Genome-Wide Characterization of Adaptation and Speciation in Tiger Swallowtail Butterflies Using De Novo Transcriptome Assemblies

Wei Zhang[1], Krushnamegh Kunte[2], and Marcus R. Kronforst[1,*]

[1]Department of Ecology & Evolution, University of Chicago

[2]National Center for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru, Karnataka, India

*Corresponding author: E-mail: mkronforst@uchicago.edu.

Accepted: May 26, 2013

Data deposition: This project has been deposited at NCBI SRA under the accession number SRP022555.

Abstract

Hybrid speciation appears to be rare in animals, yet characterization of possible examples offers to shed light on the genomic consequences of this unique phenomenon, as well as more general processes such as the role of adaptation in speciation. Here, we first generate transcriptome assemblies for a putative hybrid butterfly species, *Papilio appalachiensis*, its parental species, *P. glaucus* and *P. canadensis*, and an outgroup, *P. polytes*. Then, we use these data to infer genome-wide patterns of introgression and genomic mosaicism using both phylogenetic and population genetic approaches. Our results reveal that there is little genetic divergence among all three of the focal species, but the subset of gene trees that strongly support a specific tree topology suggest widespread sharing of genetic variation between *P. appalachiensis* and both parental species, likely as a result of hybrid speciation. We also find evidence for substantial shared genetic variation between *P. glaucus* and *P. canadensis,* which may be due to gene flow or ancestral variation. Consistent with previous work, we show that *P. applachiensis* is more similar to *P. canadensis* at Z-linked genes and more similar to *P. glaucus* at mitochondrial genes. We also identify a variety of targets of adaptive evolution, which appear to be enriched for traits that are likely to be important in the evolution of this butterfly system, such as pigmentation, hormone sensitivity, developmental processes, and cuticle formation. Overall, our results provide a genome-wide portrait of divergence and introgression associated with adaptation and speciation in an iconic butterfly radiation.

Key words: transcriptome, adaptation, hybrid speciation, introgression, *Papilio*.

Introduction

Young evolutionary radiations offer a special opportunity to explore the interplay among adaptation, speciation, and hybridization in generating biological diversity (Grant 1999; Seehausen 2006; Mallet 2009). One particular phenomenon that appears to rely on a mix of these evolutionary processes is hybrid speciation, which is the formation of a new species as a result of hybridization between two parental species (Mallet 2007; Abbott and Rieseberg 2012). Hybrid speciation is common in plants, where it frequently occurs via allopolyploidy, or a change in chromosome number between parental species and the hybrid offspring (Rieseberg 1997; Soltis PS and Soltis DE 2009; Abbott and Rieseberg 2012). In animals, hybrid speciation appears to be relatively rare and many of the examples that do exist appear to be homoploid hybrid species, having the same chromosome number as their parental taxa (Mallet 2007; Mavárez and Linares 2008).

Although hybrid speciation may only account for a small fraction of the species diversity in animals, careful study of this phenomenon can provide more general insights into the origin of reproductive isolation and the potential role of adaptive evolution in the speciation process. This is because incipient homoploid hybrid species face a variety of challenges that are likely to inhibit their persistence (Abbott and Rieseberg 2012). One of these challenges is reproductive isolation. Unlike allopolyploids, homoploid hybrid species are not immediately reproductively isolated from their parental species, and because they must originate in contact with the parental species, it may be difficult for a new hybrid lineage

to remain distinct and not simply fuse with a parental species by backcrossing. A second challenge is competitive exclusion. Those hybrid lineages that do manage to remain distinct in the face of potential gene flow with parental species must then secure resources and survive in an environment already occupied by the parental taxa. Given the factors acting against the origin of new hybrid species, examination of those hybrid species that have persisted to the present day may inform us as to how reproductive isolation and niche evolution occur on short time scales.

A recently described hybrid butterfly species, *Papilio appalachiensis*, appears to overcome both challenges by occupying a novel environment in which it has higher fitness than its parental species. The parental species, *P. glaucus* and *P. canadensis*, are sister species with parapatric distributions that share a narrow hybrid zone along the border between the United States and Canada (Hagen et al. 1991; Luebke et al. 1988). Although earlier studies considered *P. canadensis* a subspecies of *P. glaucus*, more recent work has documented pronounced reproductive isolation between them, including intrinsic postzygotic isolation (Sperling 1993; Hagen and Scriber 1995; Scriber et al. 1995). A wide variety of additional differences, including divergent habitat (Lederhouse et al. 1995) and host plant preferences (Scriber et al. 1995; Scriber 1996), larval development (Ritland and Scriber 1985), allozyme allele frequencies (Hagen and Scriber 1991), AFLP markers (Winter and Porter 2010; Kunte et al. 2011), and DNA sequence data (Kunte et al. 2011) further support *P. canadensis* as a separate species. One striking morphological difference between *P. glaucus* and *P. canadensis* involves wing pattern mimicry (fig. 1). *Papilio glaucus* females display two distinct wing patterns; a yellow, nonmimetic phenotype that looks like the males and a melanic phenotype that mimics the chemically defended Pipevine swallowtail *Battus philenor* (Brower 1958). In contrast, *P. canadensis* lacks the mimetic female morph with both males and females displaying a similar yellow wing pattern (Hagen et al. 1991). The color of *P. glaucus* females is controlled by a W-linked Mendelian locus and it is further influenced by a Z-linked enabler/suppressor locus that differs between *P. glaucus* and *P. canadensis* (Scriber and Hainze 1987; Hagen and Scriber 1989; Scriber et al. 1996).

Recently described *P. appalachiensis* (Pavulaan and Wright 2002) exists at high elevation along the Appalachian Mountains and appears to be a hybrid species (Scriber and Ording 2005). Like *P. canadensis*, it is adapted to a cooler thermal zone and it is univoltine. However, like *P. glaucus*, it displays two female morphs, one of which is a dark, mimetic form (Pavulaan and Wright 2004). This unique combination of traits allows this species to occupy a novel, high elevation habitat that is within the range of the mimicry model *B. philenor*. Using a combination of targeted DNA sequencing and AFLP genotyping, Kunte et al. (2011) recently showed that *P. appalachiensis* is a genomic mixture of *P. glaucus* and *P. canadensis* and that it is significantly differentiated from both. As a whole, these results suggest that historical hybridization between *P. glaucus* and *P. canadensis* produced a stable hybrid lineage well adapted to a novel environment,

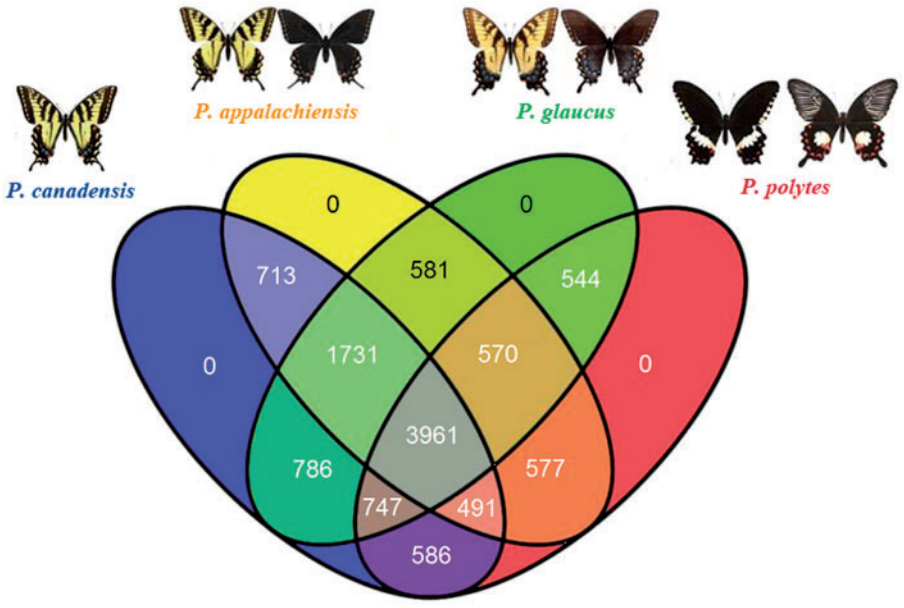

Fig. 1.—Distribution of conserved clusters among the four butterfly species. Conserved clusters were retrieved from predicted CDS data sets using Blat. A total of 3,961 clusters yielded a single sequence for each species and this set of conserved clusters was the core data set for subsequent analyses. Each species is depicted with images of female wing pattern phenotypes.

which persists today as a reproductively isolated species, *P. appalachiensis*.

