HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2015 April 16.

Published in final edited form as:

Nature. 2014 October 16; 514(7522): 317–321. doi:10.1038/nature13812.

The genetics of monarch butterfly migration and warning coloration

Shuai Zhan,
Key Laboratory of Insect Developmental and Evolutionary Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, PR China.

Department of Ecology & Evolution, University of Chicago, Chicago, IL 60637, USA.

Department of Neurobiology, University of Massachusetts Medical School, Worcester, MA 01605, USA.

Wei Zhang,
Department of Ecology & Evolution, University of Chicago, Chicago, IL 60637, USA.

Kristjan Niitepõld,
Department of Biology, Stanford University, Stanford, CA 94305, USA.

Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA

Jeremy Hsu,
Department of Biology, Stanford University, Stanford, CA 94305, USA.

Juan Fernández Haeger,
Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba, 14071 Córdoba, Spain.

Myron P. Zalucki,
School of Biological Sciences, The University of Queensland, Brisbane, Australia 4072.

Sonia Altizer,
Odum School of Ecology, University of Georgia, Athens, GA 30602, USA.

Jacobus C. de Roode,
Department of Biology, Emory University, Atlanta, GA 30322, USA.

Steven M. Reppert, and

Correspondence and requests for materials should be addressed to SZ (szhan@sibs.ac.cn) or MRK (mkronforst@uchicago.edu)..

Author Contributions

SZ designed and implemented analyses of dispersal and migration and co-wrote the manuscript. WZ performed wing color analyses. KN performed respirometry experiments. JH helped design the project and collected and prepared samples for sequencing. JFH and MPZ provided samples and interpreted results. SA, JCdR and SMR helped design the project, provided samples, and interpreted results. MRK conceived and directed the project, performed targeted population genetic analyses, and co-wrote the manuscript.

Sequence data are available from NCBI SRA (SRP045457, SRP045468)

The authors declare no competing financial interests.

Department of Neurobiology, University of Massachusetts Medical School, Worcester, MA 01605, USA.

Marcus R. Kronforst
Department of Ecology & Evolution, University of Chicago, Chicago, IL 60637, USA.

Abstract

The monarch butterfly, *Danaus plexippus*, is famous for its spectacular annual migration across North America, recent worldwide dispersal, and orange warning coloration. Despite decades of study and broad public interest, we know little about the genetic basis of these hallmark traits. By sequencing 101 monarch genomes from around the globe, we uncover the history of the monarch's evolutionary origin and global dispersal, characterize the genes and pathways associated with migratory behavior, and identify the discrete genetic basis of warning coloration. The results rewrite our understanding of this classic system, showing that *D. plexippus* was ancestrally migratory and dispersed out of North America to occupy its broad distribution. We find the strongest signatures of selection associated with migration center on flight muscle function, resulting in greater flight efficiency among migratory monarchs, and that variation in monarch warning coloration is controlled by a single myosin gene not previously implicated in insect pigmentation.

Introduction

Every year millions of monarch butterflies fly from the northern United States and southern Canada to overwinter in Mexico. Amazingly, during this portion of the annual migration, individual butterflies emerge as adults in the north, fly thousands of kilometers south, overwinter for months in reproductive diapause, and finally begin mating and flying north in the spring. Recolonization of northern latitudes takes place over the course of three to four subsequent generations, after which it is late summer again and the process repeats itself. Although most North American monarchs overwinter in Mexico, those that live west of the Rocky Mountains generally overwinter along the California coast[1,2]. Monarch migration has been studied extensively over the past few decades, with particular emphasis on tracking migration routes and overwintering sites[1,3-7], as well as characterizing the navigational mechanisms that guide this complex behavior[8-13]. A largely unappreciated aspect of this system is that not all monarchs migrate. In fact, the geographic distribution of *D. plexippus* extends far beyond North America and it does not migrate across most of its range. For instance, while the monarch undergoes extensive, long-distance migration across North America, it also exists throughout Central America, South America, and the Caribbean, where it does not migrate[14-16]. Furthermore, the monarch has recently dispersed to many locations around the globe where it does not exhibit the same long-distance migration found in North America[17,18].

The molecular genetic mechanisms that contribute to migration are likely to be complex, spanning pathways related to navigation and circadian rhythms, environmental sensing, energy production and metabolism, thermal tolerance, reproduction and longevity, neuromuscular development, and phenotypic plasticity[11,19,20]. Because *D. plexippus* exists

as both migratory and non-migratory populations, we sought to use comparative population genomics to characterize the genetic basis of migration by identifying genome-wide targets of divergent natural selection associated with shifts in migratory behavior. To do this, we sequenced the genomes of 89 butterflies - 80 *D. plexippus* and nine samples from four additional *Danaus* species - from across their worldwide distribution (Fig. 1a, b) and analyzed them using the monarch reference genome sequence[20].

To guide our analysis, we first characterized the evolutionary origin of *D. plexippus* and its history of dispersal. The genus *Danaus* is broadly tropical and non-migratory[14] which suggests non-migratory populations of *D. plexippus* in South and Central America may represent the ancestral source of the North American population[21], similar to the "southern home theory" for the origin of seasonal migration in birds[22]. More recently, monarchs dispersed across the Pacific, throughout Oceania and Australia, and they also dispersed across the Atlantic to Europe and Africa[17]. It is unknown whether the Pacific and Atlantic dispersal events were independent, but both are thought to derive from North American migratory monarchs[17,23]. There is also an enigmatic non-migratory population in south Florida[15], which may result from overwintering migratory butterflies that fail to return north.

