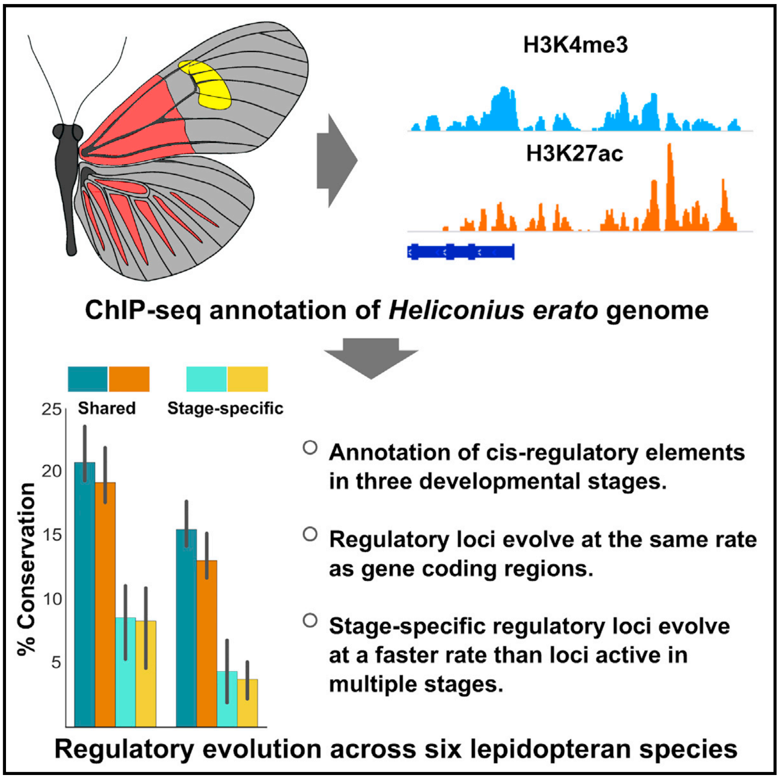
Cell Reports

Article

ChIP-Seq-Annotated *Heliconius erato* Genome Highlights Patterns of *cis*-Regulatory Evolution in Lepidoptera

Graphical Abstract



ChIP-seq annotation of *Heliconius erato* genome

○ Annotation of cis-regulatory elements in three developmental stages.

○ Regulatory loci evolve at the same rate as gene coding regions.

○ Stage-specific regulatory loci evolve at a faster rate than loci active in multiple stages.

Regulatory evolution across six lepidopteran species

Highlights

- Assembly and annotation of the *Heliconius erato* genome are provided
- ChIP-seq is used to annotate the regulatory loci for three stages of *H. erato* head development
- Lepidopteran regulatory loci evolve at approximately the same rate as coding regions
- Stage-specific regulatory loci evolve much faster than loci active in multiple stages

Authors

James J. Lewis, Karin R.L. van der Burg, Anyi Mazo-Vargas, Robert D. Reed

Correspondence

jjl336@cornell.edu

In Brief

Lewis et al. use ChIP-seq to annotate active gene regulatory elements over three periods of head development in the butterfly, *Heliconius erato*. Comparison of genomic sequences at regulatory loci across six lepidopteran genomes shows that regulatory elements evolve more rapidly if they are active at only a single stage.

Lewis et al., 2016, Cell Reports *16*, 2855–2863
September 13, 2016 © 2016 The Authors.
http://dx.doi.org/10.1016/j.celrep.2016.08.042

OPEN ACCESS
CellPress

ChIP-Seq-Annotated *Heliconius erato* Genome Highlights Patterns of *cis*-Regulatory Evolution in Lepidoptera

James J. Lewis,[1,2,*] Karin R.L. van der Burg,[1] Anyi Mazo-Vargas,[1] and Robert D. Reed[1]
[1]Department of Ecology and Evolutionary Biology, Cornell University, 215 Tower Road, Ithaca, NY 14853-7202, USA
[2]Lead Contact
*Correspondence: jjl336@cornell.edu

http://dx.doi.org/10.1016/j.celrep.2016.08.042

SUMMARY

Uncovering phylogenetic patterns of *cis*-regulatory evolution remains a fundamental goal for evolutionary and developmental biology. Here, we characterize the evolution of regulatory loci in butterflies and moths using chromatin immunoprecipitation sequencing (ChIP-seq) annotation of regulatory elements across three stages of 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. Our study provides a benchmark for genome-wide patterns of regulatory element evolution in insects, and it shows that developmental timing of activity strongly predicts patterns of regulatory sequence evolution.