The genetic data supporting a mosaic genome in *P. appalachiensis* are still rather limited. Therefore, we have focused on fully characterizing the transcriptomes of *P. appalachiensis*, its putative parental species, *P. glaucus* and *P. canadensis*, and an outgroup species, *P. polytes*. We then use these data to examine genome-wide patterns of divergence and genomic mosaicism among the tiger swallowtails using a variety of analytical approaches. We also infer rates of gene flow among the three taxa and characterize genes that have experienced recent positive selection. Our results lend strong support to the hypothesis that *P. appalachiensis* is a hybrid species and provide important insights into the potential functional genetic changes associated with speciation in this well-studied butterfly group.

Materials and Methods

RNA Isolation and Illumina Sequencing

We generated RNA-seq data for a total of eight pupal RNA samples; two *P. glaucus*, two *P. canadensis*, two *P. appalachiensis*, and two *P. polytes*. *Papilio polytes* and the ingroup taxa come from different within-*Papilio* subclades that may have diverged from one another approximately 35 Ma (Zakharov et al. 2004). The *P. polytes* samples were selected from a lab colony originating from the Philippines, whereas the other species were field-collected in Louisiana (*P. glaucus*), New Hampshire (*P. canadensis*), and West Virginia (*P. appalachiensis*). For *P. polytes*, RNA was extracted from pupal wing discs only, whereas RNA was extracted from entire pupae for the other samples. RNA was extracted with Trizol according to a standard protocol, poly-A purified, and converted to cDNA and barcoded using the Illumina Tru-Seq protocol. The cDNA libraries were then pooled and sequenced using an Illumina HiSeq 2000 sequencer (100 bp paired-end).

De Novo Transcriptome Assembly

Raw reads were demultiplexed according to their barcodes and low quality sequences were removed before assembly. After quality filtering, data were combined by species. Trinity version 2012-06-08 (Grabherr et al. 2011) was used to perform de novo transcriptome assembly and open reading frame extraction under default parameters. These parameters include the following: min_contig_length 200, min_kmer_cov 1, max_reads_per_graph 200000, max_number_of_paths_per_node 10, group_pairs_distance 500, and path_reinforcement_distance 75.

Clustering and Annotating Conserved Coding Sequences

To identify clusters of homologous sequences among transcriptomes, predicted coding sequence (CDS) regions of each species were used as queries and targets separately for Blat (Kent 2002) to search against data sets for the other three species (reciprocal best hits). The best Blat hits of the longest isoforms with *E* value lower than 10^{-6} were retrieved and only one-to-one orthologous genes existing in all four species were retained. These are the conserved clusters, or "genes," used in all further analyses. Note that adjusting the *E* value threshold to 10^{-15} reduced the final data set by only 1%. Clusters that contained two or more sequences from the same species were not analyzed further to eliminate potential issues stemming from paralogs. Conserved mitochondrial genes and rRNAs were identified using each transcriptome data set as query for Blat searches against predicted genes and rRNAs in the mitochondrial genome of *Bombyx mandarina* (NCBI Reference Sequence: AY301620.2) (Pan et al. 2008).

Multiple Alignments and Phylogenetic Analysis

Multiple Alignments

We performed two separate analyses of our conserved clusters, one based on alignment of nucleotide sequences and another based on alignment of predicted peptide sequences. Multiple sequence alignments for both data sets were performed using MUSCLE 3.8 (Edgar 2004) with default parameters.

Topological Structure Assignment

To infer the best tree topology for each conserved cluster, we estimated phylogenetic trees using both maximum-likelihood (ML) and neighbor-joining (NJ) methods. First, we used PhyML 3.0 (Guindon et al. 2010) to generate trees under specified topological constraints with either K2P (for nucleotide) or JTT (for protein) models of evolution. The three constrained trees were defined as ((A,C),G,P), ((A,G),C,P), and ((C,G),A,P) where A, C, G, and P stand for *P. appalachiensis*, *P. canadensis*, *P. glaucus,* and *P. polytes,* respectively. We used CONSEL 0.20 (Shimodaira and Hasegawa 2001) to assess the confidence of selecting the best topological structure for each cluster. CONSEL was used to generate Shimodaira and Hasegawa (SH) test *P* values for each tree topology with *P* values ≥ 0.95 indicating significant support for a particular topology. We further estimated trees for all clusters using NJ method, with 1,000 bootstrap pseudoreplicates, using PHYLIP 3.69 (Felsenstein 1989). We focused our subsequent analyses on conserved clusters for which ML, combined with the SH test, and the NJ tree yielded the same tree topology. Similar methods were used to examine tree topologies for specific clusters, which we inferred to be Z-linked, by comparison with the *Heliconius melpomene* genome sequence (*Heliconius* Genome Consortium 2012), or mitochondrial, by comparison with the *Bombyx* mitochondrial genome. All sequence descriptions are based on results of BLASTX searches against NCBI's nr protein database. Gene ontology terms were assigned to conserved clusters using Blast2GO (Conesa et al.

2005) and Fisher's exact tests were used to test for functional enrichment for clusters yielding each of the three tree topologies. Note, setting a false-discovery rate to correct for multiple testing resulted in no significant enrichment. Huang et al. (2008) suggest that multiple testing corrections may be too conservative to effectively guide initial exploratory analyses so we present the uncorrected *P* values for all GO term enrichment tests.

Detecting Gene Flow among *P. glaucus*, *P. canadensis*, and *P. appalachiensis*

We calculated Patterson's *D*-statistic (Green et al. 2010; Durand et al. 2011) to quantify gene flow among the three ingroup taxa, using *P. polytes* as an outgroup. This test examines the phylogenetic distribution of derived alleles (designated "B") at loci that display either an ABBA or BABA configuration on a four species phylogeny (fig. 2). Summed

