Please cite this article in press as: Zhan et al., The Monarch Butterfly Genome Yields Insights into Long-Distance Migration, Cell (2011), doi:10.1016/j.cell.2011.09.052

Cell

The Monarch Butterfly Genome Yields Insights into Long-Distance Migration

Shuai Zhan,[1] Christine Merlin,[1] Jeffrey L. Boore,[2] and Steven M. Reppert[1,*]
[1]Department of Neurobiology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, MA 01605, USA
[2]Genome Project Solutions, 1024 Promenade Street, Hercules, CA 94547, USA
*Correspondence: steven.reppert@umassmed.edu
DOI 10.1016/j.cell.2011.09.052

SUMMARY

We present the draft 273 Mb genome of the migratory monarch butterfly (*Danaus plexippus*) and a set of 16,866 protein-coding genes. Orthology properties suggest that the Lepidoptera are the fastest evolving insect order yet examined. Compared to the silkmoth *Bombyx mori*, the monarch genome shares prominent similarity in orthology content, microsynteny, and protein family sizes. The monarch genome reveals a vertebrate-like opsin whose existence in insects is widespread; a full repertoire of molecular components for the monarch circadian clockwork; all members of the juvenile hormone biosynthetic pathway whose regulation shows unexpected sexual dimorphism; additional molecular signatures of oriented flight behavior; microRNAs that are differentially expressed between summer and migratory butterflies; monarch-specific expansions of chemoreceptors potentially important for long-distance migration; and a variant of the sodium/potassium pump that underlies a valuable chemical defense mechanism. The monarch genome enhances our ability to better understand the genetic and molecular basis of long-distance migration.

INTRODUCTION

Each fall, millions of eastern North American monarch butterflies undergo a long-distance migration, traveling up to 4,000 km to reach their overwintering grounds in central Mexico (Brower, 1995; Reppert et al., 2010; Urquhart and Urquhart, 1978) (Figure 1A). Migratory monarchs are in reproductive diapause. Migrants also have a striking increase in longevity, increased abdominal fat stores and cold tolerance, and an overpowering drive to fly south. Diapause persists at the overwintering sites until spring, when the migrants reproduce and then fly northward to oviposit fertile eggs on newly emerged milkweed plants in the southern United States (Figure 1B). Monarchs are milkweed specialists (Figure 1C), and their evolved chemical defense mechanism has contributed to the monarch's widely known involvement in a mimicry complex with the viceroy butterfly (*Limenitis archippus*) (Ritland and Brower, 1991).

A major compass system that monarchs use for directional information for the migration is a time-compensated sun compass (Froy et al., 2003; Mouritsen and Frost, 2002; Perez et al., 1997). Sun compass components involve the eye's sensing of skylight cues for direction and the brain integration of skylight-stimulated neural responses in the central complex, the presumed site of the sun compass (Heinze and Reppert, 2011). Sun compass output in brain is time compensated by the circadian clock that allows flight direction to be constantly adjusted to maintain a southerly flight direction.

The genome of the commercial silkmoth *Bombyx mori* represents a publicly available lepidopteran genome (ISGC, 2008). Because moths are usually olfactory-centric and butterflies vision-centric, due in part to their respective nocturnal and diurnal behaviors, comparison of the genes involved in these sensory modalities may be informative.

Here, we present the draft 273 Mb genome of the migratory monarch butterfly, including its assembly, a set of 16,866 protein-coding genes, and evolutionary analyses. We focus our gene annotation on gene families likely involved in major aspects of the seasonal migration. The biological interpretation of the monarch genome advances our understanding of the genes and regulatory elements important for the remarkable fall migration.

RESULTS AND DISCUSSION

Genome Assembly and Gene Content

We used a whole-genome shotgun approach with next-generation sequencing platforms to generate the draft genome of the monarch butterfly (Table 1 and Table S1 available online). The combined assembly of 14.7 Gb pairs of Illumina reads (equal to 53.3× coverage of the whole genome) and 6.2 Gb Roche 454 reads (22.3×) resulted in 273 megabases (Mb) of genomic sequence (combined total coverage of 74.7×) (Table S1A). This was termed the v1 assembly and was used for all subsequent analyses (Table S1B). Assessment of the completeness and quality of the assembly v1 is described in the Experimental Procedures.

Please cite this article in press as: Zhan et al., The Monarch Butterfly Genome Yields Insights into Long-Distance Migration, Cell (2011), doi:10.1016/j.cell.2011.09.052

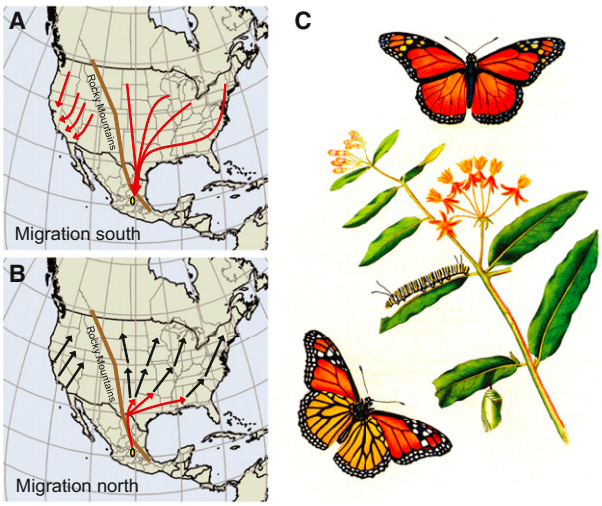

Figure 1. Natural History of the Monarch Butterfly

(A) Migration south. The eastern North American monarch butterfly undergoes a long-distance fall migration to a restricted site in central Mexico (yellow oval). The population of monarchs west of the Rocky Mountains undergoes a truncated fall migration. (Red arrows) Flight paths. From Reppert et al., 2010.

(B) Journey north. Eastern migrants remain at the overwintering areas until spring, when the same butterflies reproduce and migrate northward to lay fertilized eggs on newly emerged milkweed in the southern United States (red arrows). Successive generations of spring and summer monarchs repopulate the home range (black arrows). From Reppert et al., 2010.

(C) Life cycle. Complete metamorphosis from egg to larva (five instars) to pupa (chrysalis) to adult. The male butterfly (upper right) has visible black spots on its hind wings that are missing in females (lower left, underwing view). The larvae feed on milkweed (plants of the genus *Asclepias*). Photograph of engraving from James Edward Smith, *Natural History of the Rarer Lepidopterous Insects of Georgia; from the Observations of John Abbot*, 1797.

See also Figure S4 and Table S9.

Table 1. Features of Assembled Genome and Gene Set

	Monarch (v1)	*Bombyx*
Genome size (Mb)	273	432[c]
Number of chromosomes	29–30[a]	28[c]
Quality Control (covered by assembly)		
ESTs (%)	96	98[c]
CEGMA genes (%)	98.5	98.7
CRP genes (%)	100	100
Genomic Features		
Allele frequency (%)	0.55	ND[c]
Repeat (%)	13.1	43.6
G+C (%)	31.6	37.7
CpG (O/E)	1.13	1.14
Coding (%)	7.51	4.14
Intron (%)	22.8	16.3
Number of miRNA	116[b]	ND[c]
Number of tRNA	431	441[c]
Gene Repertoire		
Number of protein-coding genes	16,866	14,623[c]
with InterPro domains	10,999	9,892
with GO terms	11,210	10,148
Universal orthologs lost	50	194
Species-specific genes	2,511	1,598

See also Figure S1 and Tables S1A–S1J. CEGMA genes, core eukaryotic genes (Parra et al., 2007); CRP genes, cytoplasmic ribosomal protein genes; CpG[O/E], ratio of observed-to-expected CpG; GO, gene ontology.

[a] From Brown et al. (2004).

[b] Only adult miRNAs were identified.

[c] Defined independently or not determined in the latest version of *Bombyx* genome (ISGC, 2008). See also Tables S1A–S1G for details.

In comparison with the 432 Mb *Bombyx* genome (ISGC, 2008), the monarch genome had much less repeat content (13.1% of the whole-genome versus 43.6% in *Bombyx*; Table S1E) and GC content (31.6% in the monarch versus 37.7% in *Bombyx*; Table S1F; Table 1). Like *Bombyx* and the beetle *Tribolium castaneum*, but unlike the honeybee *Apis mellifera*, GC content distribution was uniform but showed a bias of occurrence in coding regions (Figures S1A and S1B). In addition, the distribution plots of CpG ratios (observed/expected CpG dinucleotide density) of the monarch, *Bombyx*, and *Tribolium* genomes clustered together, leaving *Drosophila* and *A. mellifera* with two different patterns (Figures S1C and S1D). These features are consistent with the fact that the monarch, *Bombyx*, and *Tribolium* each encode two types of DNA methyltransferases (Dnmt1 and Dnmt2), whereas *Drosophila* only has Dnmt2 and *A. mellifera* has Dnmt1-3 (Figure S1D). The monarch may thus have a *Bombyx*-like epigenetic system, with a predicted low methylation level (Xiang et al., 2010).

We estimated 16,866 protein-coding genes (Table 1) by combining both homology-based and *ab initio* methods (Table S1G), along with ~228× coverage of the monarch transcriptome. The gene model accounted for the full complement of conserved cytoplasmic ribosomal proteins genes (Marygold et al., 2007), with only one pseudogene and two incomplete predictions (Table S1H), and matched 89.1% of 5,415 manually annotated exons (Table S1G). Nearly 85% of the predicted genes detected homology in the public databases (Table S1I). Moreover, more than 93% of the monarch genes were supported by our transcriptome sequence. These attributes show that the gene models were predicted with accuracy and completeness (Table 1).

Lepidopteran Orthology and Evolution

To understand the lepidopteran proteomes of the monarch and *Bombyx* in the context of other insect species, we compared the reported gene sets of twelve insects and two mammalian species (Figure 2A). Orthology was then assigned according to the predicted evolutionary relationships. The monarch gene set contained 3,138 (18.6%) single-copy genes and 2,514 (14.9%) many-to-many universal genes, compared to 20.4% and 15.9%, respectively, for *Bombyx*. We found significant coverage bias of the transcriptome sequence for the monarch universal orthologs (Table S1J), indicating that they constitute a core set of proteins with conserved functions. Transcriptome coverage also showed a higher distribution of many-to-many universal orthologs than single-copy genes, indicating that the

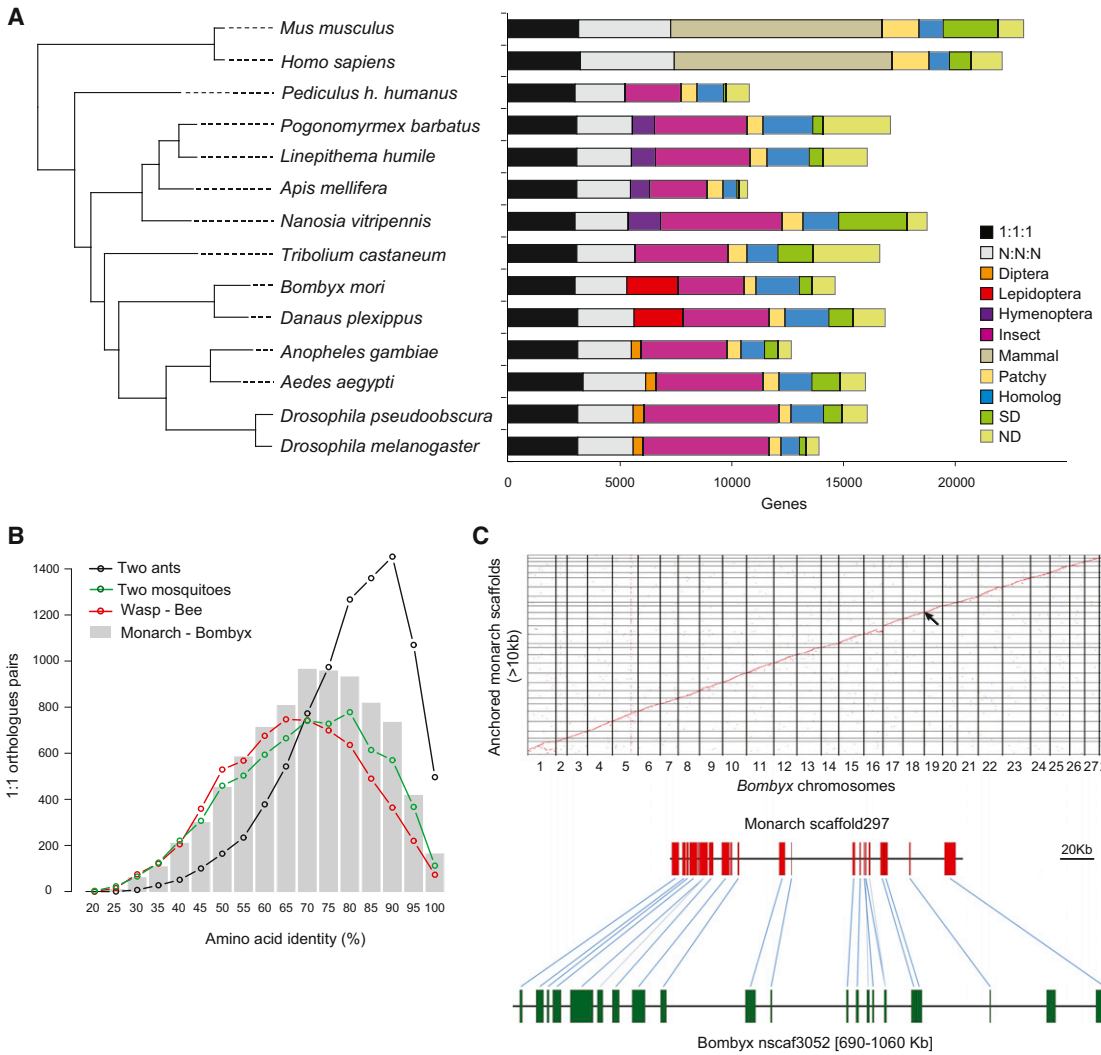

Figure 2. Lepidopteran Orthology and Evolution

(A) Orthology assignment of twelve insect and two mammal genomes. Bars are subdivided to represent different types of orthology relationships. 1:1:1 indicates universal single-copy genes, but absence and/or duplication in a single genome is tolerated. N:N:N indicates other universal genes, but absence in a single genome or two genomes within the different orders is tolerated. Diptera indicates dipteran-specific genes and presence in at least one mosquito and one fly genome. Lepidoptera indicates lepidopteran-specific genes and presence in both the monarch and *Bombyx* genomes. Hymenoptera indicates hymenopteran-specific genes and presence in at least one bee or wasp genome and one ant genome. Insect indicates all other insect-specific orthologs. Mammal indicated mammalian-specific orthologs. Patchy indicates orthologs that are present in at least one insect and one mammal genome. Homolog indicates partial homology detected with $E < 10^{-5}$ but no orthology grouped. SD, species-specific duplicated genes; ND, species-specific genes. The phylogeny on the left was calculated using maximum likelihood analyses of a concatenated alignment of 1,642 single-copy proteins from the 1:1:1 subgroup. The tree was rooted using mammals as outgroup. Bootstrap values based on 1,000 replicates are equal to 1,000 for each node.