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 (John et al., 2011), a general signature of most regulatory loci, identified over two million regulatory elements across 125 human cell lines (ENCODE Project Consortium, 2012). This discovery, coupled with the many case studies implicating *cis*-regulatory activity as a driving force of morphological evolution (Monteiro and Podlaha, 2009; Wittkopp and Kalay, 2012), clearly points to the importance of regulatory elements in shaping not only organisms but also 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 and binding motif predictions) have produced important results (Rubinstein and de Souza, 2013), but they also have limitations (Su et al., 2010; Zhen and Andolfatto, 2012). More recent efforts to incorporate functional regulatory element annotations have made use of chromatin immunoprecipitation sequencing (ChIP-seq), where antibodies targeting DNA-binding proteins of interest are used to isolate genomic sequences with regulatory activity (Schmidt et al., 2010; Villar et al., 2015). As of 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. (2010) used CEBPA ChIP assays to study conservation of transcription factor binding in livers of five vertebrate species. Human CEBPA-binding sites displayed between 15% and 2% conservation across 300 Ma 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 (Villar et al., 2015). Comparing all active regulatory loci, Villar et al. (2015) found only 1% of presumptive enhancers and 16% of presumptive promoters were conserved among all 20 species over 180 Ma 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 (Sebé-Pedrós et al., 2016; Schwaiger et al., 2014), though as of 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 govern 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 (Kharchenko et al., 2011; Nègre et al., 2011). Even the few exceptions (Menet et al., 2010; Simola et al., 2016; Slattery et al., 2011) have rarely focused on more than one developmental time point, and, to the best of our knowledge, no studies 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 frequently are reutilized between tissue-specific developmental programs. Of the 155,000 transcription factor-binding sites annotated by the model organism Encyclopedia of DNA Elements (modENCODE) consortium, assayed over a broad spectrum of developmental stages in whole *D. melanogaster*, only 35,000 binding sites were unique genomic loci (Nègre et al., 2011). 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 among 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 (ENCODE Project Consortium, 2012). 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 toward 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 it highlights a deep need for additional comparative study of in vivo regulatory activity across multiple developmental stages.

Here, we generate a portrait of genome-wide patterns of regulatory element evolution in an insect lineage, the Lepidoptera, and we 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 dataset that 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 Ma of evolution. We provide evidence of regulatory evolution at both transcription start site (TSS)-proximal and TSS-distal loci, and we 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 (~220 bp) and mate pair (3, 8, and 12 kb) libraries, made from a single, outbred female *H. erato lativitta* (*Hel*) pupa, were assembled to produce an initial assembly of 12,985 scaffolds, with scaffold and contig N50 values of 362 and 13.2 kb, respectively. The total assembly length, including scaffold gaps, was ~670 Mb. As previously reported, flow cytometry estimated a genome size of ~400 Mb for *H. erato petiverana* (Tobler et al., 2005), 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 (Huang et al., 2012), producing an assembly with a total length of ~385 Mb and considerably improving the scaffold and contig N50 values to 4.3 Mb and 15.3 kb, 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 5.5 Mb and 123 kb, 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 10 Ma (Tobler et al., 2005), and comparison of two assembled bacterial artificial chromosome (BAC) sequences for both species showed highly similar gene order (Papa et al., 2008). 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 (Davey et al., 2016). Comparisons with *Eueides isabella*, which split from the *Heliconius* genus ~18 Ma, showed 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* (Davey et al., 2016). Davey et al. (2016) also identified 21 as the ancestral chromosome number for *Heliconius* species, suggesting highly conserved chromosome content between the two species for which genomes