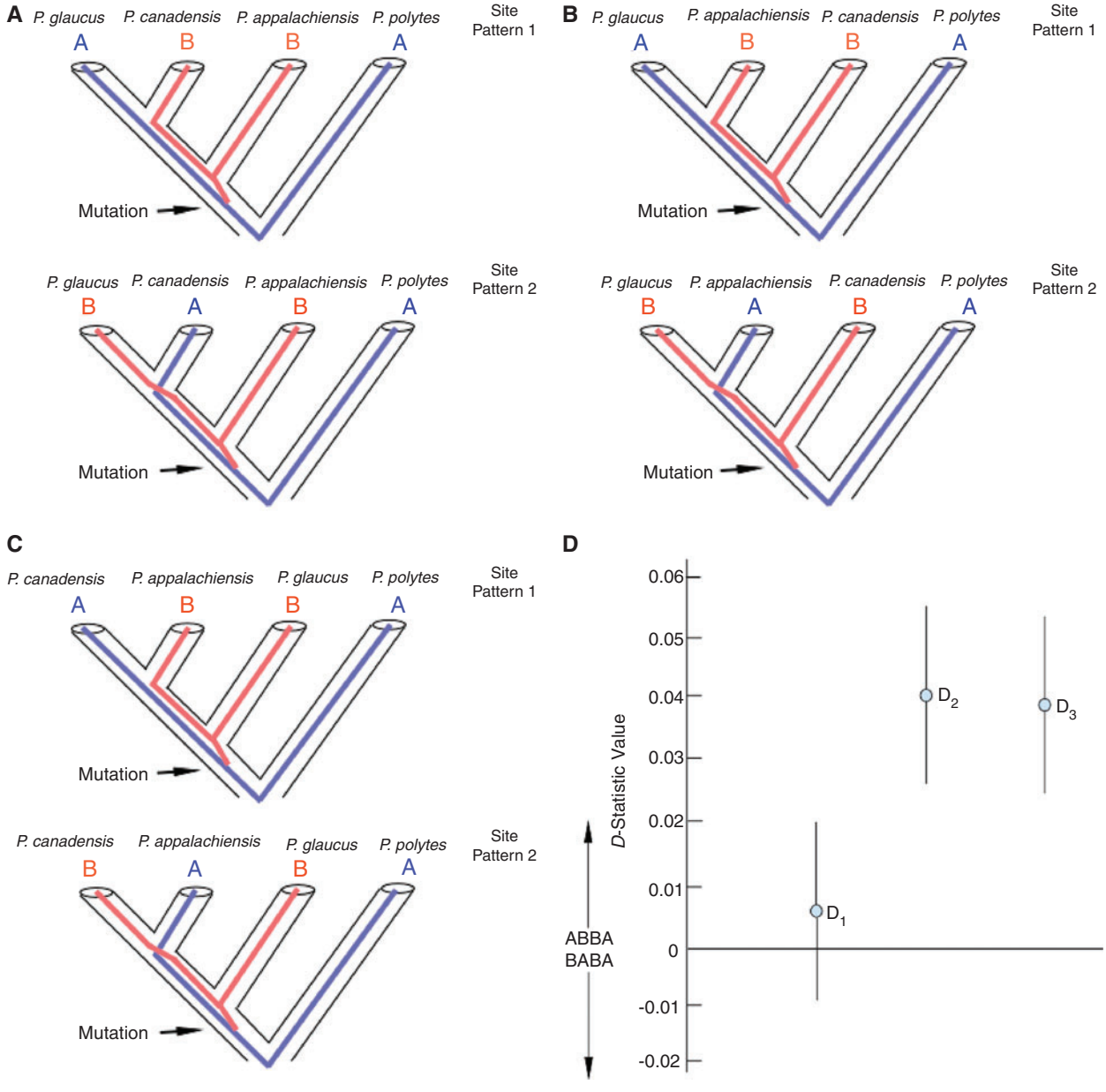

Fig. 2.—Patterson's *D*-statistic suggests widespread introgression between *Papilio appalachiensis* and the putative parental species. We calculated a transcriptome-wide *D*-statistic value for each of three tree topologies (*A*–*C*) and found evidence for significant introgression in comparisons with *P. appalachiensis* (*D*). Results suggest roughly equal introgression between *appalachiensis/canadensis*, compared with *appalachiensis/glaucus* (D_1, $P = 0.715$), but much more introgression between *appalachiensis/canadensis* and *appalachiensis/glaucus,* compared with *glaucus/canadensis* (D_2 and D_3, $P < 0.01$ for both).

across the genome, no enrichment of ABBA or BABA sites is expected as a result of random sorting of ancestral variation. Interspecific gene flow, however, is expected to result in a systematic bias of allele sharing between the two taxa exchanging alleles. For these tests, single nucleotide polymorphisms (SNPs) were extracted from each cluster using ape-package 3.0-6 (Paradis et al. 2004) and adegenet-package 1.3-5 (Jombart and Ahmed 2011). Then, separate tests were performed to detect gene flow in pairwise comparisons among our three ingroup taxa. The number of shared, derived SNPs supporting either an ABBA or BABA pattern was calculated in three comparisons: D_1 (*glaucus*, *canadensis*, *appalachiensis*, and *polytes*), D_2 (*glaucus*, *appalachiensis*, *canadensis*, and *polytes*), and D_3 (*canadensis*, *appalachiensis*, *glaucus*, and *polytes*). Finally, the leave-one-out jackknife estimate was performed using bootstrap-package 2012-04-0 (Tibshirani and Leisch 2012) to determine the standard error for each *D* value of each cluster and significant deviations from zero were tested using a two tailed *z*-test. *D*-statistic values that differ from zero are indicative of gene flow.

Chromosome Distribution of Conserved Clusters

Although there is no reference genome sequence for *Papilio* butterflies, we used the fact that synteny is highly conserved between the butterfly *H. melpomene* and the moth *B. mori* (*Heliconius* Genome Consortium 2012) to examine the genome-wide distribution of our conserved clusters (fig. 3). We also tested whether clusters with the same tree topology were clustered in the genome. To do this, we used Blat to assign conserved clusters to putative orthologs in the *B. mori* and *H. melpomene* genome sequences. *Bombyx mori* genome data were downloaded from SilkDB (http://www.silkdb.org/silkdb/, last accessed June 17, 2013) (Xia et al. 2004) and *H. melpomene* genome data were downloaded from the Butterfly Genome Database (http://www.butterflygenome.org/, last accessed June 17, 2013). We used Spearman Rank Correlation tests to compare the chromosomal-level distribution of clusters with a particular tree topology to a null distribution based on the distribution of all conserved clusters. This analysis was done twice, once using the *Bombyx* genome as a reference and once using the *Heliconius* genome.

Calculating K_a/K_s Ratios for Conserved Clusters

For each conserved cluster, we calculated nonsynonymous (K_a) and synonymous (K_s) substitution rates for every species pair. K_a and K_s were estimated using the unbiased approximation of Li (1993), implemented in seqinr-package 3.0-6 (Charif and Lobry 2007). We performed separate analyses, looking for evidence of positive selection between ingroup (*P. glaucus*, *P. canadensis*, and *P. appalachiensis*) and outgroup (*P. polytes*) taxa as well as among ingroup taxa.

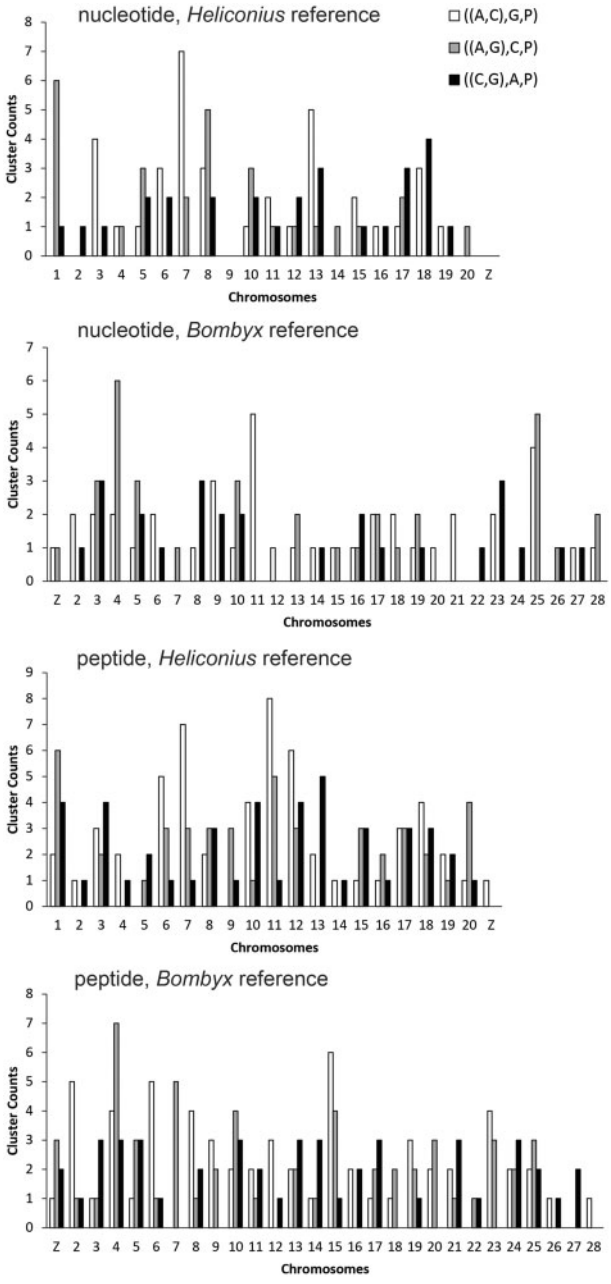

Fig. 3.—Genome-wide distribution and clustering of genes by tree topology. We mapped conserved clusters back to the genome of *Heliconius melpomene* (*A* and *C*) and *Bombyx mori* (*B* and *D*) and compared the chromosome-level distribution of clusters with a given tree topology with the null distribution given by all mapped clusters. The results of these tests are in table 6. (*A*, *B*) Tree topologies based on nucleotide alignments whereas (*C*) and (*D*) are based on peptide alignments.