(B) The distribution of pairwise amino acid identity. Histogram shows the distribution of sequence identity of 8,221 1:1 orthologs between the monarch and *Bombyx* (diverged ~65 million years ago; Grimaldi and Engel, 2005). To highlight the similar level of molecular divergence, 8,897 orthologs between two ants (*Linepithema humile* and *Pogonomyrmex barbatus*, which diverged ~100–150 million years ago; Moreau et al., 2006), 6,875 orthologs between two mosquitoes (*Anopheles gambiae* and *Aedes aegypti*, which diverged ~150 million years ago; Krzywinski et al., 2006), and 6,520 orthologs between bee and wasp (*Apis mellifera* and *Nasonia vitripennis*, which diverged ~180 million years ago; Werren et al., 2010) were plotted in black, green, and red, respectively.

(C) Microsynteny between monarch and *Bombyx* genomes. Alignment of monarch scaffolds and silkworm chromosomes is shown by pairwise dot plots based on gene homology. 1,802 monarch scaffolds (>10 kb) were anchored to the corresponding position based on the consensus order of gene homology. The arrow denotes the position of a scaffold between the monarch and *Bombyx* that showed particularly strong microsynteny, which is magnified below.

See also Figure S2.

universal orthologs with duplication contributed greatly to basic biological processes compared to the contribution from duplication of recently evolved insect genes.

To address lepidopteran-specific evolution, we identified 1,962 lepidopteran-specific orthologs, which was about twice the number of hymenopteran-specific orthologs and five times

Please cite this article in press as: Zhan et al., The Monarch Butterfly Genome Yields Insights into Long-Distance Migration, Cell (2011), doi:10.1016/j.cell.2011.09.052

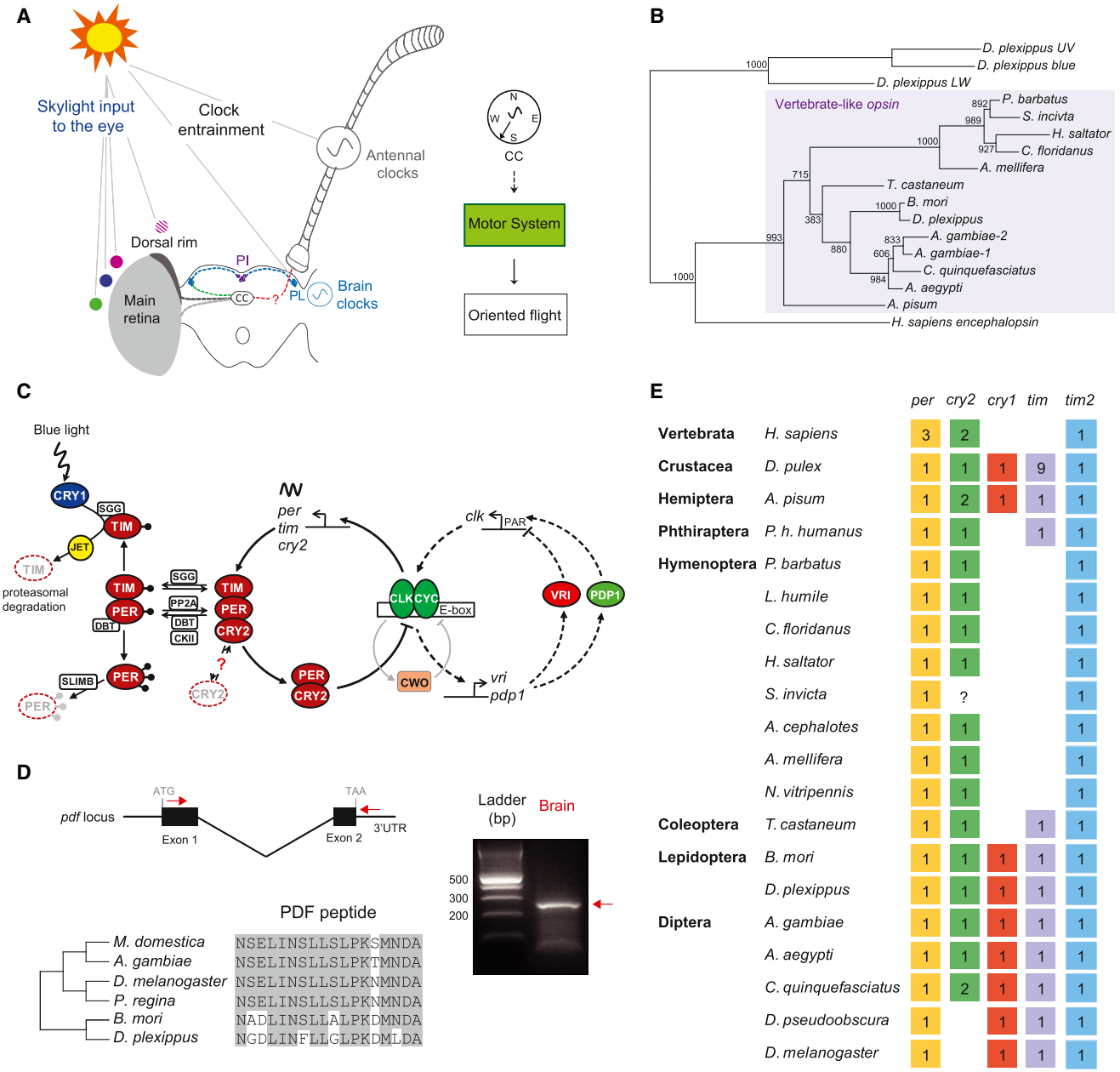

Figure 3. Sun Compass Components Focusing on the Circadian Clock

(A) Model delineating the components used for sun compass navigation. The compass mechanism involves the monarch eye sensing of skylight cues, including color gradient or the sun itself (violet, blue, and green circles) and the polarization pattern of ultraviolet (UV) light (violet circle crosshatched), and the brain integration of skylight cue-stimulated neural response in the central complex (CC; gray dashed lines). In addition, time compensation is provided by circadian clocks located in the antenna. The integrated time-compensated sun compass information is relayed to the motor system to induce oriented flight. The brain circadian clocks are located in the pars lateralis (PL) and communicate with the pars intercerebralis (PI). The PL may also communicate with the central complex. Modified from Reppert et al., 2010.

(B) Maximum likelihood phylogenetic tree of insect vertebrate-like opsins (pteropsins). The tree was rooted using the monarch UV, blue, and long-wavelength opsins.

(C) Schematic of the proposed clockwork mechanism in the monarch butterfly, including the core transcriptional/translational feedback loop (thick arrows) and the modulatory feedback loop (dashed arrows), both incorporating monarch orthologs of all described *Drosophila* clock genes (Dubruille and Emery, 2008). CLOCK (CLK) and CYCLE (CYC) heterodimers drive the transcription of *period* (*per*), *timeless* (*tim*), and type-2 *cryptochrome* (*cry2*), which upon translation form complexes and 24 hr later cycle back into the nucleus, where CRY2 inhibits CLK:CYC-mediated transcription. Light entrainment is mediated by type-1 cryptochrome (CRY1), which promotes TIM degradation. Casein kinase II (CKII), doubletime (DBT), and the protein phosphatase 2A (PP2A) are involved in the posttranslational modifications of PER and TIM, and supernumerary limbs (SLIMB) and jetlag (JET) signal their degradation. The gene(s) involved in CRY2 degradation are unknown (red question mark). The modulatory feedback loop regulates the expression of *clock* through VRILLE (VRI) and PDP1. Monarch *vri* has

the number of dipteran-specific orthologs. In addition, the lepidopteran lineage lacked 223 orthologs that exist widely in other insects and in mammals. In comparison, there were 167 and 103 orthologs missing in the Diptera and Hymenoptera, respectively, suggesting that the Lepidoptera are more derived than the other insect orders.

The Lepidoptera have rapidly evolved. The monarch and *Bombyx* shared 70.8% average amino acid identity between 8,221 1:1 orthologs (Figure 2B), comparable to two mosquitoes (69.4% for 6,875 *Anopheles gambiae*/*Aedes aegypti* orthologs) or bee-wasp identity (67.2% for 6,520 *A. mellifera*/*Nasonia vitripennis* orthologs) but significantly lower than such comparison between two ants (82.2% for 8,897 orthologs between *Linepithema humile* and *Pogonomyrmex barbatus*). Because all three of these genome pairs diverged at least 100 million years ago (Krzywinski et al., 2006; Moreau et al., 2006; Werren et al., 2010) and the monarch radiated from *Bombyx* ~65 million years ago (Grimaldi and Engel, 2005), the Lepidoptera appear to be the fastest evolving insect order sequenced to date.

The monarch and *Bombyx* genomes exhibited a surprisingly high degree of microsynteny (Figure 2C). Approximately 80% of the monarch genes have identifiable *Bombyx* homologs, whereas less than 5% of the coexisting orthologs are duplicated in the monarch or *Bombyx*. According to the consensus gene order shared between the monarch and *Bombyx*, we successfully mapped 1,802 > 10 kb monarch scaffolds spanning 142.7 Mb of the genome to the corresponding *Bombyx* scaffolds. We found strong colinearity in most of the putative chromosomes except for the sex chromosome Z (Chr. 1 in Figure 2C). A total of 8,290 monarch genes (75% of 11,017 genes located in mapped scaffolds with *Bombyx* homology) were found in microsynteny blocks, versus 75% for *A. gambiae* and *A. aegypti* (Zdobnov and Bork, 2007) and 63% for *A. mellifera* and *N. vitripennis* single-copy orthologs (Werren et al., 2010). Although we cannot exclude large-scale chromosomal rearrangements because of the lack of a monarch linkage map, the existent extensive microsynteny reveals that most regions of conserved gene neighborhood were retained after divergence.

Comparison of protein family sizes also showed prominent similarities between the monarch and *Bombyx*, from the global view of proteome domain content (Figure S2). Only 17 InterPro (IPR)-defined families had significant size differences between the two Lepidoptera (Figure S2A), most of which were related to proteinase activity (Figure S2B). Interesting IPR expansions included insect pheromone-binding proteins in the monarch and insulin-like peptides in *Bombyx* (Figure S2B). The major contribution to the lepidopteran phenotypic changes was also apparent by comparing the IPR family size of the Lepidoptera with *Drosophila* or *Tribolium* (Figure S2C). Overall, most IPR families that exhibited variation have general functions involved in transcriptional regulation, protein interactions, and cell-cell communication.

We also compared the evolutionary rate between the monarch and *Bombyx* based on amino acid substitutions, using *Drosophila* as a common outgroup. The results showed that the monarch shares similar sequence identity with *Bombyx* in 1:1 orthologs (50.4% versus 50.7%). This analysis suggests that the ~5,000 year of human domestication of the silkworm (Xiang, 1995) has not had a strong influence on the overall evolutionary rate of nonselected traits in *Bombyx*.

Sensory Input to the Sun Compass

We began our manual annotation by focusing on genes involved in the formation and function of visual input into the sun compass system (Figure 3A). In migrating monarchs, the horizontal position of the sun (solar azimuth) and the derived polarized skylight pattern provide directional cues for the sun compass (Heinze and Reppert, 2011). The solar azimuth is likely sensed by the main retina, whereas polarized light is sensed by the specialized dorsal rim area, a small region of the compound eye anatomically specialized for sensing the angle of plane-polarized skylight (Labhart and Meyer, 1999; Reppert et al., 2004) (Figure 3A).

In the monarch butterfly genome, we identified orthologs of most genes involved in eye development in *Drosophila* (Table S2A) with some notable differences. For development of the main retina, only two genes were not detected in either the monarch or *Bombyx*. The lens crystalline protein Drosocrystallin, which is restricted to the Diptera, was predictably missing from the two Lepidoptera. The other missing gene was *phyllopod*, which is involved in photoreceptor cell fate commitment. Of the 53 genes examined, seven are duplicated in *Drosophila* and none in either the monarch or *Bombyx*, supporting less genetic complexity in the Lepidoptera eye.

Many of the genes necessary for the formation of the dorsal rim area in *Drosophila* are present in the monarch butterfly genome, including two counterparts of homothorax (both also present in *Bombyx*), a transcription factor that is both necessary and sufficient for the formation of the fly dorsal rim (Table S2A) (Wernet et al., 2003). However, members of the *spalt-related* and the *iroC* gene families that are involved in *Drosophila* dorsal rim formation were not found in either the monarch or *Bombyx*. The one duplication and two gene contractions in the two

five consensus CACGTG E box elements in its promoter, and *pdp1* has five in its first intron, through which CLK and CYC could drive their transcription; each transcription factor also contains PAR DNA-binding domains that could modulate CLK transcription by binding to PAR-like binding sites present in the monarch *clk* promoter. Clockworkorange (CWO) modulates the amplitude of the clock.

(D) *pdf* expression in monarch brain. (Top) The primers used for RT-PCR amplification are shown (red arrows) on a schematic of the *pdf* locus identified from the genome assembly. (Right) Agarose gel showing the *pdf* amplicon migrating at ~250 bp (red arrow). (Bottom) Alignment of monarch PDF peptide sequence with those previously described in insects.

(E) Insights into the evolution of the arthropod circadian clock. The presence/absence of the core clock components *period* (*per*), *type-2 cryptochrome* (*cry2*), *type-1 cryptochrome* (*cry1*), *timeless* (*tim*), and *timeout* (*tim2*) has been assessed in all the published arthropod genomes, including 17 insect species. The presence of a given gene is represented by a colored box, and the numbers represent the number of copies found in the genome. The question mark represents an absence that may be due to a degree of incompleteness in the genome given the presence of this gene in all others ant genomes.

See also Figure S3 and Tables S2, S3, S4, S5, and S7.

Please cite this article in press as: Zhan et al., The Monarch Butterfly Genome Yields Insights into Long-Distance Migration, Cell (2011), doi:10.1016/j.cell.2011.09.052

Lepidoptera genomes suggest a modified pattern of dorsal rim development from that in *Drosophila.*

In contrast to eye development, there were interesting differences in the genes involved in phototransduction between the monarch and *Bombyx* (Table S2B). The majority of the genes involved in phototransduction in *Drosophila* were present in the monarch butterfly genome (Table S2B). However, there were duplications of five genes in the monarch only. Most paralogs exhibited lower expression than their counterpart (Table S2B), suggesting that these duplications are relatively recent. Because these duplications were not found in either *Bombyx* or *Drosophila*, the monarch-specific expansions may be involved in the phototransduction mechanisms for sensing skylight cues.

In addition to the monarch butterfly genes encoding each of the three major opsin subfamilies (ultraviolet, blue, and long-wavelength) previously identified (Sauman et al., 2005), we found a monarch gene encoding a vertebrate-like opsin called peropsin in *A. mellifera* (Velarde et al., 2005) (Figure 3B and Table S2B). Further examination of other insect genomes revealed its presence in *Bombyx*, mosquitoes, *Tribolium*, and several Hymenoptera (but not all) beyond *A. mellifera* (Figure 3B), suggesting that this putative light-detecting system is widespread in insects. Interestingly, a sea urchin ortholog was recently shown to play a role in photosensitive larval swimming vertical migration (Ooka et al., 2010).