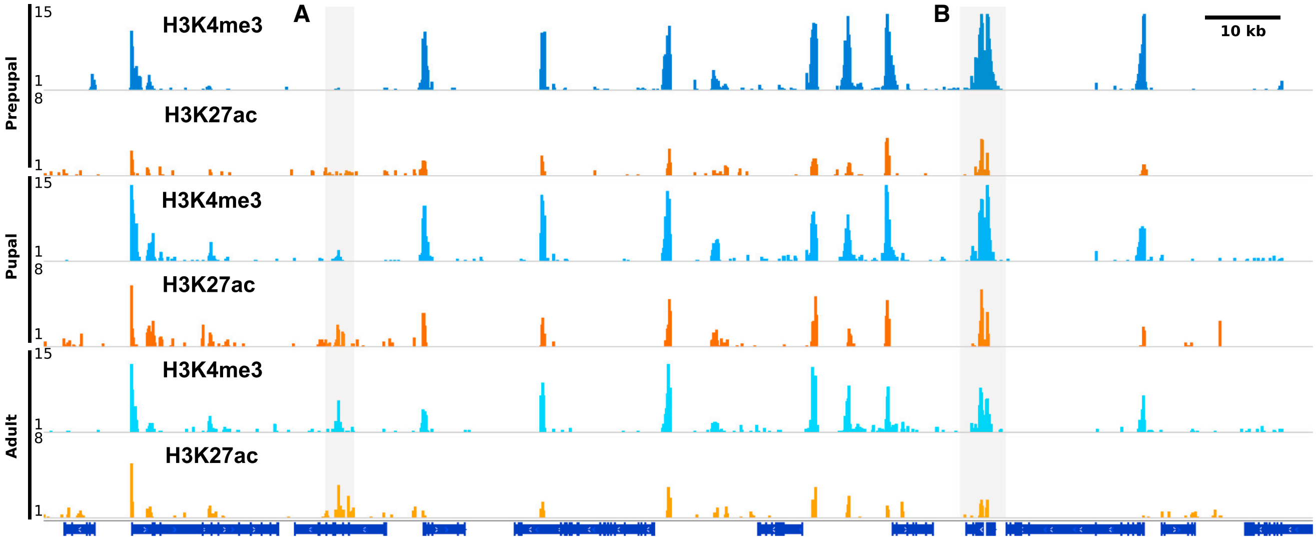


Figure 1. ChIP-Seq Signal Shows Change in Regulatory Activity during Tissue Maturation

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' are shown. (A and B) Representative examples of increasing (A) and decreasing (B) regulatory activity during head maturation are highlighted. See also Figure S1.

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 and deletions), 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 418 Mb, with a scaffold N50 of 5.48 Mb and a contig N50 of 129.8 kb. All further analyses were performed on this final genome assembly.

A total of 14,613 genes were predicted based on three iterations of MAKER (Cantarel et al., 2008), 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 gene ontology (GO) terms for 8,086 genes (Conesa et al., 2005). Analysis of genome completeness identified 95% of the 248 core CEGMA (Parra et al., 2007) genes. Our genome assemblies and annotated gene set are available for download and browsing at http://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 ChIP-seq (Figure 1; see also Figure 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 (Nègre et al., 2011; Calo and Wysocka, 2013; Core et al., 2014), and they 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 appeared to vary along with TSS proximity (Figure 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 2 kb of nearest TSS) or distal (>2 kb 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 http://butterflygenome.org.

Proximal versus 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 (~6–7 days post-pupation at 30°C, hereafter pupal), and 2-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 2A), for a total of 6,568