Clusters yielding large K_a/K_s ratios were checked manually to eliminate spurious results due to poor alignment. Blast2GO was used to test for functional enrichment of clusters displaying evidence of positive selection between ingroup and outgroup taxa.

McDonald–Kreitman Tests

As an additional test of adaptive protein evolution, we performed McDonald–Kreitman (MK) tests (McDonald and Kreitman 1991) on the subset of clusters for which we could identify orthologous transcripts for each individual. To do this, we reassembled transcriptomes for each individual using Trinity, as opposed to combining data by species. Clustering and multiple alignments were performed as described earlier for the combined analysis. We performed two analyses, first comparing among ingroup taxa using all clusters for which we identified one sequence from each of the six ingroup samples, and then comparing ingroup taxa with the outgroup at the subset of these clusters where we could also identify one sequence from each *P. polytes* sample. MK tests was done using libsequence and MKtest package (Thornton 2003) and statistical significance was inferred using Fisher's exact test (P values < 0.05). Clusters were annotated using Blast2GO.

Results

Transcriptome Assembly and Conserved Cluster Characterization

We generated between 45 million and 148 million reads per sample, yielding approximately 4.5–14.8 Gb of RNA-seq data per sample. De novo transcriptome assembly for each species yielded a large number of putative single-copy genes (table 1) and combining data among species yielded 3,961 conserved clusters for which all four species contributed a single sequence (fig. 1). The mean CDS of these conserved clusters was 1,392 bp for *P. appalachiensis*, 1,376 bp for *P. canadensis*, 1,378 bp for *P. glaucus*, and 1,336 bp for *P. polytes*. For comparison, the mean CDS is 1,258 and 1,248 bp for all genes in the reference genome sequence of *H. melpomene* and *B. mori*, respectively. Comparisons using the "ortholog hit ratio" (O'Neil et al. 2010) further suggest that our conserved clusters largely span entire genes (supplementary fig. S1, Supplementary Material online). Note that using a much more stringent threshold for ortholog detection, an E value of 10^{-15}, altered the final data set very little (3,920 clusters compared with 3,961).

Mosaic Transcriptome of *P. appalachiensis*

Previously, Kunte et al. (2011) demonstrated that Z-linked genes connected *P. appalachiensis* to *P. canadensis* while mitochondrial genes, and presumably W-linked genes (these are linked in butterflies because females are the heterogametic sex), connected *P. appalachiensis* to *P. glaucus*. We first verified these findings by surveying our conserved clusters for putatively Z-linked and mitochondrial genes, yielding 18 Z-linked genes and 14 mtDNA genes. We found that many clusters did not yield statistically significant tree topologies based on the SH test, and nucleotide and peptide alignments did not always agree on the best tree topology. However, consistent with previous results, the most frequent tree topology for Z-linked genes was that linking *P. appalachiensis* and *P. canadensis* (table 2) while the most frequent tree topology for mitochondrial genes was that linking *P. appalachiensis* and *P. glaucus* (table 3).

To examine potential mosaicism across the *P. appalachiensis* genome as a whole, we performed similar phylogenetic analysis of all 3,961 conserved clusters. Because our ingroup taxa are very closely related, the vast majority of clusters did not yield a highly supported tree topology. Indeed, only 179 clusters yielded well-supported tree topologies (SH test plus NJ tree corroboration) in our analysis of the nucleotide data and 303 clusters in the analysis of peptide data. Interestingly, in analysis of both data sets, a similar number of clusters supported all three topologies (table 4). This result is consistent with the mosaic genome expected for *P. appalachiensis* but it also suggests extensive sharing between *P. glaucus* and *P. canadensis*. This may be a result of long-term hybridization between these two species where their ranges overlap. Conserved clusters belonging to each of the three topologies were enriched for a variety of GO terms (table 5).

Table 1
Transcriptome Assembly Results

Sample	Total Transcripts	Longest Isoform	Predicted CDS	Unique Genes
P. appalachiensis	102,375	53,198	36,879	10,179
P. canadensis	146,954	76,471	48,092	10,624
P. glaucus	124,664	57,509	43,843	10,240
P. polytes	108,707	72,920	35,750	9,704

Gene Flow among *P. glaucus*, *P. canadensis*, and *P. appalachiensis*

To verify our phylogenetic signatures of shared genetic variation among the three tiger swallowtail species, we investigated potential introgression among species using Patterson's D-statistic (Green et al. 2010; Durand et al. 2011). To do this, we counted derived SNP alleles supporting either "ABBA" or "BABA" patterns among the in-group taxa and then calculated the mean D value across our conserved clusters (fig. 2). These results suggest substantial and nearly equal amounts of gene flow between *P. glaucus* and *P. appalachiensis,* compared with *P. canadensis* and *P. appalachiensis* ($P < 0.01$ for both comparisons). These results also suggest that the amount of gene flow between *P. glaucus* and *P. canadensis* is low compared with between each of these species and *P. appalachiensis.*

Chromosome Distribution of Conserved Clusters

We mapped clusters with different tree topologies back to the reference genome sequence for *H. melpomene*

Table 2
Tree Topologies of Z-linked Clusters

Cluster ID	Nucleotide Alignment		Peptide Alignment		Annotation
	Topological Structure	*P* Value	Topological Structure	*P* Value	
105	((A,G),C,P)	0.94	((A,C),G,P)	1	ww domain-containing adapter protein with coiled-coil
611	((A,C),G,P)*	0.8	((A,C),G,P)*	0.77	Scabrous protein
930	((A,G),C,P)	0.87	NA[a]	—	Putative flotillin-1
1294	((A,C),G,P)*	0.79	((A,C),G,P)*	0.83	Secernin 3
1617	((A,C),G,P)*	0.83	((A,C),G,P)*	0.96	Catalase
1660	((A,C),G,P)	0.97	NA	—	Ankyrin repeat domain-containing protein 12
2021	((A,C),G,P)*	0.95	((A,C),G,P)*	0.75	Disulfide-isomerase a5
2055	((A,C),G,P)*	0.79	((A,C),G,P)*	0.59	Hepatic leukemia factor
3130	((A,C),G,P)*	0.86	((A,C),G,P)*	0.77	Serine threonine-protein kinase osr1-like
3347	((C,G),A,P)	0.82	((A,C),G,P)*	0.63	Tyrosine hydroxylase
3361	((A,C),G,P)*	0.68	NA	—	Y-box protein
3703	((A,C),G,P)*	0.72	((A,C),G,P)*	0.9	Tyrosine-protein kinase abl-like
4566	((C,G),A,P)	0.79	((A,C),G,P)*	0.78	Acetyl-synthetase
4569	((A,C),G,P)*	0.72	((A,C),G,P)	1	Dipeptidase 1-like
4894	((A,C),G,P)	0.87	((C,G),A,P)	0.97	Serine threonine-protein kinase osr1-like
5837	((A,C),G,P)*	0.78	((A,C),G,P)*	0.83	Protein daughter of sevenless
6828	((A,C),G,P)*	0.81	((A,G),C,P)	0.73	Carboxypeptidase N subunit 2-like
6895	((A,C),G,P)*	0.94	((A,C),G,P)*	0.88	Kettin

Note.—Z-linked conserved clusters were identified by comparison with predicted CDS of Z-linked genes in the *Heliconius melpomene* genome sequence. SH *P* values were calculated based on both nucleotide and peptide alignments.

[a]NA indicates no best topology because of the same highest value assigned to more than one topological structure.