Central Processing by the Sun Compass

The central neuronal processing of skylight cues in the monarch occurs in the central complex, the sun compass structure in central brain (Heinze and Reppert, 2011) (Figure 3A). With some exceptions, most of the proteins encoding *Drosophila* genes in which mutations lead to altered structure of the central complex and/or locomotion defects were present in the monarch genome (Table S3). There were no lepidopteran homologs of *tay bridge*, whose loss causes defects in the protocerebral bridge, and *polyhomeotic*, a complex locus encoding two transcription factors that are part of the Polycomb group involved in segment identity. There were lepidopteran expansions in *fused lobes*, which encodes a hexosaminidase involved in N-glycan processing, and *SNF4*, which encodes an AMP-activated protein kinase gamma subunit. Many of the *Drosophila* genes whose mutations cause central complex defects have broad developmental defects. Nonetheless, the identified set of monarch orthologs and paralogs provides a starting point for more extensive analyses of the sun compass complex network and its development.

Several peptides (including tachykinins, allostatins, pyrokinin, and neuropeptide F) and neurotransmitters (e.g., serotonin and GABA) have been identified by immunocytochemistry in the central complex of locusts (Homberg, 2002), grasshoppers (Herbert et al., 2010), and *Drosophila* (Kahsai and Winther, 2011) that are likely to be important for neural function and circuitry. We annotated monarch genes that encode orthologs of the vast majority of neuropeptides, polypeptides (Table S4A), and the enzymes involved in biogenic amine synthesis (Table S4B) that are collectively used for neural signaling. There was good agreement between these neural signaling molecules and their corresponding G protein-coupled receptors (Tables S4C and S4D). Specific antibodies can now be developed to map the molecular substrates for central complex neural signaling.

Circadian Rhythms

Circadian clocks and their output pathways play an essential role in migratory processes (Figure 3A). Circadian clocks located in the antennae provide time compensation for the sun compass system (Merlin et al., 2009). In addition, brain clocks located in the pars lateralis of central brain are likely involved in initiating the migratory generation by sensing decreasing day length in the fall (Goehring and Oberhauser, 2002; Reppert et al., 2010).

In *Drosophila* and mammals, the clock mechanism is comprised of a core negative transcriptional feedback loop, which drives self-sustaining rhythms of essential clock components, and a modulatory, interlocking second feedback loop (Allada and Chung, 2010; Reppert and Weaver, 2002). The monarch genome contained the components of both loops (Figure 3C and Table S5). The monarch core feedback loop possesses all the critical clock genes found in *Drosophila*—*clock* (*clk*), *cycle* (*cyc*), *period* (*per*), *timeless* (*tim*), and type-1 *cryptochrome* (designated *cry1*) —but differs in that it also possesses a type-2 vertebrate-like *cry* (*cry2*), previously shown to encode the main transcriptional repressor in the monarch clock (Zhu et al., 2008b), a function fulfilled by *per* in *Drosophila* (Allada and Chung, 2010), which does not possess *cry2*. We also identified genes encoding orthologs of all of the major proteins involved in posttranslational modifications of the core clock proteins (PER and TIM) (Figure 3C). We further identified the major components of a *Drosophila*-like secondary clock feedback loop in the monarch. This included genes encoding orthologs of VRILLE and PDP1, the major regulators of CLK transcription in *Drosophila* (Cyran et al., 2003), along with the appropriate *cis*- and *trans*-regulatory elements (Figure 3C).

A special focus of our manual annotation was the identification of genes encoding pigment-dispersing factor (PDF), a circadian output signal in *Drosophila* brain essential for clock circuitry and driving locomotor activity rhythms (Helfrich-Förster et al., 2000; Shafer and Taghert, 2009), and its G protein-coupled receptor. Although PDF-like immunostaining has been detected in a many other insects, including silkmoths (Závodská et al., 2003), previous immunocytochemical studies have failed to identify PDF staining in the monarch brain (I. Sauman and S.M.R., unpublished data), likely due to the divergence in the monarch PDF sequence (Figure 3D). Although the *pdf* transcript that encodes the prepropeptide was not present in our transcriptome, we verified that it is expressed in the monarch brain (Figure 3D). Mapping monarch PDF expression and clock circuitry is now feasible.

The discovery of type-2 vertebrate-like CRYs in insects, derived from the discovery of CRY2 in monarchs (Zhu et al., 2005), altered our view of how circadian clocks of non-drosophilid insects work (Yuan et al., 2007). To further our understanding of animal clock evolution, we reinvestigated the existence of type-1 and type-2 CRYs in all arthropods in which a draft genome has been published. Virtually all possess a type-2 CRY (the fire ant, *Solenopsis invicta*, genome did not

reveal any *cry* genes), except all *Drosophila* species, which only possess the light-sensitive type-1 CRY (Figures 3E and S3). This supports the existence of both CRY types at the base of arthropod evolution. In addition, type-1 CRY and TIM appear to have been lost prior to the radiation of the hymenopterans, suggesting that the Hymenoptera have evolved different mechanism(s) for photic entrainment. Perhaps the TIMELESS paralog, TIMEOUT, which has some influence on the light input pathway in *Drosophila*, is the key (Benna et al., 2010), as it is expressed in all available insect genomes (Figure 3E).

Juvenile Hormone Regulatory System

Endocrine regulation is crucial in migratory butterflies to coordinate the multiple physiological processes required for a successful long-distance migration, including reproductive arrest, an increase in life span, and a metabolic change increasing fat stores used for flight. These traits are induced in the migratory monarch by a likely downregulation of the insulin-signaling pathway and demonstrated juvenile hormone (JH) deficiency (Herman, 1975; Herman and Tatar, 2001), as documented in flies (Flatt et al., 2005). In response to environmental factors (i.e., temperature and photoperiod), insulin signaling could be reduced through a decrease in the production and/or secretion of insulin-like peptides and/or in the expression of their associated receptors, which would reduce JH biosynthesis in the corpora cardiaca-corpora allata complex, leading to both reproductive quiescence and aging (Figure 4A) (Reppert et al., 2010). We thus focused on annotating genes involved in this endocrine regulation and comparing their expression profiles between summer and migrant butterflies, based on the previous microarray data of corresponding ESTs (GSE14041 of GEO database).

The vast majority of the genes involved in the insulin signaling pathway described in *Drosophila* were represented in the monarch genome (Table S6A). We identified seven insulin-like peptides (ILPs), a number matching that in *Drosophila* but lower than the 20 genes present in *Bombyx*, which, as mentioned previously, represents an expansion for this gene family. Monarch ILP-1 was expressed in the transcriptome at an ~10-fold higher level than any of the other six (Table S6A). In addition, ILP-1 was the only one found in the brain EST library, and its levels were generally decreased in migrants compared to summer butterflies. This suggests that ILP-1 is the main peptide involved in the monarch insulin-signaling pathway and likely the one ultimately regulating JH biosynthesis. Genes encoding downstream target molecules such as the transcription factor forkhead have been identified in the monarch (Table S6A). As reported in flies (Flatt et al., 2005), a derepression of forkhead by a decrease in insulin-like peptides in the migratory monarch would result in increased longevity by inducing JH deficiency (Figure 4A). We annotated from the monarch 71 of the 81 genes that are involved in longevity in *Drosophila* (Table S6B). An ortholog of one of the longevity genes in flies, *rosy* (which encodes xanthine dehydrogenase), has been shown previously to be upregulated in migrant brains (Zhu et al., 2009); *rosy* loss-of-function mutant flies have decreased life span (Geiger-Thornsberry and Mackay, 2004). The potential involvement of other JH-regulated genes in migrant longevity can now be evaluated more extensively.

We also identified and annotated genes involved in the biosynthesis of JH. The entire repertoire of known enzymes involved in the JH biosynthetic pathway proposed in insects (Bellés et al., 2005) is also represented in the monarch genome (Figure 4B and Table S6C). Transcriptional profiles between summer and migratory monarchs of both sexes with confirmed reproductive status revealed an unexpected sexually dimorphic pattern of JH biosynthesis regulation (Figure 4C); male migrants exhibit an overall downregulation of biosynthesis, whereas female migrants appear to use instead an increased turnover (involving JH esterase and/or the epoxide hydrolases; Figure 4B) and/or a putative regulation by JH-binding proteins (Table S6D) to maintain low JH levels. Even though JH has been previously shown to play a role in the control of sexual dimorphism in fly locomotor activity (Belgacem and Martin, 2002), our findings represent evidence for a sexual dimorphism in the molecular pathway of JH regulation itself, which could be common in insects.

Sun Compass Orientation Genes

Besides JH-regulated genes, what are the genes that show seasonal changes in expression that define the migratory state? To address this question in the monarch, microarray analysis of unique cDNA sequences in a brain EST library was recently performed (Zhu et al., 2009). By treating migrants with a JH analog, it was possible to isolate genes involved in sun compass-oriented flight (not affected by JH status) from those involved in other, JH-dependent aspects of the migration, like reproductive function and longevity. Using this approach, 40 cDNAs were identified whose differential expression in the brain correlated with sun compass-oriented flight behavior in individual migrants, independent of JH activity (Zhu et al., 2009). At the time of publication, only 25 of them could be annotated.

With the monarch genome, we have successfully annotated all 40 cDNAs (Table S7); two ESTs were found to be parts of the same gene leaving 39 orientation genes. The 14 previously unannotated genes included those upregulated in migrants that encode the transcription factors bric a brac-like protein and methoprene-tolerant protein 1, a cGMP subunit, and a monarch-specific protein of unknown function. Downregulated orientation genes in migrants included a β-arrestin and a monarch-specific protein of unknown function. This complete annotation has thus revealed two differentially expressed monarch-specific proteins of unknown function that may be unique to the sun compass orientation mechanism.

Small Noncoding RNAs and Monarch Migration

In concert with protein-coding genes, regulatory elements in the genome could be responsible for the initiation and/or maintenance of the migratory state. The primary gene-silencing regulators are microRNAs (miRNAs), small interfering RNAs, or piwi-interacting RNAs. Of 31 RNAi pathway-related genes found in *Drosophila* and/or *C. elegans*, 21 were annotated in the monarch and 18 in *Bombyx* (Table S8). Our manual annotation suggests that the Lepidoptera may possess the machinery for effective systemic RNAi-mediated gene silencing, in spite of variable success reported between and within lepidopteran species (Terenius et al., 2011).

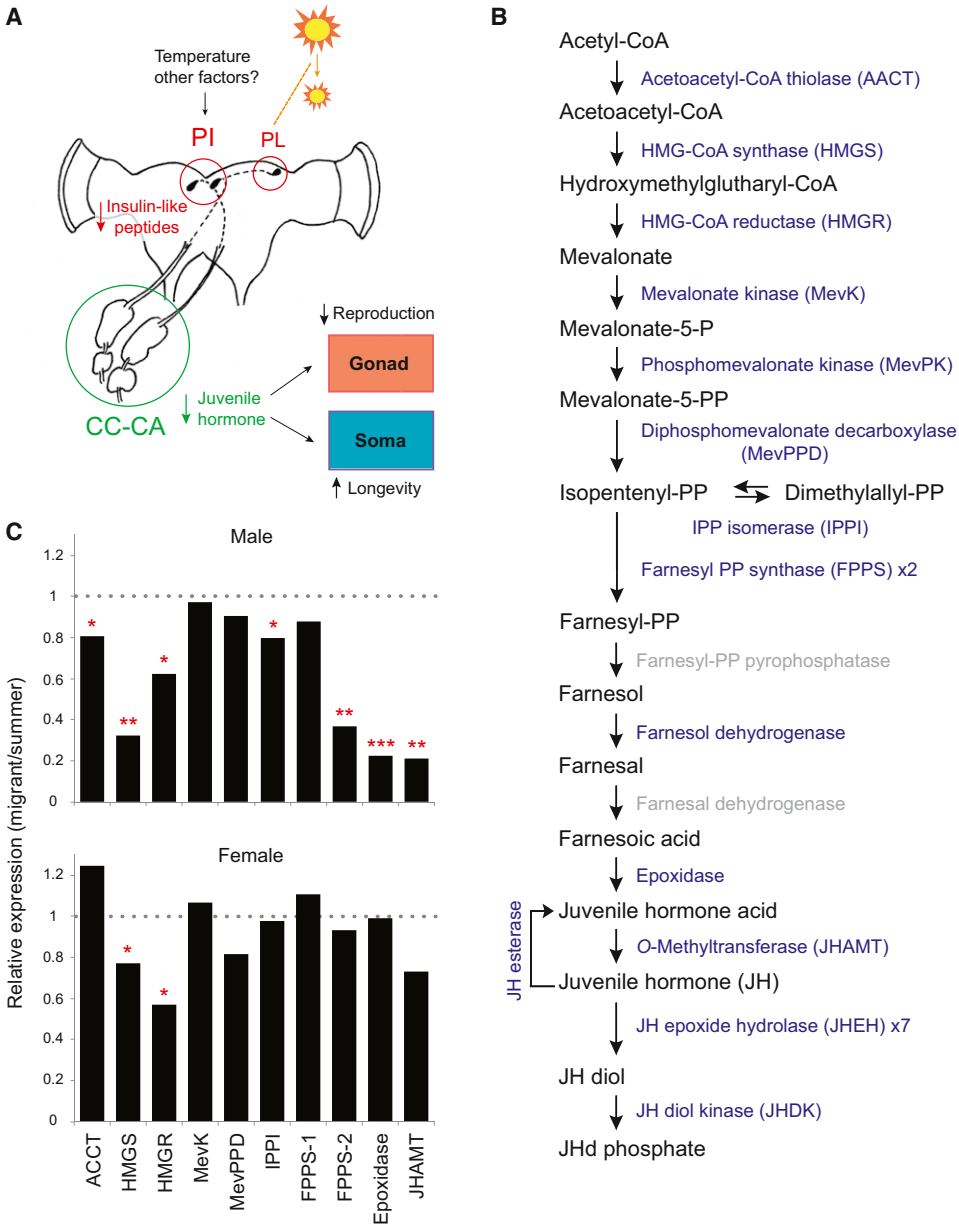

Figure 4. Juvenile Hormone Regulatory Pathway

(A) Proposed endocrine regulation of reproductive arrest and longevity in migratory monarch butterflies. Decreasing daylength (decrease in sun size) is sensed by circadian clocks in the pars lateralis (PL). This information is relayed to the pars intercerebralis (PI) in which the production and/or secretion of insulin-like peptides is decreased, resulting in a decrease of juvenile hormone (JH) biosynthesis in the corpora cardiaca-corpora allata (CC-CA) complex. JH deficiency affects target tissues, leading to reproductive quiescence and increased longevity.

(B) JH biosynthetic and degradation pathways (Bellés et al., 2005). Enzymes in the monarch genome are in blue. Gray denotes enzymes proposed to catalyze Farnesyl-PP to farnesoic acid, but their actual existence has not yet been verified.

(C) Sexually dimorphic pattern of JH biosynthesis. Bars indicate relative gene expression levels of enzymes, as annotated in (B), in migrants relative to summer counterparts (data calculated from GSE14041 of GEOdatabase; n = 5 per sex per enzyme). Significance was estimated by two-sample t tests. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

See also Table S6.