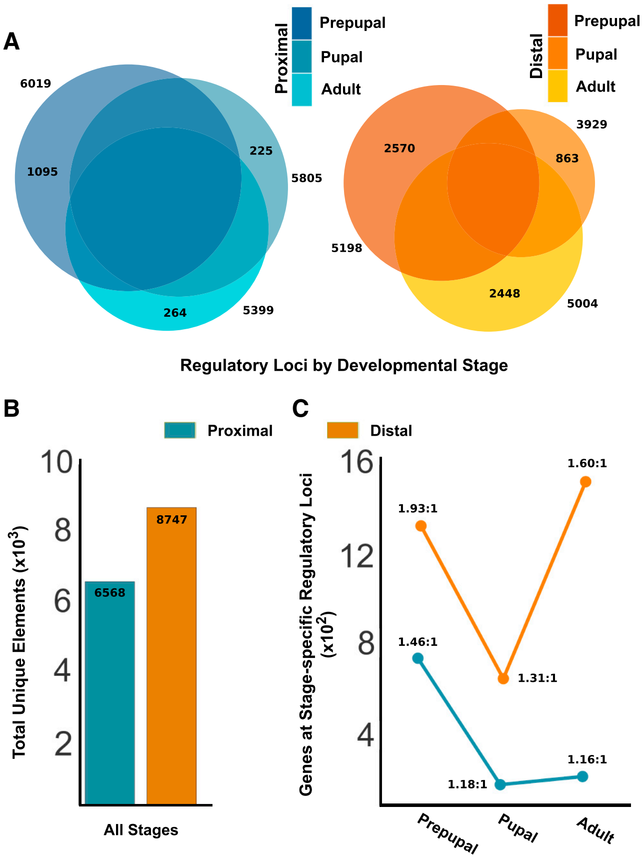


Figure 2. The *cis*-Regulatory Architecture of *Heliconius* Head Tissue Highlights Key Transitional Stages

(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 are shown.

(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 Figure S2.

and 8,747 unique proximal and distal loci across all stages (Figure 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 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 (Figure 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 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 were similar to those of *H. melpomene* and other lepidopteran species (Zhan et al., 2011; Ahola et al., 2014; Li et al., 2015; Davey et al., 2016), 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 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 ~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 ~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 prepatterning 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 S1). 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 *H. melpomene* (*Hm*), *Melitaea cinxia* (*Mc*), *Danaus plexippus* (*Dp*), *Papilio xuthus* (*Px*), and *Bombyx mori*

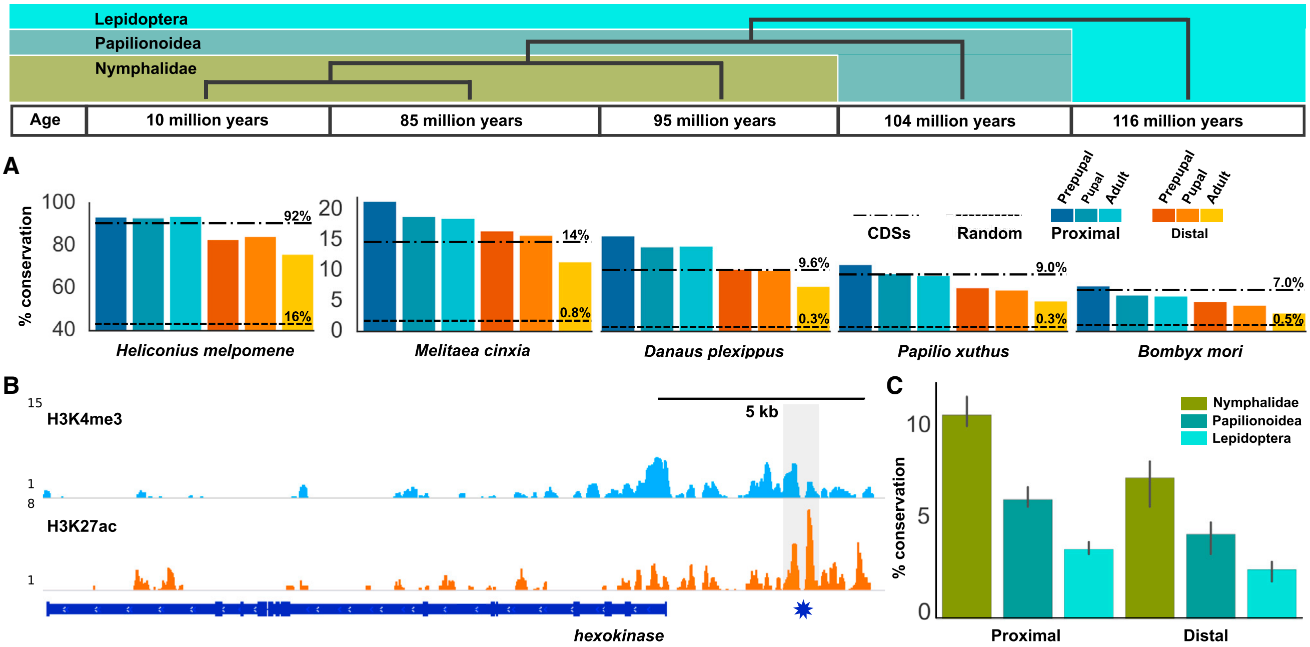


Figure 3. Evolutionary Trends in Annotated *cis*-Regulatory Elements

(A) Pairwise 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 pairwise comparison of randomly selected genomic sequences. Long dashed lines show conservation of all annotated gene CDSs. Phylogenetic scale and taxonomic groups are highlighted above.

