7339): p. 527-531.
8. Alam, J. and J.L. Cook, *Reporter genes: Application to the study of mammalian gene transcription*. Analytical Biochemistry, 1990. **188**(2): p. 245-254.
9. Galas, D.J. and A. Schmitz, *DNAase footprinting a simple method for the detection of protein-DNA binding specificity*. Nucleic acids research, 1978. **5**(9): p. 3157-3170.
10. Song, L. and G.E. Crawford, *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells*. Cold Spring Harbor Protocols, 2010. **2010**(2): p. pdb. prot5384.
11. Simon, J.M., et al., *Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA*. Nature protocols, 2012. **7**(2): p. 256-267.
12. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position*. Nat Meth, 2013. **10**(12): p. 1213-1218.
13. Song, L., et al., *Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity*. Genome research, 2011. **21**(10): p. 1757-1767.

14. Valouev, A., et al., *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data*. *Nature methods*, 2008. **5**(9): p. 829-834.
15. Kharchenko, P.V., et al., *Comprehensive analysis of the chromatin landscape in Drosophila*. *Nature*, 2011. **471**(7339): p. 480-485.
16. Schmidt, D., et al., *ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions*. *Methods*, 2009. **48**(3): p. 240-248.
17. Zaret, K.S. and J.S. Carroll, *Pioneer transcription factors: establishing competence for gene expression*. *Genes & development*, 2011. **25**(21): p. 2227-2241.
18. Sudarsanam, P. and F. Winston, *The Swi/Snf family: nucleosome-remodeling complexes and transcriptional control*. *TRENDS in Genetics*, 2000. **16**(8): p. 345-351.
19. Peterson, C.L. and M.-A. Laniel, *Histones and histone modifications*. *Current Biology*, 2004. **14**(14): p. R546-R551.
20. Wittkopp, P.J., K. Vaccaro, and S.B. Carroll, *Evolution of yellow gene regulation and pigmentation in Drosophila*. *Current Biology*, 2002. **12**(18): p. 1547-1556.
21. Wittkopp, P.J., S.B. Carroll, and A. Kopp, *Evolution in black and white: genetic control of pigment patterns in Drosophila*. *TRENDS in Genetics*, 2003. **19**(9): p. 495-504.
22. Shapiro, M.D., et al., *Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks*. *Nature*, 2004. **428**(6984): p. 717-723.
23. Kvon, Evgeny Z., et al., *Progressive Loss of Function in a Limb Enhancer during Snake Evolution*. *Cell*, 2016. **167**(3): p. 633-642.e11.
24. Reed, R.D., et al., *optix Drives the Repeated Convergent Evolution of Butterfly Wing Pattern Mimicry*. *Science*, 2011. **333**(6046): p. 1137.
25. Villar, D., et al., *Enhancer Evolution across 20 Mammalian Species*. *Cell*, 2015. **160**(3): p. 554-566.
26. Arnold, C.D., et al., *Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution*. *Nat Genet*, 2014. **46**(7): p. 685-692.
27. Degner, J.F., et al., *DNaseI sensitivity QTLs are a major determinant of human expression variation*. *Nature*, 2012. **482**(7385): p. 390-394.