*Indicates the tree topology was also supported by NJ method. Most of the tree structures not supported by NJ yielded an ((A,C),G,P) structure in the NJ tree.

Table 3
Tree Topologies of Mitochondrial Clusters

Gene	Nucleotide Alignment		Peptide Alignment	
	Topological Structure	*P* Value	Topological Structure	*P* Value
12s	((C,G),A,P)*	0.824	NA[a]	—
16s	((A,G),C,P)*	0.744	NA	—
ATP6	((A,G),C,P)*	0.749	((C,G),A,P)	0.498
COI	((A,G),C,P)*	0.673	((A,C),G,P)	0.547
COII	((A,G),C,P)*	0.748	NA	—
COIII	((C,G),A,P)	0.711	((A,C),G,P)	0.844
cytB	((A,G),C,P)*	0.797	((A,G),C,P)*	0.779
ND1	((A,G),C,P)*	0.578	((C,G),A,P)	1
ND2	((A,C),G,P)	0.986	((A,G),C,P)*	1
ND3	((C,G),A,P)*	0.866	((C,G),A,P)*	0.792
ND4	((A,G),C,P)*	0.617	((A,G),C,P)	1
ND4L	((A,G),C,P)*	0.763	NA	—
ND5	((C,G),A,P)	0.961	((A,G),C,P)	0.818
ND6	((A,G),C,P)*	0.818	NA	—

Note.—Mitochondrial conserved clusters were identified by comparison with predicted mitochondrial CDS or rRNA. SH *P* values were calculated based on both nucleotide and peptide alignments.

*Indicates the tree topology was also supported by NJ. Most of the tree structures not supported by NJ yielded ((A,G),C,P) structure in the NJ tree.

[a]NA indicates no peptide alignment because untranslated RNA sequence (12s and 16s rRNA) or no best topology because of the same highest value assigned to more than one topological structure.

Table 4
Number of Conserved Clusters with Well-Supported Tree Topologies

	Topological Structure		
	((A,C),G,P)	((A,G),C,P)	((C,G),A,P)
Nucleotide	71	58	50
Peptide	113	93	97
Shared	27	19	22

Note.—Counts were calculated based on either peptide or nucleotide alignment with the "shared" counts appearing in both groups.

and *B. mori* and then compared chromosomal clustering relative to the null hypothesis based on the chromosomal distribution of all conserved clusters. Of 3,961 conserved clusters, we were able to uniquely map 1,884 to the *Heliconius* genome and 2,101 to the *Bombyx* genome. The results of this analysis suggest that conserved clusters with the same tree topology are likely to be clustered in the *Papilio* genome (table 6). It is important to note that the results are only suggestive of true clustering because this analysis rests on extrapolating the highly conserved synteny between *Heliconius* and *Bombyx* to *Papilio*, a group for which no genome sequence currently exists.

Table 5
Functional Enrichment of Conserved Clusters with Various Topological Structures

Topology	GO Term	Category	Type[a]	*P* Value
Nucleotide alignment				
((A,C),G,P)	GO:0016301	Kinase activity	F	0.005
	GO:0016772	Transferase activity, transferring phosphorus-containing groups	F	0.005
	GO:0004672	Protein kinase activity	F	0.006
	GO:0016773	Phosphotransferase activity, alcohol group as acceptor	F	0.006
	GO:0016740	Transferase activity	F	0.016
	GO:0006091	Generation of precursor metabolites and energy	P	0.028
((A,G),C,P)	GO:0007049	Cell cycle	P	0.001
	GO:0006996	Organelle organization	P	0.001
	GO:0071842	Cellular component organization at cellular level	P	0.001
	GO:0071841	Cellular component organization or biogenesis at cellular level	P	0.001
	GO:0051716	Cellular response to stimulus	P	0.002
	GO:0050794	Regulation of cellular process	P	0.002
	GO:0007165	Signal transduction	P	0.002
	GO:0050896	Response to stimulus	P	0.005
	GO:0009987	Cellular process	P	0.006
	GO:0007005	Mitochondrion organization	P	0.008
	GO:0023052	Signaling	P	0.009
	GO:0065007	Biological regulation	P	0.009
	GO:0006811	Ion transport	P	0.010
	GO:0005215	Transporter activity	F	0.011
	GO:0030234	Enzyme regulator activity	F	0.012
	GO:0016043	Cellular component organization	P	0.012
	GO:0071840	Cellular component organization or biogenesis	P	0.012
	GO:0032501	Multicellular organismal process	P	0.020
	GO:0007275	Multicellular organismal development	P	0.020
	GO:0050789	Regulation of biological process	P	0.022
	GO:0032502	Developmental process	P	0.041
((C,G),A,P)	GO:0045182	Translation regulator activity	F	0.027
	GO:0035556	Intracellular signal transduction	P	0.040
Peptide alignment				
((A,C),G,P)	GO:0005623	Cell	C	0.000
	GO:0044464	Cell part	C	0.004
	GO:0005622	Intracellular	C	0.005
	GO:0007267	Cell–cell signaling	P	0.015
	GO:0005811	Lipid particle	C	0.019
	GO:0016209	Antioxidant activity	F	0.028
	GO:0007154	Cell communication	P	0.049
((A,G),C,P)	GO:0008283	Cell proliferation	P	0.011
	GO:0007005	Mitochondrion organization	P	0.022
	GO:0004518	Nuclease activity	F	0.023
	GO:0030528	Transcription regulator activity	F	0.027
	GO:0016032	Viral reproduction	P	0.043
	GO:0016788	Hydrolase activity, acting on ester bonds	F	0.046
((C,G),A,P)	GO:0007005	Mitochondrion organization	P	0.024

[a]F, P, and C stand for molecular function, biological process, and cellular component, respectively.

Genes under Positive Selection

A general method to test for positive selection is based on likelihood ratio tests, but this is not a powerful approach with few sequences (Anisimova et al. 2001). Because we only had four sequences in each cluster, we calculated the K_a/K_s ratio for each conserved cluster and considered those with a value more than 1 to be candidates for positive selection (Li 1993).

Table 6

Genomic Clustering of Genes Based on Inferred Tree Topology

Tree Topology	*Heliconius melpomene* Reference		*Bombyx mori* Reference	
	Nucleotide Alignment	Peptide Alignment	Nucleotide Alignment	Peptide Alignment
((A,C),G,P)	0.082	0.001	0.189	0.011
((A,G),C,P)	0.033	0.121	0.029	0.014
((C,G),A,P)	0.005	0.001	0.149	0.066

NOTE.—*P* values reported above are based on Spearman's rank correlation tests, comparing the chromosomal distributions of clusters with a given tree topology to the distribution of all clusters, using both *H. melpomene* and *B. mori* as a reference for chromosomal locations. Tree topologies were inferred using both nucleotide and peptide alignments.

In comparisons with the outgroup, 275 clusters yielded K_a/K_s ratios more than 1 in all three pairwise comparisons. The functional enrichment of these clusters yielded a variety of terms related to RNA/DNA modification, ion binding and transportation, cell cycle regulation, pigment metabolism, and hormone regulation (table 7). We further examined clusters for evidence of positive selection among the ingroup taxa, which yielded a small number of candidate genes (table 8). Interestingly, there was considerable overlap in the gene sets that emerged from our analysis comparing ingroup taxa, with those that exhibited high K_a/K_s ratios in comparisons between ingroup and outgroup species, suggesting some recurrent targets of selection. For those that did not overlap, we were particularly interested in genes showing evidence of selection in two pairwise ingroup comparisons, which would suggest adaptive evolution along a single lineage. This pattern could also result from divergent selection between *P. glaucus* and *P. canadensis* followed by introgression from one of those species into *P. appalachiensis*. Regardless, this approach yielded a list of candidate genes with some apparent enrichment of functions related to mitosis, ecdyteroid-induction, and cuticular proteins.