Evolutionary and demographic history

Our analysis, based on genome-wide SNP variation (approximately 32 million SNPs), revealed a history quite distinct from our *a priori* expectations. For instance, we recovered North American populations as the most basal lineages, with Central/South America, Pacific, and Atlantic populations each forming an independent, derived lineage (Fig. 1 c, d; Extended Data Fig. 1). We also found that all geographic sampling locations were genetically distinct, except those in North America, providing evidence of worldwide population structure but gene flow across North America (Fig. 1e, f). An exception was the non-migratory population in south Florida, which was distinct from other North American populations. We also found evidence for very recent dispersal between the North American migratory population and adjacent non-migratory populations (Supplementary Information). Our results suggest the monarch originated in North America from a migratory ancestor, a scenario consistent with the observation that all monarch populations, as well as *D. erippus*, share reproductive traits and behaviors which may have evolved in the context of mass migration[24]. We speculate the monarch originated in the southern USA or northern Mexico, where it originally undertook a shorter-distance annual migration. Three subsequent, independent dispersal events led to the monarch's current broad distribution. Towards the south, monarchs expanded from Belize to Costa Rica and into South America, as well as offshore, from south Florida to Bermuda and Puerto Rico. Westwards, they expanded into Hawaii and then to Samoa and Fiji before ending up in New Caledonia, Australia and New Zealand. Across the Atlantic, monarchs established first in Portugal and then moved to Spain and Morocco.

Our dispersal scenario was further supported by the observation that non-North American populations had elevated linkage-disequilibrium (Extended Data Fig. 2a) and minor allele frequencies (Extended Data Fig. 2b), indicative of founder effects, and heterozygosity

declined along each putative dispersal route (Extended Data Fig. 2c), as expected of step-wise dispersal. Similar step-wise dispersal is reflected in microsatellite markers[25]. The directionality index Ψ for range expansions[26] also supported North America as the monarch's ancestral origin (Extended Data Table 1). We estimated historical population sizes and divergence times using PSMC[27] (Extended Data Fig. 2d) and $\partial a \partial i$[28] (Extended Data Fig. 3) and found a concordant history among all monarch populations, but distinct from *D. erippus*, for most of the last 1 My. Approximately 20 ky ago, at the end of the last glacial maximum, the North American population began to grow, presumably fueled by increasing availability of milkweed host plants throughout the Midwestern USA and expanding monarch migration. More recently, Atlantic and Pacific populations split from North America, at which time both experienced dramatic declines in population size. Historical records suggest Atlantic and Pacific dispersal events occurred in the 1800's[17], but our results suggest an earlier timeframe (Supplementary Information).

Natural selection associated with migration

We used a modified version of the population-branch statistic (PBS)[29] to identify regions of the genome strongly differentiating North American monarchs from all three transitions to non-migratory behavior (Fig. 2a). By further limiting this search to regions of low sequence diversity within North America, we isolated 5.14 Mb (2.1% of the genome), encompassing 536 genes, significantly associated (Z-test, $P < 0.01$) with migration. This set was enriched for genes related to morphogenesis, neurogenesis, and extracellular matrix/basement membrane. Derived-allele frequency was elevated among monarchs in these migration-associated genomic regions (Fig. 2a), further suggesting a history of natural selection. Derived alleles were similarly enriched in *D. erippus*, consistent with the observation that this species also displays migratory behavior[14,17], and suggesting the common ancestor of *D. plexippus* and *D. erippus* was migratory as well.

We were surprised to find that among the approximately 5 Mb associated with migration, a single 21 kb genomic segment stood out as an extreme outlier (Fig. 2a). This region showed multiple signatures of divergent selection (Fig. 2b) as well as an enrichment of shared alleles (ABBA vs. BABA sites[30]) among Atlantic, Pacific, and Central/South American dispersal events, indicating a shared haplotype among all non-migratory populations that was highly divergent from the haplotype in North America. This signature of haplotype sharing among non-migratory populations is likely due to recurrent selection on ancestral variation, as opposed to gene flow, because the non-migratory haplotype was present at low frequency in our North American samples (e.g., sample 203 from New Jersey was heterozygous).