(B) Input normalized signal for H3K4me3 (top) and H3K27ac (bottom) in adult head tissue. 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 ~116 Ma of evolution. Black bars indicate conservation scores across developmental stages.

See also Figure S3 and Tables S2–S4.

(*Bm*) (Mita et al., 2004; Zhan et al., 2011; Ahola et al., 2014; Li et al., 2015; Davey et al., 2016). 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 ranged from recent (10 Ma) for the two *Heliconius* species to the early Cretaceous (116 Ma) for divergence between *Heliconius* and *Bombyx* lineages (Wahlberg et al., 2009, 2013).

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 timescales. *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). 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 (Dermitzakis and Clark, 2002).

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 3A; see also Table S3). A noticeable conservation threshold at the transition from the genus *Heliconius* to more distantly related lepidopteran species was observed. Within the genus *Heliconius*, ~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 frequencies of regulatory loci in these species were 19%, 14%, 9%, and 6% for proximal loci and 14%, 9%, 6%, and 4% for distal loci, for *Mc*, *Dp*, *Px*, and *Bm*, respectively. Divergence times for these lineages have been estimated at ~78 Ma (*Mc*), 90 Ma (*Dp*), 104 Ma (*Px*), and 116 Ma (*Bm*) (Wahlberg et al., 2009, 2013).

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

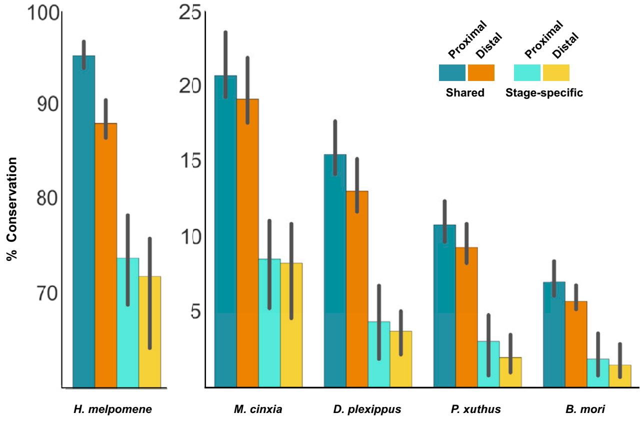


Figure 4. Stage-Specific and Shared Regulatory Elements Display Highly Dissimilar Evolutionary Patterns

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 Table S3.

(including both coding and noncoding 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 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 (Figure 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 (Figures 3B and 3C). 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 (Villar et al., 2015). Contrary to prior observations of highly reduced turnover in TSS-proximal regulatory elements (Schmidt et al., 2010; Villar et al., 2015), 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 (Taher et al., 2011). 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 (Dermitzakis and Clark, 2002). Nonetheless, adjusting our conservation counts according to the most significant results (7.1%) by Taher et al. (2011) produced negligible change in our observed evolutionary trends (Table S4). 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 4; see also Table 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 lower 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*, respectively. 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 we provide ChIP-based regulatory annotations for a butterfly, 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 will 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 we 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 timescales (Schmidt et al., 2010; Villar et al., 2015). This is perhaps unsurprising as per-generation mutation rates are similar in mammals and *Heliconius* butterflies (Keightley et al., 2015; Kumar and Subramanian, 2002). 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 historical, ecological, and evolutionary data on numerous butterfly species, combined with their amenability to functional genomic work (Markert et al., 2016; Zhang and Reed, 2016) and the availability of additional genome assemblies (Davey et al., 2016), 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 (Schmidt et al., 2010; Lowe et al., 2011; Villar et al., 2015). Thus, sorting our annotated regulatory elements by stage specificity has yielded insight 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 they 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 they 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 (Monteiro and Podlaha, 2009). 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 *H. erato lativitta* pupa. An initial assembly was produced using Allpaths-LG (Gnerre et al., 2011), resulting scaffolds were merged using HaploMerger (Huang et al., 2012), and additional scaffolding and gap filling were performed with long-read sequences using PBJelly (English et al., 2012). A Satsuma- (Grabherr et al., 2010) derived synteny map was used to produce a mostly ordered and oriented assembly of the 21 *H. erato* chromosomes (Table S5). Tophat (Trapnell et al., 2009) and Cufflinks (Trapnell et al., 2012) were used to assemble mRNA sequencing (mRNA-seq) data from head and wing tissues 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 (Cantarel et al., 2008; Bairoch and Apweiler, 1996; Lavoie, et al., 2013; Smit et al., 2013).