Transcriptome assembly and clustering at the individual level, for MK tests, yielded 2,551 conserved clusters that included one sequence for each ingroup sample. Of these 2,225 also contained a single sequence for each *P. polytes* sample and so could be used to compare ingroup and outgroup taxa. A total of 56 clusters yielded significant ($P < 0.05$) MK tests between one or more ingroup taxa, and another 18 were significant in all ingroup versus outgroup comparisons (supplementary table S1, Supplementary Material online). Interestingly, there was only a single instance of overlap between the K_a/K_s and MK results, with a gene annotated as protein phosphatase regulatory subunit b gamma appearing in both ingroup versus outgroup comparisons. One factor that may contribute to the low overlap between K_a/K_s and MK results is the modest overlap in the data sets themselves. For instance, of the 62 clusters yielding significant K_a/K_s results

Table 7

Functional Enrichment of Conserved Clusters under Positive Selection between Ingroup and Outgroup

GO Term	Category	Type[a]	*P* Value
GO:0006306	DNA methylation	P	0.004
GO:0006305	DNA alkylation	P	0.004
GO:0006304	DNA modification	P	0.004
GO:0051238	Sequestering of metal ion	P	0.004
GO:0045448	Mitotic cell cycle, embryonic	P	0.005
GO:0043169	Cation binding	F	0.008
GO:0043167	Ion binding	F	0.008
GO:0071383	Cellular response to steroid hormone stimulus	P	0.008
GO:0030003	Cellular cation homeostasis	P	0.012
GO:0030684	Preribosome	C	0.012
GO:0046915	Transition metal ion transmembrane transporter activity	F	0.012
GO:0070851	Growth factor receptor binding	F	0.012
GO:0035186	Syncytial blastoderm mitotic cell cycle	P	0.012
GO:0008173	RNA methyltransferase activity	F	0.012
GO:0004887	Thyroid hormone receptor activity	F	0.012
GO:0007394	Dorsal closure, elongation of leading edge cells	P	0.012
GO:0046914	Transition metal ion binding	F	0.013
GO:0046872	Metal ion binding	F	0.013
GO:0055080	Cation homeostasis	P	0.017
GO:0008270	Zinc ion binding	F	0.020
GO:0007050	Cell cycle arrest	P	0.021
GO:0032870	Cellular response to hormone stimulus	P	0.022
GO:0006726	Eye pigment biosynthetic process	P	0.024
GO:0031163	Metallo-sulfur cluster assembly	P	0.024
GO:0033301	Cell cycle comprising mitosis without cytokinesis	P	0.024
GO:0031099	Regeneration	P	0.024
GO:0016226	Iron–sulfur cluster assembly	P	0.024
GO:0000794	Condensed nuclear chromosome	C	0.024
GO:0072503	Cellular divalent inorganic cation homeostasis	P	0.024
GO:0007392	Initiation of dorsal closure	P	0.024
GO:0071495	Cellular response to endogenous stimulus	P	0.027
GO:0043324	Pigment metabolic process involved in developmental pigmentation	P	0.038
GO:0006497	Protein lipidation	P	0.038
GO:0042441	Eye pigment metabolic process	P	0.038
GO:0042158	Lipoprotein biosynthetic process	P	0.038
GO:0042157	Lipoprotein metabolic process	P	0.038
GO:0009826	Unidimensional cell growth	P	0.038
GO:0072507	Divalent inorganic cation homeostasis	P	0.038
GO:0043474	Pigment metabolic process involved in pigmentation	P	0.038
GO:0071156	Regulation of cell cycle arrest	P	0.038
GO:0003707	Steroid hormone receptor activity	F	0.039
GO:0005615	Extracellular space	C	0.039
GO:0004879	Ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	F	0.039
GO:0043401	Steroid hormone–mediated signaling pathway	P	0.039
GO:0048545	Response to steroid hormone stimulus	P	0.045
GO:0009755	Hormone-mediated signaling pathway	P	0.049
GO:0048066	Developmental pigmentation	P	0.049

NOTE.—GO terms enrichment of conserved clusters with K_a/K_s ratios above one in all three ingroup versus outgroup comparisons.

[a]F, P, and C stand for molecular function, biological process, and cellular component, respectively.

Table 8

Annotation of Clusters under Positive Selection among Ingroup Taxa

K_a/K_s Ratio	Cluster ID	Annotation
A vs. C >1,	869	Histone h1-like
A vs. G >1,	1537	Splicing factor arginine serine-rich 6
C vs. G <1	4761	Cuticle protein BmorCPR83 (BmEdg84A)
	5014	Polo
	6025	Spinophilin-like
	6201	Uncharacterized protein KIAA1841-like
A vs. C >1,	621	Zinc finger protein on ecdysone puffs
C vs. G >1,	1165	Vesicle associated
A vs. G <1	1475	Serine protease 14
	3028	Shaker-like potassium channel
	3111	Nuclear hormone receptor
	4179	Kinase d-interacting substrate of 220 kDa-like
	5888	NA[a]
	6892	Hypothetical protein KGM_07109 [*Danaus plexippus*]
A vs. G >1,	153	Pab-dependent poly-specific ribonuclease subunit 3-like
C vs. G >1,		
A vs. C <1	726	Tata-binding protein-associated phosphoprotein
	2649	NA
	2702	Ecdysone-induced protein 78c
	3483	40s ribosomal protein s3a
	3565	Pdz and lim domain protein 3
	4364	Follistatin
	4564	Encore protein
	6586	Tyrosine-protein kinase fps85d-like isoform 1
A vs. C >1,	24	Putative rRNA processing protein RRP7
A vs. G <1,	114	Hypothetical protein KGM_04049
C vs. G <1		[*D. plexippus*]
	497	Cuticular protein 76bd
	754	Hexokinase
	1392	Inositol-trisphosphate 3-kinase a-like
	1836	g-protein coupled receptor mth2-like
	2064	Cuticle protein BmorCPR141
	4222	Ankyrin repeat domain-containing protein 57
	4286	Hypothetical protein KGM_21585 [*D. plexippus*]
	4640	Rho guanine nucleotide exchange factor 7-like isoform 1
	4973	Adipocyte plasma membrane-associated protein
	5095	Unknown secreted protein [*Papilio xuthus*]
	5340	Katanin p80 wd40-containing subunit b1
	6084	NEDD4-binding protein 2-like
A vs. G >1,	22	Elongation factor 1 delta
A vs. C <1,	36	Chondroitin 4-sulfotransferase
C vs. G <1	854	Atp-binding cassette sub-family g member 1-like
	926	Naked cuticle-like protein
	1213	Serine proteinase-like protein 1
	2044	xpg-like endonuclease
	2069	RNA helicase-like protein
	2478	upf0712 protein c7orf64-like
	2542	DNA topoisomerase 3-beta-1
	2668	Mosc domain-containing protein mitochondrial-like
	5154	Lim domain-binding protein 3
	5774	Hypothetical protein KGM_14584 [*D. plexippus*]
	5993	DNA repair protein xp-c rad4
	6095	Tyrosine transporter
	6781	Speckle-type poz protein
C vs. G >1,	733	Serine protease
A vs. C <1,	1247	Protein sda1 homolog
A vs. G <1	1561	Unc-isoform a
	2898	Down syndrome cell adhesion molecule isoform d
	2956	Protein lethal denticleless-like
	3055	12 cysteine protein 1
	3313	Nuclear protein localization protein 4 homolog
	4064	tRNA dimethylallyltransferase mitochondrial-like
	5197	Tryptophanyl-tRNA synthetase mitochondrial-like
	6801	Acyl-CoA oxidase

(continued)

Table 8

Continued

Note—Three pairwise comparisons were made among *P. glaucus*, *P. canadensis*, and *P. appalachiensis* and clusters with one or two ratios >1 were selected. Highlighted cluster IDs also exhibited evidence of positive selection in comparisons between ingroup and outgroup taxa (table 5).