To investigate the potential for gene regulation by endogenous noncoding RNAs in the migratory process, we used Illumina sequencing and computational methods to characterize these regulators in summer and migratory butterflies. miRNAs accounted for the vast majority of the reads from the whole-body small RNA libraries (Figure 5A, inset). We identified 116 miRNAs from monarch, including 66 conserved, 15 lepidopteran-specific, and 35 novel miRNAs (Figure 5A). A total of 55

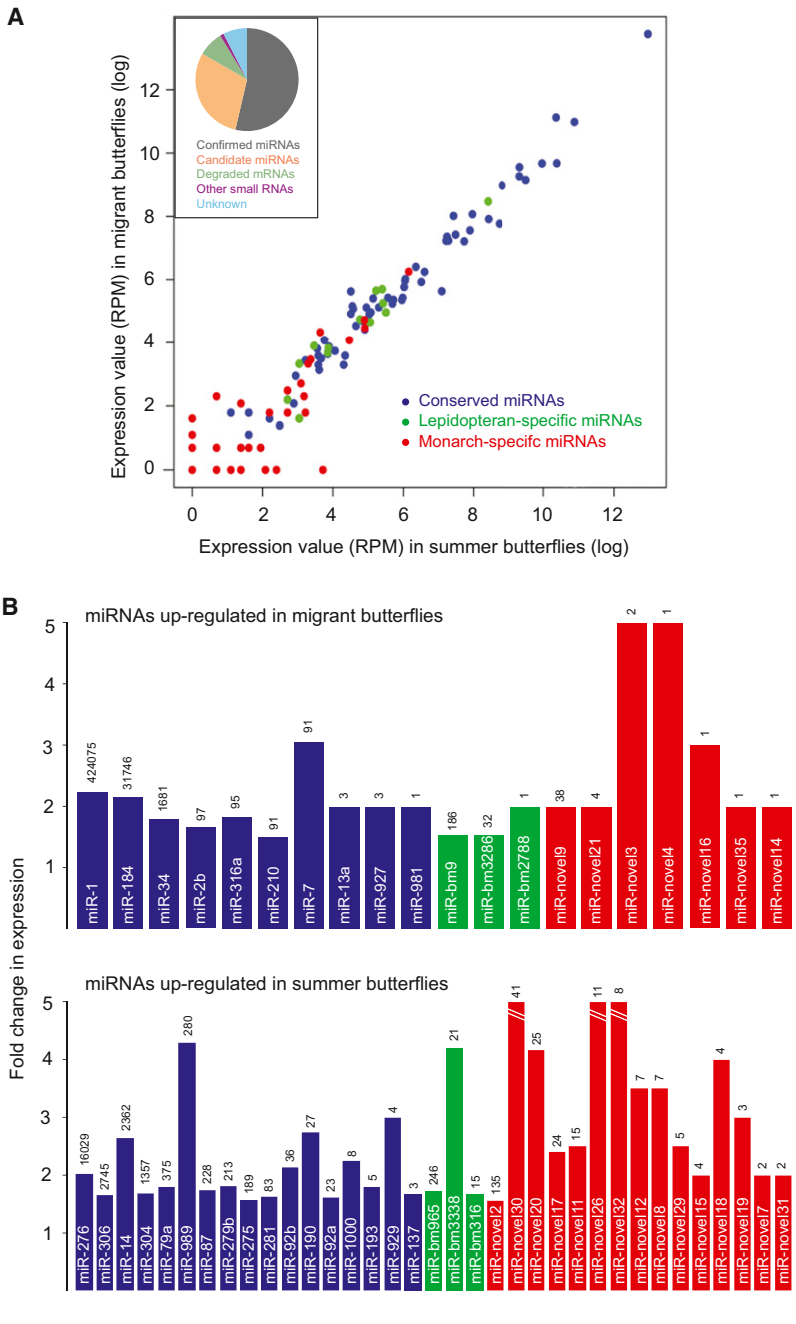

Figure 5. miRNAs in Summer and Migratory Monarchs

(A) Expression of monarch miRNAs. The relationship between the expression values in reads per million (RPM) of summer butterflies and migrants plotted on a natural logarithmic (log) scale. Each of the 116 identified miRNAs is represented as a colored dot (blue, conserved; green, lepidopteran specific; red, monarch specific). Inset is a pie chart showing the classification of noncoding small RNA sequencing reads from the merged summer and migratory samples.

(B) miRNAs expressed differentially between a pool of 10 summer butterflies and a pool of 10 migratory monarchs. (Top) Bars show the mean miRNA levels that were upregulated ≥ 1.5-fold in migrants compared to summer butterflies. The normalized expression value (in RPM) is displayed above each bar. (Bottom) Mean miRNA levels upregulated in summer butterflies compared to migrants. See also Table S8.

miRNAs had ≥ 1.5-fold differences in mean expression levels between summer and migratory monarchs, with 35 upregulated in summer butterflies and 20 upregulated in migrants (Figure 5B). The conserved miRNAs were the most highly expressed in the monarch (Figure 5A), similar to other invertebrates (Ruby et al., 2006, 2007; Wei et al., 2009).

Of the 27 conserved miRNAs whose mean levels were expressed differentially between summer and migratory monarchs (Figure 5B), three were of interest because of their reported functions in other species. miR-1, a muscle-specific miRNA that is upregulated in gregarious locusts (Wei et al., 2009), showed the highest abundance of all the monarch miRNAs, and its mean value was upregulated by 2.2-fold in migrants; miR-1 may be involved in the prolonged muscle-dependent flight required of the monarch during their long-distance migration. In *Drosophila*, miR-7 stabilizes regulatory processes against temperature perturbations (Li et al., 2009); it showed the largest upregulation of mean values of the conserved group in migrants, and it may help with the increased cold tolerance manifested by migrants and necessary for their existence at the overwintering grounds atop the transvolcanic mountains in central Mexico. We also found that miR-14, a regulator of fat metabolism in *Drosophila* (Xu et al., 2003), showed relatively high expression, and its mean value was downregulated in migrants. Because migrant butterflies have increased lipid stores that can be used as fuel during the migration, the metabolic effects of miR-14 could be important for the migration.

Similar to the low coverage of nonuniversal genes (Table S1J), most novel miRNAs showed weak expression (Figure 5A), indicating that they evolved recently and/or that their distribution is restricted. Because these novel miRNAs were confidently predicted by three independent measures, it is possible that some have functions unique to the migratory state of monarchs and require further study.

Chemoreception

Chemoreception is likely critical for a successful fall migration (Reppert et al., 2010). The detection of chemical cues is mediated by multigene families of olfactory receptors (ORs), ionotropic receptors (IRs), and gustatory receptors (Grs) (Figure 6). The molecular underpinnings of lepidopteran chemoreception

Please cite this article in press as: Zhan et al., The Monarch Butterfly Genome Yields Insights into Long-Distance Migration, Cell (2011), doi:10.1016/j.cell.2011.09.052

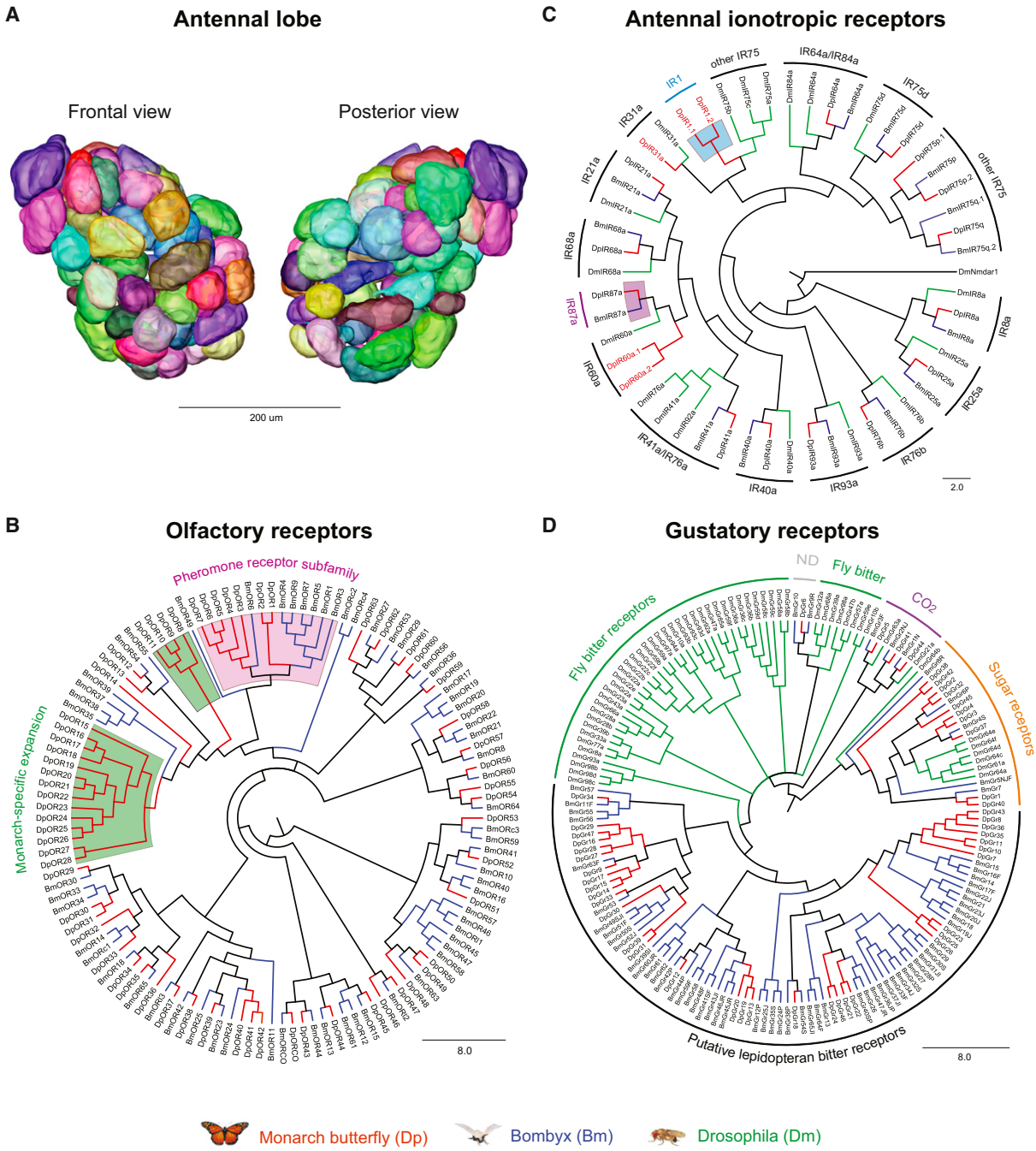

Figure 6. Insights into Chemosensory Function in the Monarch Butterfly

(A) Three-dimensional reconstruction of right antennal lobe (AL) of a female migrant monarch. Each glomerulus is highlighted with a unique color without physiological significance.

(B) Unrooted tree of candidate monarch (*Dp*, red lines) and *Bombyx* (*Bm*, blue lines) olfactory receptors (ORs). (Green boxes) monarch-specific expansions; (purple box) pheromone receptor candidates. *BmOR2* was renamed as *BmORCO*. ORl, OR-like; ORc, OR candidate.

(C) Phylogenetic relationship of the monarch antennal ionotropic receptor (IR) candidates with *Bombyx* and *Drosophila* antennal IRs. (Red) monarch; (blue) *Bombyx*; (green) *Drosophila*. The monarch IR names in red represent genes present in the monarch genome, but not in *Bombyx*.

(D) Phylogenetic relationship of the monarch gustatory receptor (Gr) candidates with *Bombyx* and *Drosophila* Grs. (Red) monarch; (blue) *Bombyx*; (green) *Drosophila*.

have been extensively studied in moths (Wanner and Robertson, 2010) but have received little attention in butterflies.

We identified and manually annotated a repertoire of 64 OR candidates (~60% are full-length) (Figure 6B). The number of ORs identified is comparable to the number of ORs found in *Bombyx* (66; Tanaka et al., 2009) and is in close agreement with the number of glomeruli in the antennal lobe of the migratory female monarch (Figure 6A). Indeed, 68 and 69 glomeruli were counted in each lobe (right one) of two respective specimens (Figure 6A). This supported the 1:1 relationship from the axonal projections of the neurons expressing a given OR to a single glomerulus (Gao et al., 2000). The identified monarch ortholog of the highly conserved *Drosophila DmOr83b* was designated *DpORCO*, based on a unified nomenclature for this coreceptor (Vosshall and Hansson, 2011). Phylogenetic analysis including *Bombyx* (Bm) ORs revealed two monarch-specific subfamily expansions (Figure 6B, green boxes) that may be used for species-specific recognition behaviors such as recognition of overwintering sites, nectaring sources, or milkweed for oviposition.

Interestingly, we also identified seven OR genes that form monarch-specific expansions clustering with the moth pheromone receptor subfamily. Unlike *Bombyx*, which relies on pheromones for sexual communication, butterflies use multisensory modalities, including vision and olfaction. However, the use of pheromone cues during monarch courtship is unclear (Pliske, 1975). We thus hypothesize that these ORs may be involved in social behavior, such as in the roosting behavior that migratory monarchs manifest at night during their migration south and at the overwintering sites (Reppert et al., 2010).

The chemosensory ionotropic receptor (IR) family may also be involved in monarch chemosensory behaviors. We identified and manually annotated 19 antennal IRs that appear to be functional genes (no pseudogenes were detected), compared to 14 in *Bombyx* (Croset et al., 2010; Olivier et al., 2011) (Figure 6C). Phylogenetic analysis revealed that 16 monarch IRs are putative orthologs of conserved antennal IRs (Croset et al., 2010). Two of the three other monarch antennal IR candidates, DpIR1.1 and DpIR1.2, cluster together in a lineage previously thought to be unique to noctuids (Olivier et al., 2011), which now appears instead to be lepidopteran specific (Figure 6C, blue box). Another IR candidate, DpIR87a, might define with its BmIR87a ortholog another subtype of lepidopteran-specific antennal IR, as proposed previously (Olivier et al., 2011) (Figure 6C, purple box).

Gustatory receptors (Grs) mediate contact chemoreception that is used by insects for feeding behavior, host plant selection, and oviposition. *Bombyx* and the monarch present similarities in that their larvae feed exclusively on mulberry and milkweed leaves, respectively. We annotated 47 monarch Gr candidates (Figure 6D), compared to 65 BmGrs (Wanner and Robertson, 2008). Phylogenetic analysis revealed that 14 putative DpGrs are from the three conserved lineages in insects: the DmGr43a protein subfamily of unknown function, the carbon dioxide receptors, and the sugar receptors subfamilies functionally characterized in flies (Dahanukar et al., 2007; Jones et al., 2007). Despite a lower number of Grs identified, the monarch possesses twice as many sugar receptors as found in *Bombyx*, which is consistent with its ecology as a flower nectaring butterfly.