ChIP of prepupal, pupal, and adult head tissues 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 (Langmead and Salzberg, 2012) and enriched loci, peaks, were called using MACS2 (Feng et al., 2012) (Table S6). Final peak sets for each histone mark and tissue were called from overlapping replicate peak sets using bedtools (Quinlan and Hall, 2010). 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 (Neph et al., 2012). GO enrichment of neighboring genes to stage-specific regulatory elements was determined using the Protein Analysis Through Evolutionary Relationships (PANTHER) database (Mi et al., 2005).

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. (2011) to identify non-aligning, functionally conserved elements. See also the Supplemental Experimental Procedures.

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

ACCESSION NUMBERS

The accession number for the genome sequencing and ChIP-seq data reported in this paper is Sequence Read Archive (SRA): SRP074347.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2016.08.042.

AUTHOR CONTRIBUTIONS

J.J.L. and R.D.R. designed the study and wrote the paper. J.J.L. performed sample preparation, genome assembly, ChIP-seq, mRNA-seq, and data analysis. K.R.L.v.d.B. conducted genome annotation. A.M.-V. performed butterfly husbandry.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation grant DEB-1354318 to R.D.R. We thank Charles Danko, Amy McCune, and members of the R.D.R. lab for their helpful comments.

Received: May 19, 2016
Revised: July 14, 2016
Accepted: August 12, 2016
Published: September 13, 2016

REFERENCES

Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., Välimäki, N., Paulin, L., Kvist, J., Wahlberg, N., et al. (2014). The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat. Commun. *5*, 4737.

Bairoch, A., and Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res. *24*, 21–25.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. *18*, 188–196.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674–3676.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. *46*, 1311–1320.

Davey, J.W., Chouteau, M., Barker, S.L., Maroja, L., Baxter, S.W., Simpson, F., Joron, M., Mallet, J., Dasmahapatra, K.K., and Jiggins, C.D. (2016). Major improvements to the Heliconius melpomene genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. G3 (Bethesda) *6*, 695–708.

Dermitzakis, E.T., and Clark, A.G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. *19*, 1114–1121.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., and Gibbs, R.A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE *7*, e47768.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat. Protoc. *7*, 1728–1740.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA *108*, 1513–1518.

Grabherr, M.G., Russell, P., Meyer, M., Mauceli, E., Alföldi, J., Di Palma, F., and Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. Bioinformatics *26*, 1145–1151.

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., Fu, Y., Yuan, S., Chen, S., and Xu, A. (2012). HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res. *22*, 1581–1588.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.-H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. *43*, 264–268.

Keightley, P.D., Pinharanda, A., Ness, R.W., Simpson, F., Dasmahapatra, K.K., Mallet, J., Davey, J.W., and Jiggins, C.D. (2015). Estimation of the spontaneous mutation rate in Heliconius melpomene. Mol. Biol. Evol. *32*, 239–243.

Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T., et al. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature *471*, 480–485.

Kumar, S., and Subramanian, S. (2002). Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. USA *99*, 803–808.

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

Lavoie, C.A., Platt, R.N., 2nd, Novick, P.A., Counterman, B.A., and Ray, D.A. (2013). Transposable element evolution in Heliconius suggests genome diversity within Lepidoptera. Mob. DNA *4*, 21.

Li, X., Fan, D., Zhang, W., Liu, G., Zhang, L., Zhao, L., Fang, X., Chen, L., Dong, Y., Chen, Y., et al. (2015). Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. Nat. Commun. *6*, 8212.

Lowe, C.B., Kellis, M., Siepel, A., Raney, B.J., Clamp, M., Salama, S.R., Kingsley, D.M., Lindblad-Toh, K., and Haussler, D. (2011). Three periods of regulatory innovation during vertebrate evolution. Science *333*, 1019–1024.

Markert, M.J., Zhang, Y., Enuameh, M.S., Reppert, S.M., Wolfe, S.A., and Merlin, C. (2016). Genomic access to monarch migration using TALEN and CRISPR/Cas9-mediated targeted mutagenesis. G3 (Bethesda) *6*, 905–915.