28. Kasowski, M., et al., *Extensive Variation in Chromatin States Across Humans*. Science, 2013. **342**(6159): p. 750-752.
29. Monteiro, A. and O. Podlaha, *Wings, Horns, and Butterfly Eyespots: How Do Complex Traits Evolve?* PLoS Biol, 2009. **7**(2): p. e1000037.
30. Wittkopp, P.J. and G. Kalay, *Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence*. Nat Rev Genet, 2012. **13**(1): p. 59-69.
31. Rubinstein, M. and F.S.J. de Souza, *Evolution of transcriptional enhancers and animal diversity*. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 2013. **368**(1632).
32. Su, J., S.A. Teichmann, and T.A. Down, *Assessing Computational Methods of Cis-Regulatory Module Prediction*. PLoS Comput Biol, 2010. **6**(12): p. e1001020.
33. Zhen, Y. and P. Andolfatto, *Methods to Detect Selection on Noncoding DNA*. Methods in molecular biology (Clifton, N.J.), 2012. **856**: p. 141-159.
34. Seb e-Pedr os, A., et al., *The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity*. Cell.
35. Schwaiger, M., et al., *Evolutionary conservation of the eumetazoan gene regulatory landscape*. Genome Research, 2014.
36. Kharchenko, P.V., et al., *Comprehensive analysis of the chromatin landscape in Drosophila melanogaster*. Nature, 2011. **471**(7339): p. 480-485.
37. Menet, J.S., et al., *Dynamic PER repression mechanisms in the Drosophila circadian clock: from on-DNA to off-DNA*. Genes & Development, 2010. **24**(4): p. 358-367.
38. Simola, D.F., et al., *Epigenetic (re)programming of caste-specific behavior in the ant Camponotus floridanus*. Science, 2016. **351**(6268).
39. Slattery, M., et al., *Genome-Wide Tissue-Specific Occupancy of the Hox Protein Ultrabithorax and Hox Cofactor Homothorax in <italic>Drosophila</italic>*. PLoS ONE, 2011. **6**(4): p. e14686.
40. Tobler, A., et al., *First-generation linkage map of the warningly colored butterfly Heliconius erato*. Heredity, 2004. **94**(4): p. 408-417.
41. Huang, S., et al., *HaploMerger: Reconstructing allelic relationships for polymorphic diploid genome assemblies*. Genome Research, 2012. **22**(8): p. 1581-1588.

42. Papa, R., et al., *Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in Heliconius butterflies*. BMC Genomics, 2008. **9**(1): p. 1-15.
43. Davey, J.W., et al., *Major Improvements to the Heliconius melpomene Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution*. G3: Genes|Genomes|Genetics, 2016. **6**(3): p. 695-708.
44. Cantarel, B.L., et al., *MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes*. Genome Research, 2008. **18**(1): p. 188-196.
45. Conesa, A., et al., *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research*. Bioinformatics, 2005. **21**(18): p. 3674-3676.
46. Parra, G., K. Bradnam, and I. Korf, *CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes*. Bioinformatics, 2007. **23**(9): p. 1061-1067.
47. Calo, E. and J. Wysocka, *Modification of Enhancer Chromatin: What, How, and Why?* Molecular Cell, 2013. **49**(5): p. 825-837.
48. Core, L.J., et al., *Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers*. Nat Genet, 2014. **46**(12): p. 1311-1320.
49. Zhan, S., et al., *The monarch butterfly genome yields insights into long-distance migration*. Cell, 2011. **147**(5): p. 1171-1185.
50. Ahola, V., et al., *The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera*. Nature communications, 2014. **5**: p. 4737-4737.
51. Li, X., et al., *Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies*. Nature Communications, 2015. **6**: p. 8212.
52. Mita, K., et al., *The Genome Sequence of Silkworm, Bombyx mori*. DNA Research, 2004. **11**(1): p. 27-35.
53. Wahlberg, N., et al., *Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary*. Proceedings of the Royal Society of London B: Biological Sciences, 2009. **276**(1677): p. 4295-4302.
54. Wahlberg, N., C.W. Wheat, and C. PeÒa, *Timing and Patterns in the Taxonomic Diversification of Lepidoptera (Butterflies and Moths)*. PLoS ONE, 2013. **8**(11): p. e80875.