The 21 kb outlier region contained three genes, the F-box protein FBXO45, an uncharacterized transmembrane protein, and collagen type IV, subunit α-1 (Fig. 2c). By comparing these three genes we found evidence for divergent selection on collagen IV α-1 (Fig. 3). Collagen IV α-1 showed striking divergence between haplotypes found in migratory and non-migratory populations (Fig. 3a, b), apparently resulting from an ancient origin of the non-migratory haplotype (Fig. 3c). Collagen IV is a central component of basement membranes and essential for muscle morphogenesis and function[31]. Mutations in the α-1 subunit result in severe myopathy in *Drosophila*[32] and myopathy-related disease in

humans[33]. Migratory and non-migratory haplotypes differed by 51 nucleotide substitutions in the coding sequence of collagen IV α-1, resulting in 15 amino acid substitutions. A subsection of the gene showed particularly high diversity within *D. plexippus*, as well as divergence between *D. plexippus* and other species, centered on the single amino acid substitution with evidence of positive selection, R1573Q (Fig. 3d). Collagen type IV is a heterotrimer composed of two α-1 chains and one α-2 chain, which bind at shared triple helix domains. *Danaus plexippus* collagen IV α-1 contains five triple helix domains and the R1573Q substitution occurred directly in the middle of one of these, suggesting a functional role related to trimerization. Interestingly, we found collagen IV subunit α-2 in a nearby genomic window, with reduced but still highly significant signatures of selection (Extended Data Table 1), providing additional evidence of selection on the interacting members of collagen IV. Furthermore, other genomic intervals strongly associated with migration overlapped with portions of another well-characterized flight muscle gene, *kettin*[34] (Extended Data Table 2).

We hypothesized that the signatures of divergence associated with these essential muscle genes reflected selection for different flight muscle function between migratory and non-migratory butterflies. Consistent with this, we found divergent expression of collagen IV α-1 and α-2, but not other linked genes, between butterflies from migratory and non-migratory populations in adult thoracic muscle tissue (Fig. 3e). Surprisingly, collagen IV subunit α-1 and α-2 were down-regulated in migratory butterflies, leading us to hypothesize that natural selection may be acting on aspects of flight efficiency with migratory populations tuned to the distinct demands of long-distance flight. This scenario is supported by evidence of distinct wing shape and size, body mass and kinematic wing loading between migratory and non-migratory populations[15], but we tested it by measuring flight metabolic rates. We found active flight to be exceptionally demanding energetically, utilizing 25 times more energy than resting. Migrating monarchs are known to offset this to some extent by gliding for periods of time on tail winds[35]. Consistent with our hypothesis that flight muscle changes have resulted in more efficient energy consumption in migratory populations, we found that flight metabolic rates were lower in butterflies from a migratory population (Massachusetts), compared to one non-migratory population (south Florida) (Fig. 3f). This increase in metabolic efficiency appears to be a result of flight muscle performance because the difference between migratory and non-migratory populations was minimal when not in flight (Fig. 3f). Furthermore, we found little sequence or gene expression divergence associated with glycolytic enzymes (Supplementary Information), which could also influence metabolic rates[36]. It is interesting that while previous work has found a link between flight metabolism and dispersal ability in other butterfly species[36-38], it has always been via glycolysis and generally in the form of higher metabolic rates yielding greater dispersal. In contrast, the extreme distances required of monarch migration appear to have generated natural selection for reduced flight metabolism, which has been mediated by alternate mechanisms. Parallel shifts to the same non-migratory collagen IV α-1 haplotype in independent dispersal events suggest that an elevated flight metabolic rate may be beneficial in the absence of long-distance migration.

The genetic basis of wing pigmentation

Another aspect of monarch biology that has attracted attention is their bold warning coloration. The monarch butterfly, like *Danaus* species generally[14], is characterized by bright orange wing coloration that warns predators of their toxicity[39]. This coloration also serves to facilitate Müllerian mimicry between *D. plexippus* and the Viceroy butterfly, *Limenitis archippus*[40]. It is not well-appreciated that the monarch is polymorphic for wing coloration (Fig. 4a). On Oahu, Hawaii, the white '*nivosus*' morph has been documented since the mid 1890's[41]. Previous breeding experiments have shown that wing coloration segregates as a single, autosomal locus with the white *nivosus* allele recessive to wild-type[42]. The *nivosus* wing coloration and inheritance pattern has led to speculation that the mutation likely disrupts the production of orange pigment[17]. Red, orange, and brown pigments on nymphalid butterfly wings are frequently ommochrome pigments[43] so we hypothesized that the *nivosus* mutation would be found somewhere in the ommochrome biosynthesis pathway.

Our population genomic data provided a means of characterizing the monarch color switch locus at a molecular level. To do this, we sequenced 12 additional Hawaiian monarch genomes, five white individuals and seven wild-type orange monarchs, all reared at the same time, in the same location. Three of these wild-type monarchs were known relatives of white monarchs, two were F1 offspring and one was an F2 offspring of a white parent. By scanning SNP genotypes for segregation patterns consistent with the Mendelian genetics (Fig. 4b), cross-referencing genotypes across the entire set of 101 genomes, and further testing co-segregation in crosses (Fig. 4c), we found that markers in one gene, the myosin gene DPOGS206617, were strongly associated with wing color. Interestingly, this gene is homologous to myosin 5a, which is responsible for the 'dilute' mouse coat color mutant[44], a phenotype resulting from reduced melanization due to impaired myosin transport of melanosomes[45]. This suggests that an alteration in pigment transport, and not pigment production, underlies the *nivosus* form. We speculate that this gene, while not previously implicated in insect pigmentation, may be an important source of color pattern variation across the butterfly subfamily Danainae because genera closely related to *Danaus* are dominated by white-winged species, and other *Danaus* species display similar orange vs. white variation[14].