[a]NA indicates no BLASTX hit against NCBI's nr protein database.

among ingroup taxa (table 8), only 22 were included in the data set used for MK tests. Similarly, 275 clusters yielded significant K_a/K_s results in comparisons between ingroup and outgroup taxa, 137 of which were in the data set used for MK tests. Additional factors could inflate or bias our test statistics, further contributing to low overlap in the results. For instance, the K_a/K_s ratio was developed to compare sequences from divergent species and it is known to perform poorly when applied to intraspecific polymorphism data (Kryazhimskiy and Plotkin 2008). Given that our ingroup taxa are closely related, and appear to be exchanging genes, some of the sequence variation we are applying to the K_a/K_s tests is likely to be polymorphism, as opposed to fixed differences between species, which may bias the test. Furthermore, our MK tests are likely to be biased toward significant departures from neutrality because we have relatively little intraspecific data from which to estimate polymorphism information (Andolfatto 2008).

Discussion

Our phylogenetic approach to transcriptome analysis is conceptually straightforward in that we simply want to track the evolutionary relationships among our three focal species on a gene-by-gene basis by comparing the fit of each gene

with the three possible evolutionary scenarios. In practice, however, this approach presents a variety of challenges that we worked hard to overcome. First, analysis of transcriptome data in the absence of a reference genome sequence presents a serious obstacle, especially in terms of identifying orthologs. Given potentially high sequence similarity among paralogs, identifying orthologous genes across species based on sequence homology alone is difficult (Pepke et al. 2009). A more powerful method for identifying orthologs is based on comparing the identity and order of genes surrounding putative orthologs (Hulsen et al. 2006), but this information is not available from our transcriptome data. Therefore, we applied very stringent filters to our homology-based pipeline which should remove virtually all sequence clusters in which paralogs might be an issue. In particular, we assembled our conserved clusters based on sequence homology, using a stringent matching threshold for comparisons within and between species, and then we discarded any clusters in which one or more species contributed two or more sequences. Our assembly statistics suggest that this approach was successful. Filtered data sets of individual species yielded approximately 10,000 unique gene sequences, which is a slightly less than the 12,669 predicted genes in *H. melpomene* or the 16,866 predicted genes in *Danaus plexippus*. Furthermore, combining data among species yielded 3,961 conserved clusters for which all four species contributed a single sequence.

A second challenge posed by the lack of a reference genome sequence emerges when trying to infer physical dynamics associated with evolutionary genomic phenomena. In particular, the evolutionary processes giving rise to well-resolved gene trees are likely to act on a scale larger than individual genes. For instance, genomic mixing between *P. glaucus* and *P. canadensis* in the formation of *P. appalachiensis* likely involved exchange of large portions of chromosomes, as has been documented in sunflower hybrid species (Rieseberg et al. 1995; Buerkle and Rieseberg 2007). However, without a genome sequence, we cannot test whether similar tree topologies are shared among linked genes. As a workaround, we used the fact that synteny is highly conserved between *Heliconius* and *Bombyx* to do a preliminary analysis, first mapping our conserved clusters back to the *Heliconius* genome and then to *Bombyx*. This approach verified that our conserved clusters really do represent a genome-wide sampling of markers. Subsequently, we tested the hypothesis that genes with the same tree topology were clustered in the genome, at the level of chromosomes. Although the results differed somewhat between the *Heliconius* and *Bombyx* reference, overall they suggested that particular chromosomes are enriched for specific tree topologies.

A third challenge that emerged from our analysis relates to the information content of nucleotide versus peptide alignments, and the unexpected finding that results from these two data sets were not always concordant. For instance, we found that 178 clusters yielded well-supported tree topologies in our analysis of the nucleotide data and 303 clusters in the analysis of peptide data. Surprisingly, there was relatively little overlap in these data sets, with only 68 clusters appearing in both. Although initially concerning, our follow-up analyses revealed that when nucleotide and peptide alignments for the same cluster both yielded statistical support for a topology, it was always the same topology. The real inconsistency then, was in the fact that a given cluster generally would only yield significant support for a given topology based on one of the two alignments. This issue is perhaps not surprising given the recent origin of all three ingroup species and the large amount of genetic variation shared across the group. Furthermore, despite the low overlap, the results of the nucleotide and peptide analyses were largely concordant, yielding approximately equal proportions of clusters for each of the three tree topologies. This suggests that the underlying biological processes giving rise to distinct tree topologies are being captured by both data sets.

Hybrid Speciation and Genomic Mosaicism

Our hypothesis of hybrid speciation with widespread genomic mosaicism in *P. appalachiensis* makes a clear prediction about gene tree topologies; we expect a substantial number of gene trees to support both ((A,C),G,P) and ((A,G),C,P) topologies. Consistent with this hypothesis, we found that Z-linked genes generally supported an ((A,C),G,P) topology while mitochondrial genes generally supported an ((A,G),C,P) topology. This pattern, which is consistent with prior results (Kunte et al. 2011), may help to explain the evolutionary origin of *P. appalachiensis* and its long-term maintenance as a separate species. A rich history of work in this system has revealed that female mimicry phenotype in *P. glaucus* is controlled primarily by a W-linked locus (Hagen and Scriber 1989; Scriber et al. 1996) and many of the thermal adaptations that differ between *P. glaucus* and *P. canadensis* are Z-linked (Hagen and Scriber 1995, 1989). The unique phenotype of *P. appalachiensis* combines the female mimetic polymorphism of *P. glaucus* with the cold-adapted traits of *P. canadensis*. Although we have not specifically traced relationships for W-linked markers (none have been identified in these species), our analysis of mtDNA suggests that the maternally inherited mitochondrion and W chromosome of *P. appalachiensis* are derived from *P. glaucus* while the Z chromosome is derived from *P. canadensis*. This is exactly the scenario predicted based on the mixed phenotype of *P. appalachiensis*.

We followed up on this analysis by examining tree topologies for the remaining, presumably autosomal, clusters. We found that approximately 38% of the conserved clusters that yielded a well-supported topology favored ((A,C),G,P) and approximately 31% favored ((A,G),C,P). However, we

also found that almost 30% of the tree topologies supported the third topology, ((C,G),A,P), linking the putative parental species, with *P. appalachiensis* as their shared sister group. There are at least two potential explanations for widespread sharing of genetic variation between *P. canadensis* and *P. glaucus*. First, these two species have a well-characterized hybrid zone where their distributions meet near the border between Canada and the United States. It is very likely that there is substantial gene flow between these two species across this hybrid zone and that may contribute to the evidence of shared genetic variation we detected. To test this possibility, we calculated Patterson's *D*-statistic (Green et al. 2010; Durand et al. 2011), a measure of shared genetic variation, and found that evidence for introgression between *P. glaucus* and *P. canadensis* is low relative to introgression into *P. appalachiensis*. Therefore, contemporary gene flow between *P. glaucus* and *P. canadensis* may not explain the roughly equal number of ((C,G),A,P) trees, compared with ((A,C),G,P) and ((A,G),C,P) trees. A second possible explanation is that the signature of shared variation between *P. glaucus* and *P. canadensis* derives from ancestral variation that predates the three species radiation. This scenario is very plausible given that *P. glaucus* and *P. canadensis* diverged only ≈600,000 years ago (Kunte et al. 2011) and both species probably have a much larger population size than *P. appalachiensis*.

Interestingly, although the number of clusters that support each topology is very similar, there may be a small excess of clusters linking *P. appalachiensis* to *P. canadensis*. This suggests that *P. canadensis* may have contributed slightly more to the *P. appalachiensis* genome than did *P. glaucus*. This scenario is consistent with the fact that prior to being described as a separate species, *P. appalachiensis* was often referred to as "giant *canadensis*" (Pavulaan and Wright 2002).

Adaptive Evolution

In addition to tracing the evolutionary history of genome-wide markers, we also used our data to perform a broad survey of adaptive protein evolution. Although this analysis is preliminary, some interesting patterns emerge from our initial lists of candidate genes. For instance, genes that showed evidence of adaptive evolution between ingroup and outgroup species were enriched for biological functions that we might expect to be important in this group, such as pigmentation, hormonal sensitivity, and developmental processes. Furthermore, candidates for adaptive evolution among the ingroup taxa point to characteristics such as cuticle formation, which is likely to play a role in thermal adaptation (Futahashi et al. 2008). Much work remains to be done but this data set provides a first-pass list of potential targets for future functional study, and moreover, it provides an initial survey of loci that may have played an important role in the tiger swallowtail radiation.