55. Dermitzakis, E.T. and A.G. Clark, *Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover*. *Molecular Biology and Evolution*, 2002. **19**(7): p. 1114-1121.
56. Taher, L., et al., *Genome-wide identification of conserved regulatory function in diverged sequences*. *Genome Research*, 2011. **21**(7): p. 1139-1149.
57. Keightley, P.D., et al., *Estimation of the Spontaneous Mutation Rate in *Heliconius melpomene**. *Molecular Biology and Evolution*, 2014.
58. Kumar, S. and S. Subramanian, *Mutation rates in mammalian genomes*. *Proceedings of the National Academy of Sciences*, 2002. **99**(2): p. 803-808.
59. Markert, M.J., et al., *Genomic Access to Monarch Migration Using TALEN and CRISPR/Cas9-Mediated Targeted Mutagenesis*. *G3: Genes|Genomes|Genetics*, 2016. **6**(4): p. 905-915.
60. Zhang, L. and R.D. Reed, *Genome editing in butterflies reveals that spalt promotes and Distal-less represses eyespot colour patterns*. *Nat Commun*, 2016. **7**.
61. Lowe, C.B., et al., *Three Periods of Regulatory Innovation During Vertebrate Evolution*. *Science*, 2011. **333**(6045): p. 1019-1024.
62. Gnerre, S., et al., *High-quality draft assemblies of mammalian genomes from massively parallel sequence data*. *Proceedings of the National Academy of Sciences*, 2011. **108**(4): p. 1513-1518.
63. English, A.C., et al., *Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology*. *PLoS ONE*, 2012. **7**(11): p. e47768.
64. Grabherr, M.G., et al., *Genome-wide synteny through highly sensitive sequence alignment: Satsuma*. *Bioinformatics*, 2010. **26**(9): p. 1145-1151.
65. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq*. *Bioinformatics*, 2009. **25**(9): p. 1105-1111.
66. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks*. *Nat. Protocols*, 2012. **7**(3): p. 562-578.
67. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. *Nat Meth*, 2012. **9**(4): p. 357-359.
68. Feng, J., et al., *Identifying ChIP-seq enrichment using MACS*. *Nature protocols*, 2012. **7**(9): p. 10.1038/nprot.2012.101.

69. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics, 2010. **26**(6): p. 841-842.
70. Neph, S., et al., *BEDOPS: high-performance genomic feature operations*. Bioinformatics, 2012. **28**(14): p. 1919-1920.
71. Mi, H., et al., *The PANTHER database of protein families, subfamilies, functions and pathways*. Nucleic Acids Research, 2005. **33**(suppl 1): p. D284-D288.
72. Wray, G.A., *The evolutionary significance of cis-regulatory mutations*. Nat Rev Genet, 2007. **8**(3): p. 206-216.
73. Albert, F.W. and L. Kruglyak, *The role of regulatory variation in complex traits and disease*. Nat Rev Genet, 2015. **16**(4): p. 197-212.
74. Kasowski, M., et al., *Variation in Transcription Factor Binding Among Humans*. Science, 2010. **328**(5975): p. 232-235.
75. McVicker, G., et al., *Identification of Genetic Variants That Affect Histone Modifications in Human Cells*. Science, 2013. **342**(6159): p. 747-749.
76. Li, X.Y., et al., *The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding*. Genome Biol, 2011. **12**(4): p. R34.
77. Flanagan, N.S., et al., *Historical demography of Müllerian mimicry in the neotropical Heliconius butterflies*. Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(26): p. 9704-9709.
78. Supple, M.A., et al., *Divergence with gene flow across a speciation continuum of Heliconius butterflies*. BMC Evolutionary Biology, 2015. **15**: p. 204.
79. Joron, M., et al., *Heliconius wing patterns: an evo-devo model for understanding phenotypic diversity*. Heredity, 2006. **97**(3): p. 157-167.
80. Hoyal Cuthill, J. and M. Charleston, *Phylogenetic Codivergence Supports Coevolution of Mimetic Heliconius Butterflies*. PLoS ONE, 2012. **7**(5): p. e36464.
81. Jiggins, C.D., et al., *What can hybrid zones tell us about speciation? The case of Heliconius erato and H. himera (Lepidoptera: Nymphalidae)*. Biological Journal of the Linnean Society, 1996. **59**(3): p. 221-242.
82. Brower, A.V., *Rapid morphological radiation and convergence among races of the butterfly Heliconius erato inferred from patterns of mitochondrial DNA evolution*. Proceedings of the National Academy of Sciences, 1994. **91**(14): p. 6491-6495.