Discussion

We have leveraged fantastic natural variation and extensive genome sequencing to characterize the monarch butterfly's evolutionary origin and history of dispersal, genome-wide signatures of divergent selection associated with migratory behavior, and the discrete genetic basis of warning coloration. Our results yielded unexpected answers in all three aspects. Not only did we re-polarize the ancestral migratory character state and geographic origin for the monarch, but we also found evidence for recurrent, divergent selection on flight muscle function during shifts in migratory behavior, likely mediated by their role in influencing flight efficiency. Surprisingly, as monarchs have reverted to a non-migratory state, which is an ancestral state that predates their own species and that of their common ancestor with *D. erippus*, in the case of collagen IV α-1 at least, they appear to have used

old genetic variation to do so. Furthermore, wing color variation is mediated by a gene with no prior known role in insect pigmentation but with an analogous effect in vertebrates, but one that influences a different pigment in a distinct morphological structure.

Unfortunately, the monarch migration is currently experiencing a devastating decline and there is fear the phenomenon may disappear entirely. Recent monitoring shows an alarming downward trend in monarch numbers from eastern North America, with 2013 marking the lowest number of overwintering monarchs in recorded history[46]. This decline has been driven by multiple factors, including deforestation, drought, and a precipitous drop-off in the number of milkweed host plants across North America[46]. Our results emphasize the importance of ongoing conservation efforts to preserve the migration and extend the extraordinary evolutionary history of this iconic butterfly.

Methods

Sampling and sequencing

We sampled a total of 101 butterflies, including 92 *D. plexippus* and nine butterflies from other *Danaus* species (Supplementary Table 1). The North America population of monarchs undergoes a yearly migration from the United States and southern Canada to central Mexico (east of the Rocky Mountains) and coastal California (west of the Rocky Mountains). Our sampling sites for the migratory group covered several stopover points and the overwintering grounds for both eastern and western populations (Fig. 1a, green points). For comparison, we sampled residential, non-migratory populations from three major geographic regions. In the south, non-migratory populations were sampled from south Florida (around the city of Miami), throughout the Caribbean, and Central and South America (Fig. 1a, red points). Out of the Americas, we also sampled monarch populations across the Pacific (Fig. 1a, blue points), including Hawaii and Oceania, and populations across the Atlantic (Fig. 1a, purple points), spanning Iberia and North Africa.

It is important to note that *D. plexippus* is known to move seasonally outside of North America. For instance, seasonal movement through mountain passes has been recorded in Costa Rica[47] and seasonal movement between inland and coastal locations is well-known in Australia[48]. While this behavior is properly referred to as migration, it is very different from the migration of North American monarchs and these phenomena are routinely distinguished in the literature. First, these phenomena differ by orders of magnitude in their scope, with millions to a billion individuals moving thousands of kilometers in North America and hundreds to thousands individuals moving tens to hundreds of kilometers elsewhere. Second, outside of North America overwintering biology, such as sexual diapause, is highly labile or not known to exist. Similarly, other *Danau*s species, such as *D. chrysippus* are well-known to move seasonally[49] but the scale is relatively small in comparison to *D. plexippus* in North America. There is, however, evidence that *D. erippus* migrates in South America in a way that mirrors *D. plexippus* migration in North America[17]. Critically, our genetic data support these expectations and distinctions by showing that particular genetic signatures distinguish taxa (populations/species) with differing migratory behavior based on the literature. In our study, we refer to *D. plexippus* populations from North America

(excluding south Florida) as ‘migratory’ and *D. plexippus* populations from other geographic locations as ‘non-migratory’ to capture this distinction in their biology.

All samples were sequenced on the Illumina sequencing platform (HiSeq 2000). Paired-end libraries were prepared using an Illumina paired-end library kit. We combined between 4-8 samples in a sequencing lane (2×100 bp) to generate approximately 10X and 20X raw coverage for *D. plexippus* and outgroup species, respectively. A total of 384.6 Gb of paired-end sequence data were generated (Supplementary Table 2).

Alignment, SNP calling, and genotyping

Before mapping, all reads were processed for quality control and filtered using Seqtk (https://github.com/lh3/seqtk). 366.9 Gb high-quality read pairs were kept and mapped to the latest version monarch assembly (Supplementary Table 2). Based on the result of our preliminary test, we chose Stampy v1.0.21[50] as our main mapping software, but we also applied other mapping methods as independent quality controls (Supplementary Table 3). A randomly selected subset of reads was mapped in advance to estimate the appropriate parameters for insert size and substitution rate for each library. Mapping results were subsequently processed by sorting, indel realignment, duplicate marking, and low quality filtering using functions in Picard v1.8 (http://picard.sourceforge.net) and GATK2[51]. Sequencing coverage and depth for each sample were calculated using the “DepthOfCoverage” module of GATK2.