Supplementary Material

Supplementary table S1 and figure S1 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Matthew Aardema, Harry Pavulaan, and David Wright for providing butterfly pupae for this analysis. They thank Nicholas Crawford and Sean Mullen for advice regarding phylogenetic methods and thank the reviewers for comments on the manuscript. This work was supported by the National Science Foundation grant DEB-1316037 to M.R.K.

Literature Cited

Abbott RJ, Rieseberg LH. 2012. Hybrid speciation. Encyclopaedia of Life Sciences (eLS). Chichester (UK): John Wiley & Sons.

Andolfatto P. 2008. Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. Genetics 180:1767–1771.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol. 18:1585–1592.

Brower JVZ. 1958. Experimental studies of mimicry in some North American butterflies: Part II. *Battus philenor* and *Papilio troilus*, *P. polyxenes* and *P. glaucus*. Evolution 12:123–136.

Buerkle CA, Rieseberg LH. 2007. The rate of genome stabilization in homoploid hybrid species. Evolution 62:266–275.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural approaches to sequence evolution: molecules, networks, populations. New York: Springer Verlag. p. 207–232.

Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol. 28: 2239–2252.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Felsenstein J. 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics 5:164–166.

Futahashi R, et al. 2008. Genome-wide identification of cuticular protein genes in the silkworm, *Bombyx mori*. Insect Biochem Mol Biol. 38: 1138–1146.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29: 644–652.

Grant PR. 1999. Ecology and evolution of Darwin's finches. Princeton (NJ): Princeton University Press.

Green RE, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Hagen RH, Lederhouse RC, Bossart JL, Scriber JM. 1991. *Papilio canadensis* and *P. glaucus* (Papilionidae) are distinct species. J Lep Soc. 45: 245–258.

Hagen RH, Scriber J. 1991. Systematics of the *Papilio glaucus* and *P. troilus* species groups (Lepidoptera: Papilionidae): inferences from allozymes. Ann Entomol Soc Am. 84:380–395.
Hagen RH, Scriber JM. 1989. Sex-linked diapause, color, and allozyme loci in *Papilio glaucus*: linkage analysis and significance in a hybrid zone. J Heredity. 80:179–185.
Hagen RH, Scriber JM. 1995. Sex chromosomes and speciation in tiger swallowtails. In: Scriber JM, Tsubaki Y, Lederhouse RC, editors. Swallowtail butterflies: their ecology and evolutionary biology. Gainesville (FL): Scientific Publishers, Inc. p. 211–227.
Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487: 94–98.
Huang DW, Sherman BT, Lempicki RA. 2008. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37:1–13.
Hulsen T, Huynen MA, de Vlieg J, Groenen PM. 2006. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 7:R31.
Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070–3071.
Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12: 656–664.
Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. PLoS Genet. 4:e1000304.
Kunte K, et al. 2011. Sex chromosome mosaicism and hybrid speciation among tiger swallowtail butterflies. PLoS Genet. 7:e1002274.
Lederhouse RC, Ayres MP, Scriber JM, Tsubaki Y. 1995. Physiological and behavioral adaptations to variable thermal environments in North American swallowtail butterflies. In: Scriber JM, Tsubaki Y, Lederhouse RC, editors. Swallowtail butterflies: their ecology and evolutionary biology. Gainesville (FL): Scientific Publishers, Inc. p. 71–81.
Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 36:96–99.
Luebke HJ, Scriber JM, Yandell BS. 1988. Use of multivariate discriminant analysis of male wing morphometrics to delineate a hybrid zone for *Papilio glaucus glaucus* and *P. g. canadensis* in Wisconsin. Am Midl Nat. 119:366–379.
Mallet J. 2007. Hybrid speciation. Nature 446:279–283.
Mallet J. 2009. Rapid speciation, hybridization and adaptive radiation in the *Heliconius melpomene* group. In: Butlin R, Bridle J, Schluter D, editors. Speciation and patterns of diversity. Sheffield (UK): Cambridge University Press. p. 177–194.
Mavárez J, Linares M. 2008. Homoploid hybrid speciation in animals. Mol Ecol. 17:4181–4185.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652–654.
O'Neil ST, et al. 2010. Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. BMC Genomics 11:310.
Pan M, et al. 2008. Characterization of mitochondrial genome of Chinese wild mulberry silkworm, *Bombyx mandarina* (Lepidoptera: Bombycidae). Sci China C Life Sci. 51:693–701.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.
Pavulaan H, Wright DM. 2002. *Pterourus appalachiensis* (Papilionidae: Papilioninae), a new swallowtail butterfly from the Appalachian region of the United States. Taxon Rep Internal Lepidoptera Survey. 3:1–20.
Pavulaan H, Wright DM. 2004. Discovery of a black female form of *Pterourus appalachiensis* (Papilionidae: Papilioninae) and additional observations of the species in West Virginia. Taxon Rep Internal Lepidoptera Survey. 6:1–10.
Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. Nat Methods. 6:S22–S32.
Rieseberg LH. 1997. Hybrid origins of plant species. Annu Rev Ecol Syst. 28: 359–389.
Rieseberg LH, Van Fossen C, Desrochers AM. 1995. Hybrid speciation accompanied by genomic reorganization in wild sunflowers. Nature 375:313–316.
Ritland DB, Scriber JM. 1985. Larval developmental rates of three putative subspecies of tiger swallowtail butterflies, *Papilio glaucus*, and their hybrids in relation to temperature. Oecologia 65:185–193.
Scriber JM. 1996. A new 'Cold Pocket' hypothesis to explain local host preference shifts in *Papilio canadensis*. Entomol Exp Appl. 80: 315–319.
Scriber JM, Hagen RH, Lederhouse RC. 1996. Genetics of mimicry in the tiger swallowtail butterflies, *Papilio glaucus* and *P. canadensis* (Lepidoptera: Papilionidae). Evolution 50:222–236.
Scriber JM, Hainze J. 1987. Geographic variation in host utilization and the development of insect outbreaks. In: Barbosa P, Schultz JC, editors. Insect outbreaks: ecological and evolutionary processes. New York: Academic Press. p. 433–468.
Scriber JM, Lederhouse RC, Dowell RV. 1995. Hybridization studies with North American swallowtails. In: Scriber JM, Tsubaki Y, Lederhouse RC, editors. Swallowtail butterflies: their ecology and evolutionary biology. Gainesville (FL): Scientific Publishers, Inc. p. 269–281.
Scriber JM, Ording GJ. 2005. Ecological speciation without host plant specialization; possible origins of a recently described cryptic *Papilio* species. Ent Exp Appl. 115:247–263.
Seehausen O. 2006. African cichlid fish: a model system in adaptive radiation research. Proc Biol Sci. 273:1987–1998.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.
Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. Annu Rev Plant Biol. 60:561–588.
Sperling FAH. 1993. Mitochondrial DNA variation and Haldane's rule in the *Papilio glaucus* and *P. troilus* species group. Heredity 71:227–227.
Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19:2325–2327.
Tibshirani R, Leisch F. 2012. Bootstrap: functions for the book "An introduction to the bootstrap" by B. Efron and R. Tibshirani, 1993. New York: Chapman & Hall.
Winter CB, Porter AH. 2010. AFLP linkage map of hybridizing swallowtail butterflies, *Papilio glaucus* and *Papilio canadensis*. J Hered. 101: 83–90.
Xia Q, et al. 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 306:1937–1940.
Zakharov EV, Caterino MS, Sperling FAH. 2004. Molecular phylogeny, historical biogeography, and divergence time estimates for swallowtail butterflies of the genus *Papilio* (Lepidoptera: Papilionidae). Syst Biol. 53:193–215.

Associate editor: George Zhang