Remarkably, the 33 remaining monarch Gr candidates cluster in our phylogenetic analysis with the 55 BmGrs that form a monophyletic subfamily distinct from those of other insects. This subfamily has been proposed to be putative silkmoth bitter receptors for secondary plant compounds (Wanner and Robertson, 2008) (Figure 6D). Our annotation extends this discovery to butterflies and therefore supports the hypothesis of a specialization in deterrent bitter compounds detection basal to the lepidopteran lineage. *Bombyx* and monarch putative bitter Grs exhibit species-specific small expansions that could reflect different specificity in host plant recognition (mulberry versus milkweed).

Chemical Defense

The P-type Na^+/K^+-ATPase is an essential enzyme that maintains the proper balance of ions on opposite sides of the cell that is critical for normal cellular function (Skou, 1998). As a milkweed specialist, monarch larvae are exposed to cardiac glycosides that are sequestered in adults, making the butterfly bitter and toxic. Although these cardenolide glycosides block the sodium/potassium pump and cause death (Prassas and Diamandis, 2008), the monarch enzyme is completely resistant to inhibition by the cardiac glycoside ouabain (Holzinger et al., 1992). The molecular basis for this was originally proposed to be a point mutation in the α subunit, changing Asn-193 to His (numbering based on the full-length monarch protein; Figure S4), which is a critical residue for ouabain binding. In fact, mutating Asn to His at the homologous position in *Drosophila* converts the fly enzyme to the monarch version highly resistant to ouabain binding (Holzinger and Wink, 1996). In addition, there was no mutation at this residue in the DNA of the α subunit of other *Danaus* species whose larvae also feed on milkweed (Mebs et al., 2000). It is possible that the α subunit variant allows for the higher sequestration of cardenolides that are found in the monarch, compared to the lower sequestration in the nonmigrating Queen butterfly (*D. gilippus*) (Cohen, 1985).

We have now been able to obtain the entire sequence of the coding region of the α subunit (1193 aa) and its genomic structure (Figure S4). We have established an additional amino acid replacement in the coding sequence, which would confer even greater resistance to ouabain binding than the original Asn193His change; no other amino acid changes exist among the conserved regions of the monarch, *Drosophila*, and sheep proteins. We previously identified a Glu182Val change (Zhu et al., 2008a); amino acid substitutions at both 182 and 193 (Figure S4) confer a higher degree of resistance to ouabain binding (Price et al., 1990). This Glu182Val change was missed in previous work (Holzinger and Wink, 1996; Mebs et al., 2000) because only genomic DNA was amplified and the splicing of the intron 3′ to position 182 was incorrectly predicted (Figure S4). We therefore have a full explanation for the difference in what we and others have reported for residue 182. Furthermore, our transcriptome revealed both the Glu182Val and Asn193His changes in all > 1,000× transcriptome coverage (Table S9). We also annotated two α subunit paralogs in the monarch genome (Table S9), but neither had a conserved

ouabain-binding site, and the transcriptome coverage of both was low ($\leq$3).

Because a functional Na^+/K^+-ATPase depends on heterodimerization between α and β subunits, we identified in the monarch all three forms of the β subunit described in *Drosophila* (nervana 1, 2, and 3). There were four homologs of nervana 2, of which two were highly expressed in the transcriptome (Table S9). Moreover, the most highly expressed β subunit was nervana 3, which has been recently shown in *Drosophila* to be exclusively expressed in the nervous system, especially sensory neurons (Baumann et al., 2010).

The unique structure of the major α subunit of Na^+/K^+-ATPase provides a molecular substrate for the ability of the monarch butterfly to sequester toxic cardenolides. This would help protect the monarch against predation during its migration and overwintering period. We also propose that this molecular substrate has allowed the monarch to participate in the well-known mimicry complex with the viceroy butterfly. This mimicry system was first described as Batesian, with the monarch being unpalatable and toxic and the viceroy, first defined as palatable, exploiting the model species through its shared coloration pattern and display behavior to predators (Brower, 1958). This view was modified with the finding that the monarch and viceroy bodies are equally unpalatable to birds, suggesting a Müllerian mimicry, in which both species are comimics (Ritland and Brower, 1991).

Conclusions

We have performed deep sequencing and de novo assembly of the monarch genome to provide, to our knowledge, the first characterized genome of a butterfly and of a long-distance migratory species. Overall, the attributes of the monarch genome and its proteome provide a treasure trove for furthering our understanding of monarch butterfly migration; a solid background for population genetic analyses between migratory and nonmigratory populations; and a basis for future genetic comparison of the genes involved in navigation yet to be discovered in other long-distance migrating species, including vertebrates like migratory birds.

EXPERIMENTAL PROCEDURES

See the Extended Experimental Procedures for detailed protocols.

Genome Sequencing

We used wild-caught, migratory female butterflies (the heterogametic sex) for sequencing; laboratory-generated butterfly lines were not available. Although the use of a single butterfly for all sequencing runs would have been optimal, it was precluded by the need of different libraries for the different sequencing runs from different platforms and vendors. Genomic DNA (34–65 μg) was isolated from individual thoraces using standard protocols with RNase treatment. We employed both Illumina technology and Roche 454 sequencing technology (Table S1A). For deep sequence coverage, we used Illumina sequencing. Short insert paired-end (200 bp; SIPES) and long insert mate-pair (3–5 kb; LIPES) libraries were constructed from the DNA of female F-2 (34 μg DNA yield). Sequencing runs from three lanes of SIPES and three lanes of LIPES were performed by Eureka Genomics (Hercules, CA, USA). To overcome the probable repetitive regions, we also employed Roche 454 sequencing to obtain longer reads. From female F-9 (63 μg DNA), 12 shotgun fragment runs were performed on the 454 FLX/titanium platform (conducted by Virginia Bioinformatics Institute, Blacksburg, VA, USA), as well as three runs from a 20 kb insert paired-end library. From a third female (F-4; 65 μg DNA), we generated DNA for two Roche sequencing runs from an 8 kb insert paired-end library.

Genome Assembly

Initial assemblies were generated by CLC bio's de novo assembler (Katrinebjerg, Denmark) and Newbler (Roche, Inc.) for Illumina and Roche 454 reads, respectively (Table S1B). We then used the Illumina paired-end reads, step by step from 200 bp to 5 kb insert size, to join the initial Illumina contigs into scaffolds by SSPACE 1.0 (Boetzer et al., 2011). Remaining gaps within these scaffolds were iteratively filled with paired-end SIPES reads and the Roche 454 contigs using GapCloser available in SOAPdenovo (Li et al., 2010). The resulting v1 assembly included all scaffolded contigs and had a final scaffold N50 length of 53,032 bp (spanning 272.7 Mb) and contig N50 length of 50,721 bp (spanning 272.2 Mb). A second version of the assembly (v2) provided additional extension of the scaffolds using reads from the Roche 8 kb and 20 kb paired-end libraries, improving the scaffold assembly to a N50 length of 207,025 bp (spanning 277.7 Mb) (Table S1B). In the v2 assembly, 7,780 scaffolds contain 3,598 gaps, spanning 5,393,193 bp. We continue to improve the assembly and will update as appropriate.

We evaluated the completeness of coverage of our assembly using homologs of other insects. The monarch assembly covered 457 core eukaryotic genes (CEGMA) (Parra et al., 2007) (TBLASTN, $E < 10^{-5}$) of *Drosophila* at a level comparable to four other well-organized insect genomes, *Bombyx*, *Tribolium*, *P. barbatus*, *and S. invicta* (Table S1C), even though those genomes have substantially larger scaffold sizes. Moreover, the fraction of bases in the CEGMA genes present in single scaffolds was also very similar among the five species (Table S1D). We also aligned the entire *Drosophila* and *Tribolium* gene sets to the monarch and *Bombyx* assembled genomes (TBLASTN, $E < 10^{-5}$) (Table S1C). Both Lepidoptera showed very similar levels of coverage (using genblastA v1.0.4) and percentage of mapped genes located in a single scaffold. Coverage of above alignments was automatically calculated using genblastA v1.0.4 with "-e 1e-5 -a 0.5 -r 1 -c 0.5" option. In addition, all 79 conserved cytoplasmic ribosomal protein genes were completely present in the v1 assembly (Table S1H), with only one gene that was lacking 30 amino acids. We also assessed the completeness and accuracy of our assembly using 9,484 independently sequenced and assembled monarch ESTs (Zhu et al., 2008a). A total of 9,072 ESTs ($\sim$96%) could be mapped to the assembly (BLASTN, $E < 10^{-50}$), and none of them mapped to more than one scaffold. In terms of accuracy, we found that only 28 exons (0.3% of all mapped ESTs) were located in the opposite orientation with their neighboring exons, as candidates inverted assembly. There were 64,380 single-base mismatches (0.94%) and 5,660 indels (0.08%) found in the 6.85 Mb region mapped. Taking into account the high level of heterozygosity of the monarch genome (0.55%), our assembly exhibits a low level of assembly error. Taken together, the monarch genome assembly appears quite complete and accurate compared to the gene coverage in other genomes.

Genome Annotation

A total of 5.4 Gb transcriptome sequence was generated by Illumina RNA-seq (The National Center for Genome Resources, Santa Fe, NM, USA), representing all stages of monarch development. The official gene set (OGS1.0; Table S1G) was based on a GLEAN consensus model (Elsik et al., 2007), which combined transcriptome, homology, and five *ab-initio* sets (Table S1G). Automatic orthology was determined using the OrthoMCL pipeline (Li et al., 2003). More than 1,000 genes of biological interest were manually annotated. Genomes and gene sets for comparative analysis were listed in Table S1I. To identify and profile miRNAs, Illumina small RNA-seq was performed independently for summer butterflies and migrants samples, each being a pool of 10 animals. We primarily used the miRDeep algorithm (Friedländer et al., 2008) for miRNA prediction.

ACCESSION NUMBERS

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AGBW00000000. The version described in

this paper is the first version, AGBW01000000. Further details are available at the MonarchBase web portal (http://monarchbase.umassmed.edu).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, and nine tables and can be found with this article online at doi:10.1016/j.cell.2011.09.052.

ACKNOWLEDGMENTS

We thank Stanley Heinze and Surainder Asokaraj for providing the three-dimensional reconstruction images of the monarch antennal lobe shown in Figure 6A; Alan Ritacco for use of the computing server; and Susan Fuerstenberg, Amy Casselman, and Lauren Foley for assistance. S.M.R. initiated and oversaw the project. J.L.B. designed the sequencing strategy, provided sequence quality control, and generated the v0 assemblies. S.Z. performed the v1 and v2 assemblies, generated the gene sets, and provided quality control and the data for all figures, except Figure 6A. C.M. performed the genomic DNA preparations, collected and prepared RNA for the transcriptome library, helped with the annotation of the chemosensory receptors, and properly formatted all the figures. S.Z., C.M., J.L.B., and S.M.R. performed data analyses. S.Z., C.M., and S.M.R. wrote the paper. This work was supported by NIH grant GM086794-02S1 and the Higgins Foundation.

Received: July 8, 2011
Revised: August 26, 2011
Accepted: September 6, 2011
Published: November 23, 2011

REFERENCES

Allada, R., and Chung, B.Y. (2010). Circadian organization of behavior and physiology in Drosophila. Annu. Rev. Physiol. *72*, 605–624.

Baumann, O., Salvaterra, P.M., and Takeyasu, K. (2010). Developmental changes in beta-subunit composition of Na,K-ATPase in the Drosophila eye. Cell Tissue Res. *340*, 215–228.

Belgacem, Y.H., and Martin, J.R. (2002). Neuroendocrine control of a sexually dimorphic behavior by a few neurons of the pars intercerebralis in Drosophila. Proc. Natl. Acad. Sci. USA *99*, 15154–15158.

Bellés, X., Martín, D., and Piulachs, M.D. (2005). The mevalonate pathway and the synthesis of juvenile hormone in insects. Annu. Rev. Entomol. *50*, 181–199.

Benna, C., Bonaccorsi, S., Wülbeck, C., Helfrich-Förster, C., Gatti, M., Kyriacou, C.P., Costa, R., and Sandrelli, F. (2010). Drosophila timeless2 is required for chromosome stability and circadian photoreception. Curr. Biol. *20*, 346–352.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.

Brower, J.V.Z. (1958). Experimental studies of mimicry in some North American butterflies. Part 1. The monarch, Danaus plexipus, and the viceroy, Limenitis archippus archippus. Evolution *12*, 32–47.

Brower, L.P. (1995). Understanding and misunderstanding the migration of the monarch butterfly (Nymphalidae) in North America: 1857–1995. J. Lepid. Soc. *49*, 304–385.

Brown, K.S., Jr., Von Schoultz, B., and Suomalainen, E. (2004). Chromosome evolution in Neotropical Danainae and Ithomiinae (Lepidoptera). Hereditas *141*, 216–236.

Cohen, J.A. (1985). Differences and similarities in cardenolide contents of queen and monarch butterflies in florida and their ecological and evolutionary implications. J. Chem. Ecol. *11*, 85–103.

Croset, V., Rytz, R., Cummins, S.F., Budd, A., Brawand, D., Kaessmann, H., Gibson, T.J., and Benton, R. (2010). Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet. *6*, e1001064.

Cyran, S.A., Buchsbaum, A.M., Reddy, K.L., Lin, M.C., Glossop, N.R., Hardin, P.E., Young, M.W., Storti, R.V., and Blau, J. (2003). vrille, Pdp1, and dClock form a second feedback loop in the Drosophila circadian clock. Cell *112*, 329–341.

Dahanukar, A., Lei, Y.T., Kwon, J.Y., and Carlson, J.R. (2007). Two Gr genes underlie sugar reception in Drosophila. Neuron *56*, 503–516.

Dubruille, R., and Emery, P. (2008). A plastic clock: how circadian rhythms respond to environmental cues in Drosophila. Mol. Neurobiol. *38*, 129–145.

Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S., and Weinstock, G.M. (2007). Creating a honey bee consensus gene set. Genome Biol. *8*, R13.

Flatt, T., Tu, M.P., and Tatar, M. (2005). Hormonal pleiotropy and the juvenile hormone regulation of Drosophila development and life history. Bioessays *27*, 999–1010.

Friedländer, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. Nat. Biotechnol. *26*, 407–415.

Froy, O., Gotter, A.L., Casselman, A.L., and Reppert, S.M. (2003). Illuminating the circadian clock in monarch butterfly migration. Science *300*, 1303–1305.

Gao, Q., Yuan, B., and Chess, A. (2000). Convergent projections of Drosophila olfactory neurons to specific glomeruli in the antennal lobe. Nat. Neurosci. *3*, 780–785.

Geiger-Thornsberry, G.L., and Mackay, T.F. (2004). Quantitative trait loci affecting natural variation in Drosophila longevity. Mech. Ageing Dev. *125*, 179–189.

Goehring, L., and Oberhauser, K.S. (2002). Effects of photoperiod, temperature, and host plant age on induction of reproductive diapause and development time in Danaus plexippus. Ecol. Entomol. *27*, 674–685.

Grimaldi, D., and Engel, M.S. (2005). Evolution of the Insects (New York: Cambridge University Press).

Heinze, S., and Reppert, S.M. (2011). Sun compass integration of skylight cues in migratory monarch butterflies. Neuron *69*, 345–358.