Since we had a wide range of sequencing coverage among samples (Supplementary Table 2), we carried out SNP calling and genotyping at two separate stages to balance the power between low and high coverage samples. We first discovered variants on a population-scale using a variety of independent pipelines, which both introduced different alignment inputs and covered most popular SNP calling algorithms (Supplementary Table 3). By comparing methods, we determined a core set of SNPs according to the sensitivity and specificity of each pipeline, as well as using combinations of pipelines (Supplementary Table 4 and Supplementary Table 5). Using the consensus set of SNPs, we went back to each sample to estimate the corresponding genotype likelihoods from all alignment sources. Based on the comparison, genotypes called from the stampy alignment showed the overall minimum difference with other independent methods (Supplementary Table 6). We further filtered out variants from regions with abnormal sequencing coverage and constructed a core SNP matrix. In this final dataset, 99.1% of the genotypes were independently supported by additional evidence, suggesting a reliable input for the subsequent population genetic analysis. Also, unlike results obtained using a single SNP scoring pipeline, this method substantially reduced the correlation between the identified number of SNPs and sequencing depth.

Outgroup species

Since divergence between *D. plexippus* and other *Danaus* species was high, genotypes within highly divergent regions were likely to be miscalled or filtered out. We therefore performed *de novo* assemblies for outgroup species for analyses that were sensitive to the quality of consensus sequence. Contigs were assembled for each outgroup sample, a

minimum length of 300 bp were kept and processed with a redundancy filter step as described previously[20].

We chose the highest quality assemblies (Chry_AUS_113_M, Eres_FL_27M, Erip_BRA_16005_F, and Gili_FL_28_F) to infer the species phylogeny based on 7,251 single-copy universal orthologs that were identified previously among lepidopteran genomes (*D. plexippus*[52], *H. melpomene*[53], and B. *mori*[54]). The OGS2.0 gene models of *D. plexippus* were used for homology search by TBLASTN. The high-scoring pairs with E<10-5 were then processed by genblastA v1.0.4[55] and gene structures were determined by GeneWise v2.2.0[56]. 3,714 proteins that were recovered with >= 50% coverage in all four outgroup species were kept, conserved blocks were extracted using Gblocks v0.91b[57], and were concatenated to seven super genes with 908,188 amino acids. The species phylogenetic tree was calculated using PhyML v3[58] with the JTT model and 100 replicates of bootstrap analyses. Our species tree resulted in clear separation among all sequenced *Danaus* species, including putative sister-species, *D. plexippus* and *D. erippus*.

Population structure

We used all bi-allelic and high quality SNPs to infer phylogeography and population structure for *D. plexippus*. For phylogeny, pairwise genetic distances were calculated among all samples as described previously[59] and a tree was subsequently generated (Fig. 1c) using the neighbor-joining (NJ) method implemented in PHYLIP v3.695[60]. A second frequency tree was also generated (Fig. 1d) based on 1000 bootstrap replicates using the consense module of PHYLIP.

We also inferred a population-level phylogeny using the maximum likelihood approach implemented in Treemix[61]. This method was designed specifically to infer patterns of population splitting events from genome-wide allele frequency data. For this analysis, we excluded samples that appeared to have recently dispersed among our geographic locations and we filtered out singleton SNP sites (MAF < 0.05). All subsequent data analysis was performed with Treemix v1.11 using parameters “-global” to generate the ML tree (Extended Data Fig. 1).

Population genetic structure and individual ancestry proportions (admixture) were inferred using FRAPPE v1.1[62]. We increased the pre-defined genetic clusters from *K*=2 to *K*=11 and ran analysis with 10,000 maximum iterations. We also performed principal component analysis (PCA) using the package EIGENSOFT v5.0[63]. A Tracey-Widom test was used to determine the significance level of the eigenvectors.

Based on the complete NJ tree, FRAPPE results, and PCA clustering, it was immediately clear that three *D. plexippus* samples had recently dispersed among geographic locations. Sample Plex_BLZ_4_M was collected in Belize but clustered with North American samples, Plex_FLs_MIA16_M was collected in south Florida but clustered with migratory North American samples, and Plex_MA_HI032_M was collected in Massachusetts but clustered with Bermuda. Since these are all putative exchanges among geographically proximate locations, we suspect they represent real dispersal events as opposed to sample mix-ups. It is also worth noting that Plex_FLs_MIA16_M is the only sample we included

from south Florida that was collected during the winter which suggests that North American migratory monarchs end up in south Florida where they are in contact with the genetically distinct non-migratory population[64]. This may also explain why this sample, and no other south Florida samples, emerged as a recent dispersal event. A small number of additional North American samples were population outliers in either FRAPPE or PCA suggesting potential admixture or contamination. We removed these samples, those with low sequence coverage, and the three recent dispersers from subsequent analyses (Supplementary Table 7). We note however that we performed a second analysis including all samples and found that we were still able to clearly identify all regions of the genome strongly associated with migration (below).

Demographic analysis

We compared patterns of linkage disequilibrium (LD) and minor allele frequency (MAF) among populations. To estimate linkage disequilibrium, we calculated r^2 using Haploview v4.2[65] with parameters "-maxdistance 160 -dprime -minGeno 0.6 -minMAF 0.1 -hwcutoff 0.001". Global patterns of LD and MAF were compared for the four main clusters of monarchs (Extended Data Fig. 2a and 2b). We found high LD and MAF in Atlantic and Pacific populations, consistent with inbreeding and population bottlenecks. Central/South America was intermediate between these populations and North America. We also estimated heterozygosity for each sample, calculated as the ratio of heterozygous to homozygous variants for each sample (Extended Data Fig. 2c).