Menet, J.S., Abruzzi, K.C., Desrochers, J., Rodriguez, J., and Rosbash, M. (2010). Dynamic PER repression mechanisms in the Drosophila circadian clock: from on-DNA to off-DNA. Genes Dev. *24*, 358–367.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., et al. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. *33*, D284–D288.

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., Namiki, N., Kitagawa, M., Yamashita, H., Yasukochi, Y., et al. (2004). The genome sequence of silkworm, Bombyx mori. DNA Res. *11*, 27–35.

Monteiro, A., and Podlaha, O. (2009). Wings, horns, and butterfly eyespots: how do complex traits evolve? PLoS Biol. *7*, e37.

Nègre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R., et al. (2011). A cis-regulatory map of the Drosophila genome. Nature *471*, 527–531.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920.

Papa, R., Morrison, C.M., Walters, J.R., Counterman, B.A., Chen, R., Halder, G., Ferguson, L., Chamberlain, N., Ffrench-Constant, R., Kapan, D.D., et al. (2008). Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in Heliconius butterflies. BMC Genomics *9*, 345.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rubinstein, M., and de Souza, F.S.J. (2013). Evolution of transcriptional enhancers and animal diversity. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20130017.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science *328*, 1036–1040.

Schwaiger, M., Schönauer, A., Rendeiro, A.F., Pribitzer, C., Schauer, A., Gilles, A.F., Schinko, J.B., Renfer, E., Fredman, D., and Technau, U. (2014). Evolutionary conservation of the eumetazoan gene regulatory landscape. Genome Res. *24*, 639–650.

Sebé-Pedrós, A., Ballaré, C., Parra-Acero, H., Chiva, C., Tena, J.J., Sabidó, E., Gómez-Skarmeta, J.L., Di Croce, L., and Ruiz-Trillo, I. (2016). The dynamic regulatory genome of Capsaspora and the origin of animal multicellularity. Cell *165*, 1224–1237.

Simola, D.F., Graham, R.J., Brady, C.M., Enzmann, B.L., Desplan, C., Ray, A., Zwiebel, L.J., Bonasio, R., Reinberg, D., Liebig, J., and Berger, S.L. (2016). Epigenetic (re)programming of caste-specific behavior in the ant Camponotus floridanus. Science *351*, aac6633.

Slattery, M., Ma, L., Négre, N., White, K.P., and Mann, R.S. (2011). Genome-wide tissue-specific occupancy of the Hox protein Ultrabithorax and Hox cofactor Homothorax in Drosophila. PLoS ONE *6*, e14686.

Smit, A.F.A., Hubley, R., and Green, P. 2013. RepeatMasker Open-4.0. http://www.repeatmasker.org/.

Su, J., Teichmann, S.A., and Down, T.A. (2010). Assessing computational methods of cis-regulatory module prediction. PLoS Comput. Biol. *6*, e1001020.

Taher, L., McGaughey, D.M., Maragh, S., Aneas, I., Bessling, S.L., Miller, W., Nobrega, M.A., McCallion, A.S., and Ovcharenko, I. (2011). Genome-wide identification of conserved regulatory function in diverged sequences. Genome Res. *21*, 1139–1149.

Tobler, A., Kapan, D., Flanagan, N.S., Gonzalez, C., Peterson, E., Jiggins, C.D., Johntson, J.S., Heckel, D.G., and McMillan, W.O. (2005). First-generation linkage map of the warningly colored butterfly Heliconius erato. Heredity (Edinb) *94*, 408–417.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.

Wahlberg, N., Leneveu, J., Kodandaramaiah, U., Peña, C., Nylin, S., Freitas, A.V.L., and Brower, A.V.Z. (2009). Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. Proc. Biol. Sci. *276*, 4295–4302.

Wahlberg, N., Wheat, C.W., and Peña, C. (2013). Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). PLoS ONE *8*, e80875.

Wittkopp, P.J., and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. *13*, 59–69.

Zhan, S., Merlin, C., Boore, J.L., and Reppert, S.M. (2011). The monarch butterfly genome yields insights into long-distance migration. Cell *147*, 1171–1185.

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

Zhen, Y., and Andolfatto, P. (2012). Methods to detect selection on noncoding DNA. Methods Mol. Biol. *856*, 141–159.