Helfrich-Förster, C., Täuber, M., Park, J.H., Mühlig-Versen, M., Schneuwly, S., and Hofbauer, A. (2000). Ectopic expression of the neuropeptide pigment-dispersing factor alters behavioral rhythms in Drosophila melanogaster. J. Neurosci. *20*, 3339–3353.

Herbert, Z., Rauser, S., Williams, L., Kapan, N., Güntner, M., Walch, A., and Boyan, G. (2010). Developmental expression of neuromodulators in the central complex of the grasshopper Schistocerca gregaria. J. Morphol. *271*, 1509–1526.

Herman, W.S. (1975). Endocrine regulation of posteclosion enlargement of the male and female reproductive glands in Monarch butterflies. Gen. Comp. Endocrinol. *26*, 534–540.

Herman, W.S., and Tatar, M. (2001). Juvenile hormone regulation of longevity in the migratory monarch butterfly. Proc. Biol. Sci. *268*, 2509–2514.

Holzinger, F., and Wink, M. (1996). Mediation of cardiac glycoside insensitivity in the monarch butterfly (Danaus plexippus): Role of an amino acid substitution in the ouabain binding site of Na+,K+-ATPase. J. Chem. Ecol. *22*, 1921–1937.

Holzinger, F., Frick, C., and Wink, M. (1992). Molecular basis for the insensitivity of the Monarch (Danaus plexippus) to cardiac glycosides. FEBS Lett. *314*, 477–480.

Homberg, U. (2002). Neurotransmitters and neuropeptides in the brain of the locust. Microsc. Res. Tech. *56*, 189–209.

ISGC (International Silkworm Genome Consortium). (2008). The genome of a lepidopteran model insect, the silkworm Bombyx mori. Insect Biochem. Mol. Biol. *38*, 1036–1045.

Jones, W.D., Cayirlioglu, P., Kadow, I.G., and Vosshall, L.B. (2007). Two chemosensory receptors together mediate carbon dioxide detection in Drosophila. Nature *445*, 86–90.

Kahsai, L., and Winther, A.M. (2011). Chemical neuroanatomy of the Drosophila central complex: distribution of multiple neuropeptides in relation to neurotransmitters. J. Comp. Neurol. *519*, 290–315.

Krzywinski, J., Grushko, O.G., and Besansky, N.J. (2006). Analysis of the complete mitochondrial DNA from Anopheles funestus: an improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. Mol. Phylogenet. Evol. *39*, 417–423.

Labhart, T., and Meyer, E.P. (1999). Detectors for polarized skylight in insects: a survey of ommatidial specializations in the dorsal rim area of the compound eye. Microsc. Res. Tech. *47*, 368–379.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.

Li, X., Cassidy, J.J., Reinke, C.A., Fischboeck, S., and Carthew, R.W. (2009). A microRNA imparts robustness against environmental fluctuation during development. Cell *137*, 273–282.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. *20*, 265–272.

Marygold, S.J., Roote, J., Reuter, G., Lambertsson, A., Ashburner, M., Millburn, G.H., Harrison, P.M., Yu, Z., Kenmochi, N., Kaufman, T.C., et al. (2007). The ribosomal protein genes and Minute loci of Drosophila melanogaster. Genome Biol. *8*, R216.

Mebs, D., Zehner, R., and Schneider, M. (2000). Molecular studies on the oubain binding site of the Na+, K+-ATPase in milkweed butterflies. Chemoecology *10*, 201–203.

Merlin, C., Gegear, R.J., and Reppert, S.M. (2009). Antennal circadian clocks coordinate sun compass orientation in migratory monarch butterflies. Science *325*, 1700–1704.

Moreau, C.S., Bell, C.D., Vila, R., Archibald, S.B., and Pierce, N.E. (2006). Phylogeny of the ants: diversification in the age of angiosperms. Science *312*, 101–104.

Mouritsen, H., and Frost, B.J. (2002). Virtual migration in tethered flying monarch butterflies reveals their orientation mechanisms. Proc. Natl. Acad. Sci. USA *99*, 10162–10166.

Olivier, V., Monsempes, C., François, M.C., Poivet, E., and Jacquin-Joly, E. (2011). Candidate chemosensory ionotropic receptors in a Lepidoptera. Insect Mol. Biol. *20*, 189–199.

Ooka, S., Katow, T., Yaguchi, S., Yaguchi, J., and Katow, H. (2010). Spatiotemporal expression pattern of an encephalopsin orthologue of the sea urchin Hemicentrotus pulcherrimus during early development, and its potential role in larval vertical migration. Dev. Growth Differ. *52*, 195–207.

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

Perez, S.M., Taylor, O.R., and Jander, R. (1997). A sun compass in monarch butterflies. Nature *387*, 29.

Pliske, T.E. (1975). Courtship behavior of the monarch butterfly, *Danaus plexippus* L. Ann. Entomol. Soc. Am. *68*, 143–151.

Prassas, I., and Diamandis, E.P. (2008). Novel therapeutic applications of cardiac glycosides. Nat. Rev. Drug Discov. *7*, 926–935.

Price, E.M., Rice, D.A., and Lingrel, J.B. (1990). Structure-function studies of Na,K-ATPase. Site-directed mutagenesis of the border residues from the H1-H2 extracellular domain of the alpha subunit. J. Biol. Chem. *265*, 6638–6641.

Reppert, S.M., and Weaver, D.R. (2002). Coordination of circadian timing in mammals. Nature *418*, 935–941.

Reppert, S.M., Zhu, H., and White, R.H. (2004). Polarized light helps monarch butterflies navigate. Curr. Biol. *14*, 155–158.

Reppert, S.M., Gegear, R.J., and Merlin, C. (2010). Navigational mechanisms of migrating monarch butterflies. Trends Neurosci. *33*, 399–406.

Ritland, D.B., and Brower, L.P. (1991). The viceroy butterfly is not a batesian mimic. Nature *350*, 497–498.

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell *127*, 1193–1207.

Ruby, J.G., Stark, A., Johnston, W.K., Kellis, M., Bartel, D.P., and Lai, E.C. (2007). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome Res. *17*, 1850–1864.

Sauman, I., Briscoe, A.D., Zhu, H., Shi, D., Froy, O., Stalleicken, J., Yuan, Q., Casselman, A., and Reppert, S.M. (2005). Connecting the navigational clock to sun compass input in monarch butterfly brain. Neuron *46*, 457–467.

Shafer, O.T., and Taghert, P.H. (2009). RNA-interference knockdown of Drosophila pigment dispersing factor in neuronal subsets: the anatomical basis of a neuropeptide's circadian functions. PLoS ONE *4*, e8298.

Skou, J.C. (1998). Nobel Lecture. The identification of the sodium pump. Biosci. Rep. *18*, 155–169.

Tanaka, K., Uda, Y., Ono, Y., Nakagawa, T., Suwa, M., Yamaoka, R., and Touhara, K. (2009). Highly selective tuning of a silkworm olfactory receptor to a key mulberry leaf volatile. Curr. Biol. *19*, 881–890.

Terenius, O., Papanicolaou, A., Garbutt, J.S., Eleftherianos, I., Huvenne, H., Kanginakudru, S., Albrechtsen, M., An, C., Aymeric, J.L., Barthel, A., et al. (2011). RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. J. Insect Physiol. *57*, 231–245.

Urquhart, F.A., and Urquhart, N.R. (1978). Autumnal migration routes of the eastern population of the monarch butterfly (Danaus p. plexippus L.; Danaidae; Lepidoptera) in North America to the overwintering site in the Neovolcanic Plateau of Mexico. Can. J. Zool. *56*, 1759–1764.

Velarde, R.A., Sauer, C.D., Walden, K.K., Fahrbach, S.E., and Robertson, H.M. (2005). Pteropsin: a vertebrate-like non-visual opsin expressed in the honey bee brain. Insect Biochem. Mol. Biol. *35*, 1367–1377.

Vosshall, L.B., and Hansson, B.S. (2011). A unified nomenclature system for the insect olfactory coreceptor. Chem. Senses *36*, 497–498.

Wanner, K.W., and Robertson, H.M. (2008). The gustatory receptor family in the silkworm moth Bombyx mori is characterized by a large expansion of a single lineage of putative bitter receptors. Insect Mol. Biol. *17*, 621–629.

Wanner, K.W., and Robertson, H.M. (2010). Lepidopteran chemoreceptors. In Molecular Biology and Genetics of the Lepidoptera, M.R. Goldsmith and F. Marek, eds. (Boca Raton, FL: CRC Press), pp. 153–168.

Wei, Y., Chen, S., Yang, P., Ma, Z., and Kang, L. (2009). Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust. Genome Biol. *10*, R6.

Wernet, M.F., Labhart, T., Baumann, F., Mazzoni, E.O., Pichaud, F., and Desplan, C. (2003). Homothorax switches function of Drosophila photoreceptors from color to polarized light sensors. Cell *115*, 267–279.

Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., et al; Nasonia Genome Working Group. (2010). Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science *327*, 343–348.

Xiang, H., Zhu, J., Chen, Q., Dai, F., Li, X., Li, M., Zhang, H., Zhang, G., Li, D., Dong, Y., et al. (2010). Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat. Biotechnol. *28*, 516–520.

Xiang, Z. (1995). Genetics And Breeding Of The Silkworm (Beijing, PR China: Chinese Agriculture Press).

Xu, P., Vernooy, S.Y., Guo, M., and Hay, B.A. (2003). The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. Curr. Biol. *13*, 790–795.

Yuan, Q., Metterville, D., Briscoe, A.D., and Reppert, S.M. (2007). Insect cryptochromes: gene duplication and loss define diverse ways to construct insect circadian clocks. Mol. Biol. Evol. *24*, 948–955.

Závodská, R., Sauman, I., and Sehnal, F. (2003). Distribution of PER protein, pigment-dispersing hormone, prothoracicotropic hormone, and eclosion hormone in the cephalic nervous system of insects. J. Biol. Rhythms *18*, 106–122.

Zdobnov, E.M., and Bork, P. (2007). Quantification of insect genome divergence. Trends Genet. *23*, 16–20.

Zhu, H., Yuan, Q., Briscoe, A.D., Froy, O., Casselman, A., and Reppert, S.M. (2005). The two CRYs of the butterfly. Curr. Biol. *15*, R953–R954.

Zhu, H., Casselman, A., and Reppert, S.M. (2008a). Chasing migration genes: a brain expressed sequence tag resource for summer and migratory monarch butterflies (Danaus plexippus). PLoS ONE *3*, e1345.

Zhu, H., Sauman, I., Yuan, Q., Casselman, A., Emery-Le, M., Emery, P., and Reppert, S.M. (2008b). Cryptochromes define a novel circadian clock mechanism in monarch butterflies that may underlie sun compass navigation. PLoS Biol. *6*, e4.

Zhu, H., Gegear, R.J., Casselman, A., Kanginakudru, S., and Reppert, S.M. (2009). Defining behavioral and molecular differences between summer and migratory monarch butterflies. BMC Biol. *7*, 14.

EXTENDED EXPERIMENTAL PROCEDURES

Animals

Monarchs used for genomic DNA isolation were female migrants. One female was caught in October, 2008 near Eagle Pass, TX, USA (latitude 28°71′N, longitude 100°49′W) by Carol Cullar, and two females were caught in October, 2008 near Greenfield, Massachusetts, USA (latitude 42°59′N, longitude 72°60′W) by Fred Gagnon.

Genomic Features

Illumina SIPES reads were aligned to assembly using Bowtie v0.12.7 (Langmead et al., 2009) to obtain the best alignment per read pair with the "-k 1 –best" option. The alignment output was then processed by samtools v0.1.15 (Li et al., 2009) to detect single nucleotide polymorphisms using the suggested parameter values. We identified repetitive sequences and transposable elements using RepeatMasker v3.2.9 (http://www.repeatmasker.org) against a de novo repeat library that was built by RepeatModeler v1.0.4 (http://www.repeatmasker.org), as well as the arthropod set of Repbase v20090604 (Lowe and Eddy, 1997). Non-interspersed repeat sequences were also identified by RepeatMasker with the "-noint" option. We predicted transfer RNAs (tRNA) on the repeat-masked genome using tRNAscan-SE-1.23 (Lowe and Eddy, 1997). Distribution of GC content was analyzed in 500-bp non-overlapped windows. CpG ratio, CpG[O/E], is defined as CpG[O/E] = P[CpG]/(P[C]*P[G]), in which P[CpG] is the frequency of CpG dinucleotides, P[C] the frequency of C nucleotides, and P[G] the frequency of G nucleotides.

Transcriptome Analysis

To construct the cDNA library for transcriptome analysis, monarchs from all stages of development were used to ensure a good representation of transcripts: 50 one- to two-days old eggs, one second instar larva raised on milkweed plants, one fifth instar larva raised on diet, one five-day old pupa, and male and female adults from both summer (reproductive) and migrant (non-reproductive) butterflies. To avoid plant contaminants, larvae were dissected in 0.5X RNAlater (Ambion) and their guts were emptied. Heads without antennae, legs, thoraces and abdomens from one male and one female of each state (summer or migrant) were used. Antennae were from two males and two females of each state. Male and female migrant butterflies from which heads, legs and thoraces were used were caught in October, 2008 near Eagle Pass, Texas, USA by Carol Cullar, and those from which antennae and abdomens have been used were caught in October, 2008 near Greenfield, Massachusetts, USA by Fred Gagnon. Summer butterflies were either obtained from Edith Smith (Shady Oak Butterfly Farm, Florida, USA) for all tissues except the antennae, which were obtained from butterflies provided by Orley Taylor (Kansas University, USA). All butterflies were housed in the laboratory in glassine envelopes in incubators with controlled temperature (25°C), humidity (70%), and daily lighting conditions (12h light: 12h dark). Each was fed 25% honey every other day for a week or two prior to collections. To confirm the reproductive status of the butterflies, female abdomens were dissected. Abdomens from reproductively active summer females contained mature oocytes, while those from migrants did not.

Total RNA was extracted from each developmental stage and for each tissue described above using RNeasy extraction kits (QIAGEN; RNeasy Mini kit for eggs and antennae; RNeasy Midi kit for heads, legs and second instar larva; RNeasy Maxi kit for thoraces, abdomens, fifth instar larva and pupa). For heads, thoraces, abdomens and larvae, an additional acidic phenol extraction step was added before binding to the column. Equal amounts of RNA from all preparations were pooled and the sample was stored at −80C until further use. PolyA+ RNA extraction, reverse transcription and cDNA library construction were carried out by The National Center for Genome Resources (Santa Fe, New Mexico, USA).