We used the directionality index Ψ[26] to test the occurrence of a range expansion and to infer the origin of the range expansion (Extended Data Table 1). For this analysis, we used biallelic SNPs showing consensus genotypes across all outgroup individuals, and we defined the ancestral allele as that consensus outgroup allele. After excluding sites where one or both of the two focal populations (S1 and S2) was fixed for the ancestral allele, we calculated a 2D site derived allele frequency spectrum between populations as described previously[26].

We inferred demographic history for *D. plexippus* using the Pairwise Sequentially Markovian Coalescence (PSMC) model[27]. To ensure the quality of consensus sequence, we only used representative samples of high sequencing depth for each geographic region. Processed alignments of individuals were transformed to the whole-genome diploid consensus sequence using SAMTOOLS[66]. Bases of low sequencing depth (a third of the average depth) or high depth (twice of the average) were masked. We then used "fq2psmcfa" to transform the consensus sequence into a fasta-like format where the i-th character in the output sequence indicates whether there is at least one heterozygote in the bin of 20 bp. Parameters were set as follows: "-p 4+5*3+13*2+5*3+4 -r 2". The monarch generation time (g) was set as an estimate of 0.3 years. We used a standard mutation rate (μ) of 8.4×10^{-9}, from *Drosophila*[67]. Note, if we use a lower estimate of the mutation rate[68], all results from the PSMC and ∂a∂i analyses (below) remain qualitatively the same but inferred divergence times are older and effective population sizes are larger.

We also inferred the demographic history of the major geographic regions using diffusion approximation for demographic inference (∂a∂i[28]), which employs SNP frequency data for populations rather than recombination events within individual genomes. For this analysis,

we analyzed the North American population alone and then considered only pairwise comparisons between North America and each of the major dispersal populations (South/Central America, Atlantic, Pacific). In an attempt to avoid selected sites, we only used SNPs from intergenic regions on autosomal scaffolds. We calculated folded frequency spectra since there is no trinucleotide substitution matrix that can be used for statistical correction. As suggested, we specified simple models first and fit the model by increasing complexity gradually. A likelihood ratio test was used to optimize model selection, with the best model pictured in Extended Data Fig. 3a, although we note it is hard to rule out more complex demographic scenarios. Scaled parameters from the most likely model were transformed using the same g and μ as above. We also performed nonparametric bootstrapping (100 times) to determine the variance of each parameter (Extended Data Fig. 3).

Historical records suggest Atlantic and Pacific dispersal events of the monarch butterfly occurred in the 1800's[17,69]. These records are largely based on sightings from early European explorers who do not note the monarch in locations at certain times and then others who note the monarch in abundance only years later. Our demographic analyses based on genome sequence data are inconsistent with this timing. For instance, our PSMC analyses suggest Pacific and Atlantic dispersal events may have occurred as early as 2-3 ky ago. Because this timing was unexpected, we performed a follow-up analysis using *∂a∂i*[28] and this also yielded split times of 2-3 ky ago between North America and the Atlantic and Pacific populations (Extended Data Fig. 3). Interestingly, the *∂a∂i* analysis further indicated recent bottleneck recovery in the past 200-500 years, perhaps pointing to trans-oceanic dispersal events that were initially seeded thousands of years ago but which spread widely only within the last 200 years. This may provide some link between the genetic data and the historical records. If our demographic inference based on the genomic data is correct, where has the monarch been for all this time? The most common monarch host plants are recent introductions in the Pacific but there are native host plants in Southeast Asia. In addition, there is apparently a long history of the monarch butterfly in New Zealand. For instance, the indigenous Māori people of New Zealand believe the monarch is native to New Zealand, and unlike other Pacific locations, they have a traditional name for the monarch butterfly[69]. On the Atlantic side, it is possible the monarch has co-occurred with congener *D. chrysippus* in North Africa and on the Canary Islands, using the same host plants. We stress that the ancient Atlantic and Pacific dispersal scenarios we outline here are speculative, but plausible, and they would be in line with our genomic results.

Identification of migration-associated genomic regions

We applied a sliding window approach (5 kb windows sliding in 500 bp steps) to identify genomic regions associated with migration. Several statistical features were considered and compared. Based on the evolutionary scenario, we employed a modified population branch statistic (PBS) approach, which originally showed power to detect incomplete selective sweeps over short divergence times[29], a scenario that is highly relevant here.

Our approach was to search for genomic regions separating North America from all three independent, losses of migration (South/Central America, Atlantic, Pacific). Based on the original PBS algorithm, we specifically modified the formula as PBS = ($T^{N\text{-}C}$ + $T^{N\text{-}P}$ + $T^{N\text{-}A}$

- T^{C-P} - T^{C-A}) / 3, where T^{A-B} is the log transformed F_{ST} between population A and B (N, North America; C, Central/South America, P, Pacific; A, Atlantic). We further restricted this to windows in the lowest quartile distribution of pairwise nucleotide polymorphism (π[70]) within North America. At a significance of $P < 0.01$ (Z test), we identified a total of 5.14 Mb (2.1%) of the genome, including 536 predicted genes, which were associated with migration (Supplementary Table 8). If we did not restrict our gene set by low π in North America, our list of migration-associated genes included an additional 154 genes (Supplementary Table 8). We calculated a variety of other statistics, including Tajima's D[71], LD, difference in the number of ABBA vs. BABA[72] sites, and derived allele frequency (DAF) for each sliding window. To estimate DAF, we inferred ancestral alleles using the consensus sequence of the outgroup taxa. We found that DAF was notably enriched in our migration-associated genomic regions, relative to the rest of the genome, and this was true in both *D. plexippus* and *D. erippus*.