83. Boyle, A.P., et al., *High-Resolution Mapping and Characterization of Open Chromatin across the Genome*. Cell, 2008. **132**(2): p. 311-322.
84. Nord, A.S., et al., *Rapid and Pervasive Changes in Genome-Wide Enhancer Usage During Mammalian Development*. Cell, 2013. **155**(7): p. 1521-1531.
85. Lewis, James J., et al., *ChIP-Seq-Annotated *Heliconius erato* Genome Highlights Patterns of cis-Regulatory Evolution in Lepidoptera*. Cell Reports, 2016. **16**(11): p. 2855-2863.
86. Karlić, R., et al., *Histone modification levels are predictive for gene expression*. Proceedings of the National Academy of Sciences, 2010. **107**(7): p. 2926-2931.
87. Hines, H.M., et al., *Transcriptome analysis reveals novel patterning and pigmentation genes underlying *Heliconius* butterfly wing pattern variation*. BMC Genomics, 2012. **13**: p. 288.
88. Orsi, G.A., et al., *High-resolution mapping defines the cooperative architecture of *Polycomb* response elements*. Genome Res, 2014. **24**(5): p. 809-20.
89. Martin, S.H., et al., *Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies*. Genome Research, 2013. **23**(11): p. 1817-1828.
90. Aoki, K.F. and M. Kanehisa, *Using the KEGG database resource*. Curr Protoc Bioinformatics, 2005. **Chapter 1**: p. Unit 1.12.
91. Kanehisa, M., Y. Sato, and K. Morishima, *BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences*. J Mol Biol, 2016. **428**(4): p. 726-31.
92. Valentine, M. and S. Collier, *Planar cell polarity and tissue design: Shaping the *Drosophila* wing membrane*. Fly, 2011. **5**(4): p. 316-321.
93. Ma, W., W.S. Noble, and T.L. Bailey, *Motif-based analysis of large nucleotide data sets using MEME-ChIP*. Nature protocols, 2014. **9**(6): p. 1428-1450.
94. Counterman, B.A., et al., *Genomic Hotspots for Adaptation: The Population Genetics of Müllerian Mimicry in *Heliconius erato**. PLoS Genet, 2010. **6**(2): p. e1000796.
95. Nadeau, N., et al., *Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato**. Genome Research, 2014.
96. Zhang, Y., et al., *Model-based Analysis of ChIP-Seq (MACS)*. Genome Biology, 2008. **9**(9): p. 1-9.

97. Boyle, A.P., et al., *F-Seq: a feature density estimator for high-throughput sequence tags*. *Bioinformatics*, 2008. **24**(21): p. 2537-2538.
98. Landt, S.G., et al., *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia*. *Genome Research*, 2012. **22**(9): p. 1813-1831.
99. Piechota, M., et al., *Seqinspector: position-based navigation through the ChIP-seq data landscape to identify gene expression regulators*. *BMC Bioinformatics*, 2016. **17**: p. 85.
100. Conover, W.J., *A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions*. *Journal of the American Statistical Association*, 1972. **67**(339): p. 591-596.
101. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions*. *Genome Biol*, 2013. **14**(4): p. R36.
102. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data*. *Bioinformatics*, 2015. **31**(2): p. 166-9.
103. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. *Genome Biology*, 2014. **15**(12): p. 550.
104. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. *Nature methods*, 2012. **9**(4): p. 357-359.
105. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data*. *Nat Genet*, 2011. **43**(5): p. 491-498.
106. Danecek, P., et al., *The variant call format and VCFtools*. *Bioinformatics*, 2011. **27**(15): p. 2156-2158.
107. Mathelier, A., et al., *JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles*. *Nucleic Acids Research*, 2013.
108. Smit, A., Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015; Available from: <http://www.repeatmasker.org/>.
109. Lavoie, C.A., et al., *Transposable element evolution in Heliconius suggests genome diversity within Lepidoptera*. *Mobile DNA*, 2013. **4**(1): p. 1-10.
110. Bairoch, A. and R. Apweiler, *The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL*. *Nucleic Acids Research*, 1996. **24**(1): p. 21-25.
111. John, S., et al., *Chromatin accessibility pre-determines glucocorticoid receptor binding patterns*. *Nat Genet*, 2011. **43**(3): p. 264-268.