Gene Models

Approximately 5.4 Gb RNA-seq sequence was employed to generate the transcriptome-based gene models using TopHat v1.2.0 (Trapnell et al., 2009) and Cufflinks v0.9.3 (Trapnell et al., 2010). The invertebrate set of NCBI RefSeq proteins was used for homology search by TBLASTN. The high-scoring pairs (HSP) with $E < 10^{-5}$ were then processed by genblastA v1.0.4 (She et al., 2009) and gene structures determined by GeneWise v2.2.0 (Birney et al., 2004). Another five homology-based gene sets were developed independently using EXONERATE v2.2.0 (Slater and Birney, 2005) with gene sets of *Bombyx*, *Drosophila*, *A. gambiae*, *Tribolium*, and *A. mellifera* (Table S1G). Our *ab initio* gene sets were generated from five different predictors: AUGUSTUS v2.5 (Stanke et al., 2006), GeneMark v3.9d (Lomsadze et al., 2005), Genscan (Burge and Karlin, 1997), GlimmerHMM v3.0.1 (Majoros et al., 2004), and SNAP v2006-07-28 (Korf, 2004) (Table S1G). To train the predictors, we also manually curated 282 gene models based on unique monarch ESTs (Zhu et al., 2008a). All above individual gene models were integrated to a consensus gene set using GLEAN (Elsik et al., 2007) and Maker v2.08 (Cantarel et al., 2008), respectively. We evaluated sensitivity for each gene models using 20 cloned monarch genes and 784 manually annotated monarch genes based on *Bombyx* homology. Because GLEAN was superior to all the other gene sets, our official gene set (OGS1.0) was based on the non-redundant GLEAN models, with additional removal of genes that were flagged as repeat elements or were not supported by either homology or the transcriptome.

Orthology and Evolution

All used protein sets of other species are listed in Table S1I. First, we removed very short proteins (<30aa) and filtered out redundant splice variants to keep the longest isoform for each protein set. Next, all-against-all protein comparisons were performed using

BLASTP with $E < 10^{-5}$. We used orthomclSoftware-v2.0.2 to process HSPs and MCL v10-201 (Li et al., 2003) to define the final orthologs, inparalogs, and co-orthologs, following the suggested parameter values. Multiple alignments of protein sequences for each orthology group were performed using Muscle v 3.8.31 (Edgar, 2004) and the conserved blocks of these alignments were extracted using Gblocks v 0.91b (Talavera and Castresana, 2007). Conserved blocks of 1,642 proteins that have single copies in all species were concatenated to 14 super genes with 377,961 amino acids, which were used to quantify the phylogeny of the 12 insect species. The species tree was calculated using PhyML v2.4.4 (Guindon et al., 2010) with the JTT model. The values of statistical support were obtained from 1,000 replicates of bootstrap analyses. Muscle alignments were also processed by pal2nal v13 (Suyama et al., 2006), the resulting codon alignments were subjected to the calculation of synonymous (dS) and non-synonymous (dN) substitution rates with F3X4 codon frequency, using codeml from the PAML package v Jan-09-2011 (Yang, 2007).

Synteny

Bombyx genomic scaffolds were first concatenated with 500-bp Ns to 28 chromosome sequence according to the information shown in SilkDB 2.0 (Wang et al., 2005). Monarch genes were anchored based on the position of the best BLASTP hit found in the *Bombyx* gene set. For mapping long monarch scaffolds (>10 kb), more than half of the genes within a scaffold that show the consensus position is required to determine the corresponding position on *Bombyx* chromosomes. Pairwise whole genome alignment between the monarch and *Bombyx* was performed using LASTZ v 1.02 with HSP chaining (http://www.bx.psu.edu/~rsharris/lastz/). Because of the 'draft' status of the monarch genome, we only focused on micro-synteny, not chromosome-scale rearrangements.

Quantification of Gene Expression

Based on the transcriptome data, we estimated the general expression value for most predicted genes, except for neuropeptide-related genes, which were of short length that was beyond the library size, and antennal chemoreceptors, because of their general low expression and limited expression in specific cell types. Each predicted coding sequence was extended with 500-bp upstream and downstream regions. Paired-end transcriptome reads were mapped to the extended gene set using Bowtie with up to one alignment report per pair. Sequence coverage was defined as D = N*300/L, in which N is the number of mapped pairs of reads, and L is the length of the gene (we estimated the insert size of the RNAseq library as 300 bp). We also mapped the previously identified ESTs (Zhu et al., 2008a) to the extended gene models using BLASTN (both $E < 10^{-10}$ and identity > 92% are required). Expression levels for summer and migratory states were calculated based on the raw microarray data (GSE14041 of GEOdatabase). The independent two-sample t test was used to compare expression values between summer and migratory groups in males and females, respectively.

Annotation of Coding Genes

For automatic annotation, we searched the homology by querying the *Bombyx*, *Drosophila*, and NCBI RefSeq invertebrate protein sets, as well as Gene Ontology and KEGG databases. A local run of InterProScan (IPR) search (Hunter et al., 2009) with all implemented methods was also carried out to identify the conserved domains for gene sets of the monarch, *Bombyx*, *Drosophila*, and *Tribolium*. All above databases were updated to April 2011 for annotation. Species-specific expansion/contraction was determined with the significance of pairwise comparison of the IPR-defined family sizes, which was estimated by the Chi-square test with respect to the predicted number of genes with IPR domains. Several IPR families that are usually found in transposons or are problematic for automatic prediction were omitted, including reverse transcriptase, integrase, zinc finger proteins, and olfactory receptors. Species-specific families, which were missed in all other three species, were also not included in the list. For lepidopteran-specific expansion/contraction, species-specific changed families were first excluded and then families were ordered based on the size difference between the sum of genes in the two lepidopteran and the two non-lepidopteran species.

More than 1,000 genes of biological interest were manually annotated, using *Drosophila*, human, and some well-characterized *Bombyx* orthologs available on NCBI GenBank as queries in most cases. Part of the functional information of *Drosophila* homologs was referred to The Interactive Fly (http://www.sdbonline.org/fly/aimain/1aahome.htm) and GenAge (http://genomics.senescence.info/genes/models.html). Genes with incomplete structures or inappropriate concatenation were identified based on multiple alignments by ClustalX 2.1 (Larkin et al., 2007). If the target homology was not identified in the gene set, additional searches in the genome assembly (by TBLASTN) or raw reads (by Bowtie) was carried out to confirm gene loss. Actually, we have not found, to date, any target gene that only exists in the genome and is not represented in the gene set, which confirms the completeness of our gene model.

Circadian Genes

Drosophila and human sets of clock genes were both utilized to BLASTP search the monarch gene set and other arthropod gene sets. In addition to reciprocal blast, an initial round of phygenetic analysis was performed for cryptochrome (CRY) families to remove the members in (6-4)-photolyase and Cry-DASH clades. This method was also used to differentiate timeless and timeout orthologs. Phylogenetic analysis was performed using PhyML.

We identified the monarch pigment-dispersing factor gene (*pdf*) based on PF06324 domain (PDF domain in Pfam), as this gene is very short and highly divergent in the N-terminal part of the protein sequence. Because our current transcriptome did not capture the

transcript(s) of *pdf*, we performed additional polymerase chain reaction (PCR) amplication of cDNA to verify its expression in brain. Total RNA was extracted from a male butterfly brain using RNeasy Mini extraction kit (QIAGEN) and cDNA was synthesized using SuperScript II reverse transcriptase (Invitrogen). The primers were designed to span a 1.4 Kb intron and the full-coding region of the peptide, as follow: pdfF, 5′-GCTCTCCCAGCTACGAACTCTA-3′; pdfR, 5′-GATATTCCCGCCATAGACTTG-3′. PCR conditions were as follow: after 5 min at 94°C, five cycles of 30 s at 94°C, 30 s at 49°C, 45 s at 72°C, then 35 cycles of 30 s at 94°C, 30 s at 52°C, 45 s at 72°C, then 5 min of final elongation step at 72°C.

Chemosensory Receptors

Because chemosensory receptor genes are difficult to identify from automated predictions, we identified this class of genes in the genome assembly using TBLASTN searches with *Bombyx*, *Drosophila*, and the moth *Spodoptera littoralis* (only for ionotropic receptors) homologs as queries, followed by iteration. For the genomic loci with significant hits ($E < 10^{-3}$), we compared all independent gene sets or re-annotated the exons using GeneWise. Multiple alignments of selected protein sequences were performed using ClustalX. The well-aligned regions were analyzed for phylogenetic analysis using protdist software from the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) with 1,000 replicates of bootstrap analysis.

miRNAs

Migrant butterflies were caught in October, 2010 near Eagle Pass, Texas, USA by Carol Cullar. Total RNA was extracted from 10 summer butterflies and from 10 migrants with Trizol (Invitrogen) and equally pooled from each individual of the two sets for two independent miRNA sequencing lanes (summer and migrant). miRNAs separation, library construction, and Illumina sequencing were conducted by Eureka Genomics. Processed small RNA reads were aligned against the monarch genome by Bowtie, allowing one mismatch. Secondary structures were predicted using RNAfold v1.8.4 (Hofacker, 2003). miRNAs were primarily analyzed by miRDeep pipeline (Friedländer et al., 2008) and manually sorted to remove redundancy. Conserved miRNAs were named according to the unified nomenclature system of miRBase release16 (Kozomara and Griffiths-Jones, 2011). Another two rounds of prediction were conducted using miRTRAP v1.0 (Hendrix et al., 2010) and mireap v0.2 (http://sourceforge.net/projects/mireap/) pipelines. Novel miRNAs that were predicted by all three methods were considered as monarch specific. Remaining mapped reads were aligned to monarch gene models and Rfam r10.0 (Gardner et al., 2009) to identify degraded mRNAs and other non-coding RNAs, respectively. The miRNA expression value for each of the two profiles (summer versus migrant) was normalized to the total number of valid RNA sequence reads per profile.

SUPPLEMENTAL REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Arensburger, P., Megy, K., Waterhouse, R.M., Abrudan, J., Amedeo, P., Antelo, B., Bartholomay, L., Bidwell, S., Caler, E., Camara, F., et al. (2010). Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. Science *330*, 86–88.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988–995.

Bonasio, R., Zhang, G., Ye, C., Mutti, N.S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C., et al. (2010). Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. Science *329*, 1068–1071.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Cantarel, B.L., Korf, I., Robb, S.M., 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.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., et al; Drosophila 12 Genomes Consortium. (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203–218.

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., et al. (2011). The ecoresponsive genome of Daphnia pulex. Science *331*, 555–561.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., and Bateman, A. (2009). Rfam: updates to the RNA families database. Nucleic Acids Res. *37* (*Database issue*), D136–D140.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Hendrix, D., Levine, M., and Shi, W. (2010). miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. Genome Biol. *11*, R39.

HGSC (Honeybee Genome Sequencing Consortium). (2006). Insights into social insects from the genome of the honeybee Apis mellifera. Nature *443*, 931–949.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res. *31*, 3429–3431.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., et al. (2002). The genome sequence of the malaria mosquito Anopheles gambiae. Science *298*, 129–149.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al. (2009). InterPro: the integrative protein signature database. Nucleic Acids Res. *37* (*Database issue*), D211–D215.

IAGC (International Aphid Genomics Consortium). (2010). Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biol. *8*, e1000313.

IHGSC (International Human Genome Sequencing Consortium). (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., Clark, J.M., Lee, S.H., Robertson, H.M., Kennedy, R.C., Elhaik, E., et al. (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc. Natl. Acad. Sci. USA *107*, 12168–12173.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. *39* (*Database issue*), D152–D157.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. *33*, 6494–6506.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics *20*, 2878–2879.

Nene, V., Wortman, J.R., Lawson, D., Haas, B., Kodira, C., Tu, Z.J., Loftus, B., Xi, Z., Megy, K., Grabherr, M., et al. (2007). Genome sequence of Aedes aegypti, a major arbovirus vector. Science *316*, 1718–1723.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. (2005). Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res. *15*, 1–18.

Richards, S., Gibbs, R.A., Weinstock, G.M., Brown, S.J., Denell, R., Beeman, R.W., Gibbs, R., Beeman, R.W., Brown, S.J., Bucher, G., et al; Tribolium Genome Sequencing Consortium. (2008). The genome of the model beetle and pest Tribolium castaneum. Nature *452*, 949–955.

She, R., Chu, J.S., Wang, K., Pei, J., and Chen, N. (2009). GenBlastA: enabling BLAST to identify homologous gene sequences. Genome Res. *19*, 143–149.

Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

Smith, C.D., Zimin, A., Holt, C., Abouheif, E., Benton, R., Cash, E., Croset, V., Currie, C.R., Elhaik, E., Elsik, C.G., et al. (2011a). Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). Proc. Natl. Acad. Sci. USA *108*, 5673–5678.

Smith, C.R., Smith, C.D., Robertson, H.M., Helmkampf, M., Zimin, A., Yandell, M., Holt, C., Hu, H., Abouheif, E., Benton, R., et al. (2011b). Draft genome of the red harvester ant Pogonomyrmex barbatus. Proc. Natl. Acad. Sci. USA *108*, 5667–5672.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. *34* (*Web Server issue*), W435-W439.

Suen, G., Teiling, C., Li, L., Holt, C., Abouheif, E., Bornberg-Bauer, E., Bouffard, P., Caldera, E.J., Cash, E., Cavanaugh, A., et al. (2011). The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle. PLoS Genet. *7*, e1002007.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34* (*Web Server issue*), W609–W12.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564–577.

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

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Wang, J., Xia, Q., He, X., Dai, M., Ruan, J., Chen, J., Yu, G., Yuan, H., Hu, Y., Li, R., et al. (2005). SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic Acids Res. *33* (*Database issue*), D399–D402.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al; Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Wurm, Y., Wang, J., Riba-Grognuz, O., Corona, M., Nygaard, S., Hunt, B.G., Ingram, K.K., Falquet, L., Nipitwattanaphon, M., Gotzek, D., et al. (2011). The genome of the fire ant Solenopsis invicta. Proc. Natl. Acad. Sci. USA *108*, 5679–5684.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

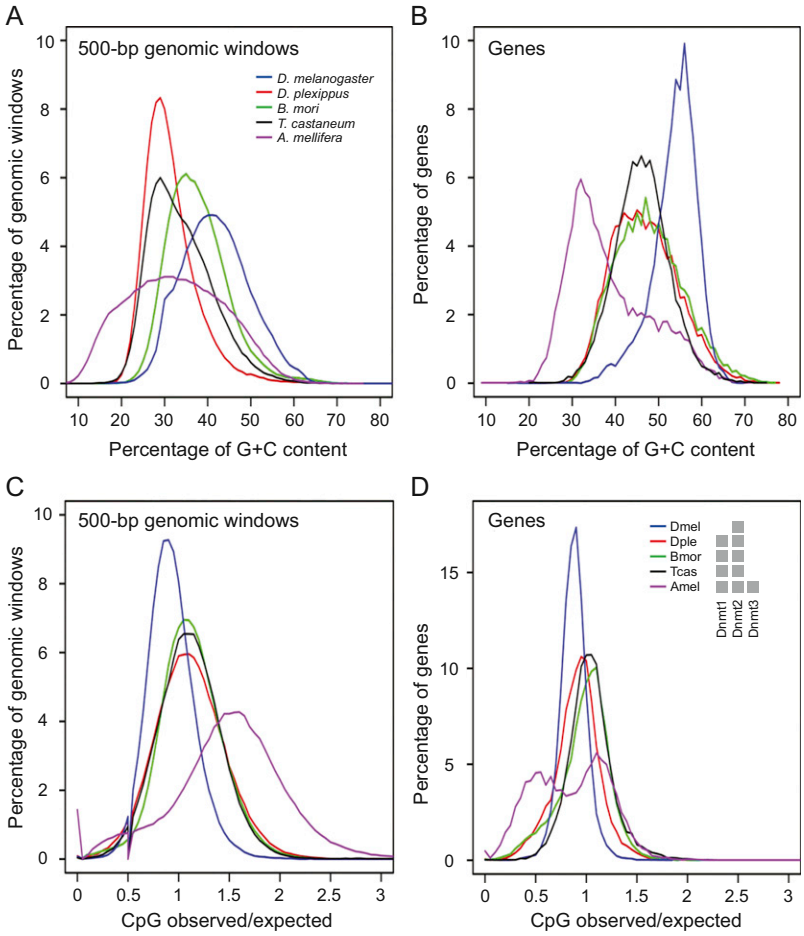

Figure S1. Comparisons of GC Content, CpG Ratios, and DNA Methylation Potential, Related to Table 1

Shown are GC content in the genome (A) and coding regions (genes, B), and CpG ratios in the genome (C), and the coding regions (D). Values are plotted against the frequency. For the genome, GC content and CpG ratios were calculated using 500-bp sliding windows of genomic sequence. Red, monarch; green, *Bombyx*; blue, *Drosophila;* black, *Tribolium*; purple, *A. mellifera*. Grey squares in (D) show the existence of the corresponding member(s) of DNA methyltransferase (Dnmt) family for each species.