Annotation

We annotated genes in migration-associated genomic regions using the monarch OGS2.0 gene models and related information from MonarchBase[52] (Extended Data Table 2, Supplementary Table 8). We additionally annotated the genes within functional categories based on the corresponding *Drosophila melanogaster* orthologs using DAVID online platform v6.7[73]. Functional enrichments are presented in Supplementary Table 9 and Supplementary Table 10. We also specifically examined PBS and gene expression in glycolysis enzymes (Supplementary Table 11).

Targeted gene analysis

We performed a targeted population genetic analysis of three adjacent genes on scaffold DPSCF300190; FBX045, DPOGS206536, and collagen IV α-1. For each gene, CDS was extracted for samples from the population resequencing data and average pairwise sequence divergence and F_{ST} were estimated between migratory and non-migratory populations using Arlequin v3.5[74]. We used the program Network v4.612 (Fluxus Engineering) to generate gene networks and MEGA v6[75] to infer a maximum-likelihood gene tree for collagen IV α-1 under the GTR+I+G model. We used DnaSP v5.10[76] to compare sequence polymorphism among the three genes using the HKA test[77]. We also performed sliding window analyses of sequence polymorphism (π) and between species (*plexippus*, *chrysippus*) divergence (Dxy) along collagen IV α-1 CDS using DnaSP. Codon-based models of adaptive protein evolution were implemented using the Datamonkey webserver (datamonkey.org). Specifically, we ran FUBAR[78], MEME[79], and FEL[80] tests on an alignment of lepidopteran collagen IV α-1 sequences. The average dN/dS across collagen IV α-1 was 0.16 but all tests were suggestive of positive selection on the R1573Q substitution while other sites were found to be under negative or no selection (FUBAR: dN = 10.5, dS = 1, Normalized dN/dS = 9.5, Bayes factor = 87; MEME: dN(β^+) = 110 (P [$\beta = \beta+$] = 0.46), dS = 0.47, P = 0.08; FEL: dN = 55, dS = 10^{-6}, Normalized dN/dS = 83, P = 0.10).

RNA-seq

We extracted total RNA from the adult thoracic muscle tissue of six *D. plexippus* samples, one male and one female from Massachusetts, Hawaii, and Costa Rica. The samples from Hawaii and Costa Rica were from non-migratory populations and the samples from Massachusetts were from a non-migratory summer generation of the migratory North American population. RNA-seq libraries were prepared using an Illumina TruSeq protocol, individually indexed and sequenced on one lane of HiSeq 2000. After QC processing of raw data, differential gene expression of target genes was analyzed using TopHat v2.0.7[81] and CuffLinks v2.1.1[82].

Respirometry

We measured resting and flight metabolic rates using flow-through respirometry[36]. The material consisted of 60, 2-day old females, originating from Massachusetts (n=19), south Florida (n=19), and Costa Rica (n=22). The samples from south Florida and Costa Rica were from non-migratory populations and the samples from Massachusetts were from a non-migratory summer generation of the migratory North American population. Individuals were placed in a 2-L cylindrical, transparent respirometry chamber and covered with a dark cloth. We pumped dried, CO_2-free air through the chamber and used a mass flow meter and controller (Sierra Instruments, Monterey, CA, USA) to regulate the STP-corrected flow rate of 1.8 L min^{-1}. We used a regularly calibrated differential, infrared CO_2 analyzer (Li-Cor 7000, Li-Cor, Lincoln, NE, USA) to measure the CO_2 emission rate. We converted the signals from analog to digital using a Sable Systems Universal Interface (UI-2) and collected the data using ExpeData (Sable Systems, Reno, NV, USA). After the individual had rested motionless in the darkened chamber for ca. 15 min and the CO_2 emission curve had reached a stable baseline, we started recording resting metabolic rate (RMR). We recorded a minimum of 10 min of stable, undisturbed RMR. The measurements took place in a temperature controlled room. We used a NTC thermistor probe (Sable Systems) to continuously measure the temperature inside the chamber. The mean temperature across the RMR measurements was 32.2°C (s.e.m. 0.07).

After the RMR recording, we removed the dark cloth and exposed the butterfly to bright light (two 25 W UV/visible light bulbs and one 26 fluorescent light bulb). We allowed the butterfly to adjust to the light for 30 sec after which we began to stimulate it to fly by sharply shaking the chamber after which the individual took off. Once it alighted, we shook it up in the air again. The shaking continued for 10 min with the aim of producing continuous flight. After 10 min, we stopped the stimulation and covered the chamber with the dark cloth. We allowed the CO_2 curve to return to the baseline. We performed an instantaneous or Z-correction[83] on all metabolic rate data to compensate for delayed washout of CO_2 in the respirometry chamber. To characterize flight performance, we focused on total flight metabolic rate (FMR), which consists of the total volume of CO_2 produced during the 10-min flight experiment.