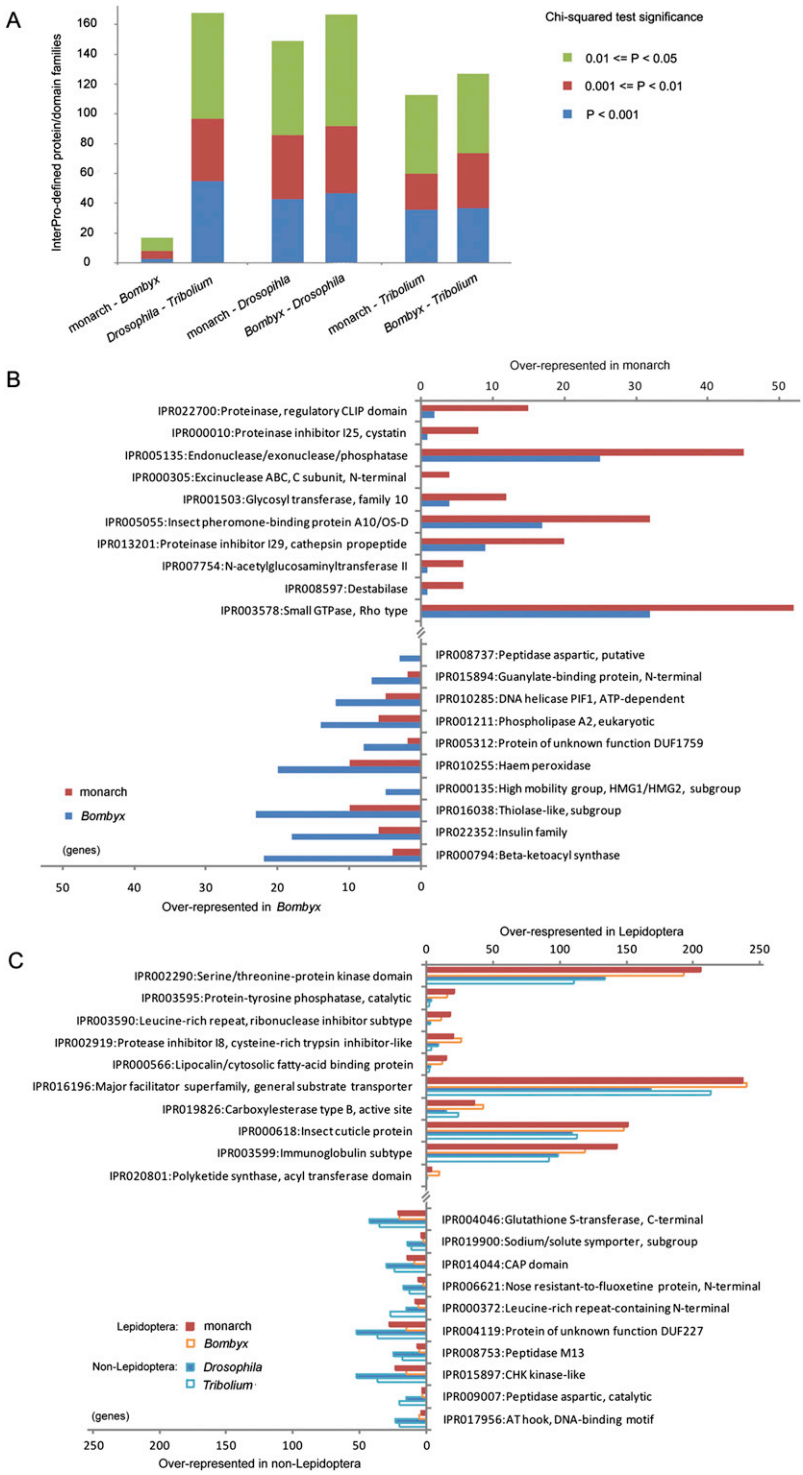

Figure S2. Comparison of Protein Domains, Related to Figure 2

(A) Pairwise comparison of the InterPro (IPR)-defined family sizes. Bars indicate the number of significantly differing families between each pair of species; color denotes degree of significance. The significance was determined by the Chi-square test with respect to the predicted number of genes with IPR domains.
(B) The ten most prominent expansions (upper) and contractions (lower) of monarch IPR families compared to *Bombyx*, listed in decreasing order of significance.
(C) The ten most prominent expansions and contractions of lepidopteran IPR families compared to two non-lepidopteran insect species, *Drosophila* or *Tribolium*.
See also Extended Experimental Procedures for the definitions of expansion and contraction.

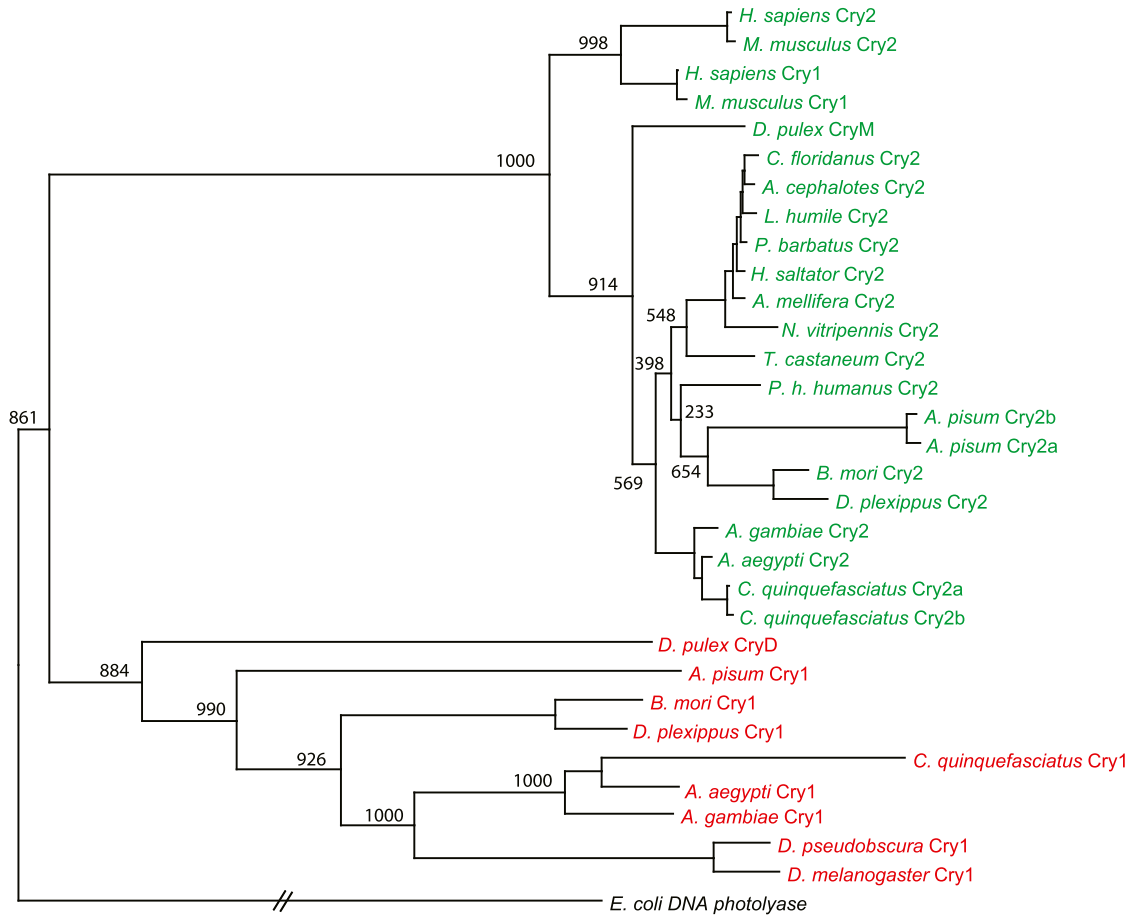

Figure S3. Animal CRYPTOCHROME Phylogeny, Related to Figure 3

Maximum likelihood phylogenetic tree showing the evolution of the type-1 (*Drosophila*-like, red lettering) and type 2 (vertebrate-like, green) CRYs in all the arthropods for which draft genomes are available. The tree was rooted with the *E. coli* DNA photolyase. Boostrap values based on 1000 replicates are represented at the nodes. *A. aegypti*: *Aedes aegypti*; *A. cephalotes*: *Atta cephalotes*; *A. gambiae*: *Anopheles gambiae*; *A. mellifera*: *Apis mellifera*; *A. pisum*: *Acyrthosiphon pisum*; *B. mori*: *Bombyx mori*; *C. floridanus*: *Camponotus floridanus*; *C. quinquefasciatus*: *Culex quinquefasciatus*; *D. melanogaster*: *Drosophila melanogaster*; *D. plexippus*: *Danaus plexippus*; *D. pseudoobscura*: *Drosophila pseudoobscura*; *D. pulex*: *Daphnia pulex*; *E. coli*: *Escherichia coli*; *H. saltator*: *Harpegnathos saltator*; *H. sapiens*: *Homo sapiens*; *L. humile*: *Linepithema humile*; *M. musculus*: *Mus musculus*; *N. vitripennis*: *Nanosia vitripennis*; *P. barbatus*: *Pogonomyrmex barbatus*; *P. h. humanus*: *Pediculus humanus humanus*; *T. castaneum*: *Tribolium castaneum*.

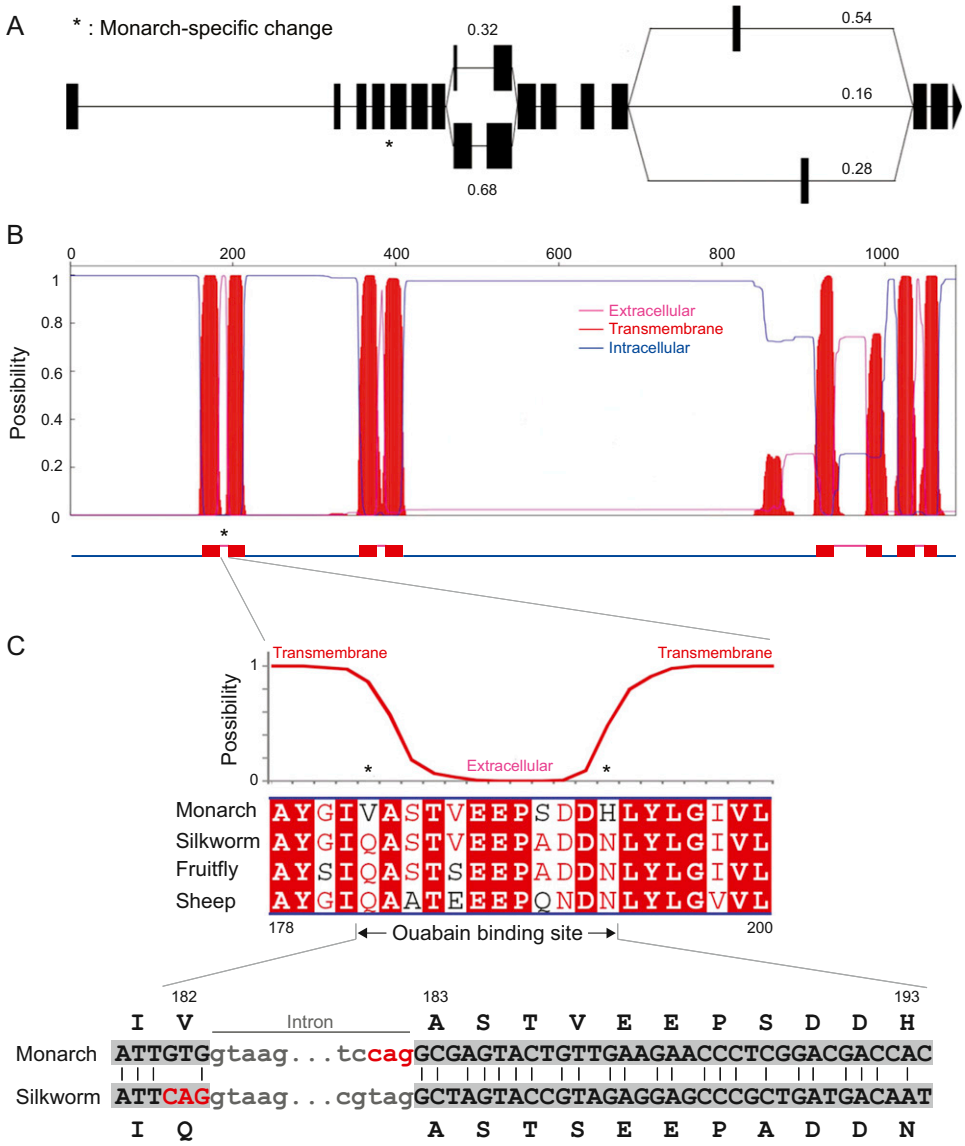

Figure S4. Major α Subunit Gene of Monarch P-Type Na^+/K^+ ATPase, Related to Figure 1

(A) Schematic of the genome structure of the major sodium/potassium pump α subunit gene. Black boxes indicate exons and alternative splicing patterns, which were manually curated using transcriptome sequence. The fraction of each splicing pattern is shown around the corresponding positions. Asterisk indicates the position of monarch-specific changes.

(B) Hypothetical secondary structure of the α subunit. The secondary structure is based on the topology prediction method, TMHMM Server v. 2.0 (Krogh et al., 2001). The predicted extracellular, intercellular, and transmembrane domains were plotted. Eight major hydrophobic (transmembrane) regions are shown as red peaks.

(C) Monarch-specific mutations within the α subunit of Na^+/K^+ ATPase. Multiple alignment of the entire sequence revealed only two monarch-specific mutations, Q(Glu)182V(Val) and N(Asn)193H(His), which are indicated by asterisks. The previous work (Holzinger et al., 1992) was based on DNA sequencing only and missed Q182V because of the mis-splicing of CAG in the intron (in red in lowercase) to the coding region. The magnified region of the first extracellular domain shows the correct splicing pattern.