For analysis, we compared the three populations using ANCOVA. Time of the day, temperature, and body size (pupal mass) were added as covariates to the models. There were significant differences in RMR among the three populations ($F_{2,51} = 6.11$, $P = 0.004$). Pupal

mass had a positive effect on RMR ($F_{1,51}$ = 14.98, P = 0.0003). There was a marginally significant quadratic time effect, suggesting that RMR peaked in early afternoon (Time: $F_{1,51}$ = 4.09, P = 0.048; Time^2: $F_{1,51}$ = 4.03, P = 0.0499). The measurement temperature main effect was non-significant ($F_{1,51}$ = 1.08, P = 0.30) but there was a significant interaction between population and temperature ($F_{1,51}$ = 6.30, P = 0.004), suggesting a positive relationship between temperature and RMR in Florida and Costa Rica, but a negative relationship in Massachusetts. A Tukey's HSD test revealed no significant RMR differences in pairwise comparisons among populations (P > 0.05 in all pairwise comparisons). FMR differed among populations ($F_{2,56}$ = 6.43, P = 0.003). Pupal mass had a positive effect on FMR ($F_{1,56}$ = 9.38, P = 0.0034). A Tukey's HSD test showed that mass-independent FMR was different between Massachusetts and south Florida (P = 0.0025) but not between Massachusetts and Costa Rica (P = 0.5837).

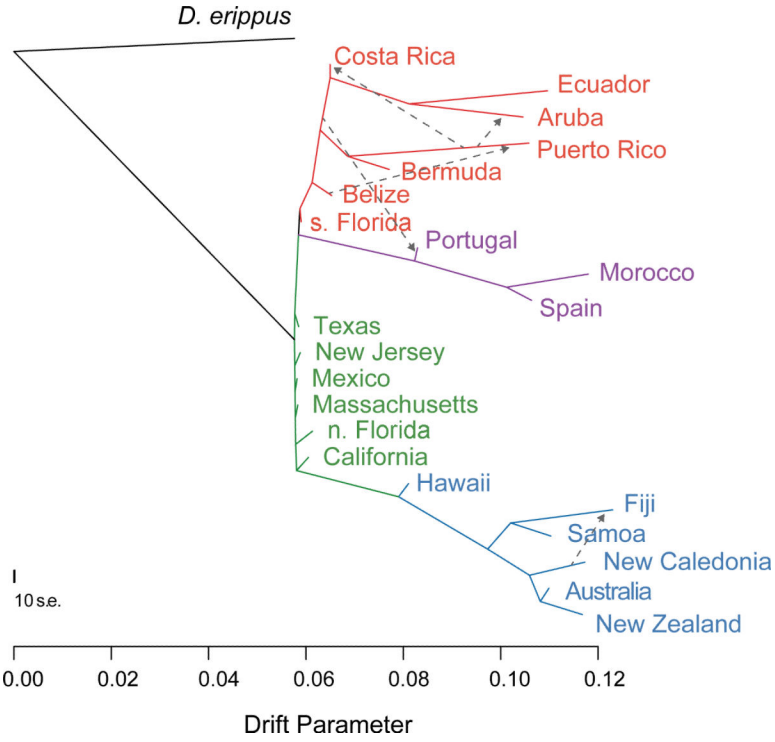
Association mapping of color gene

We extracted genomic DNA from 12 individuals (Supplementary Table 1) and constructed Illumina paired-end libraries using the Illumina TruSeq protocol. 12 libraries were indexed and pooled into 2 lanes and sequenced using Illumina Hiseq2000. Only high quality reads that passed QC step were used for downstream analyses. Genome resequencing data were aligned to *D. plexippus* reference genome sequence using Bowtie2 v2.1.0[84] with parameter –very-sensitive-local and then were re-ordered and sorted by Picard v1.84 (http://picard.sourceforge.net). Realigner Target Creator and Indel Realigner in GATK v2.1 were used to realign indels and Unified Genotyper was used to call genotypes across 12 individuals using following parameters: heterozygosity 0.01, stand_call_conf 50, stand_emit_conf 10, dcov 250. 10,034,303 SNPs and 1,434,642 indels supported by more than 10 individuals and with good quality (Q>30) were kept for further analysis. Association tests were performed using PLINK v1.07[85] and variants with P < 0.005 (Fisher's exact test) were selected. Candidate loci were checked using customized scripts and those with strong linkage patterns (containing >10 associated variants) within gene regions and satisfied the known genotypes were picked up for PCR verification (Supplementary Table 12). We also repeated this analysis after excluding 4 of the 12 samples with lower genome sequence coverage and the results were identical, yielding the highest genome-wide SNP associations in the myosin gene DPOGS206617.

We further tested potentially associated polymorphisms in families and field-collected samples provided by John Stimson[41,42]. We extracted genomic DNA from 58 historic specimens (Supplementary Table 13) and performed whole genome amplification using Genome Plex Complete Whole Genome Amplification Kit (Sigma). We designed primers to span polymorphisms on five potentially associated scaffolds (Supplementary Table 12). Amplification efficiency was low because of the age of these specimens. However, after Sanger sequencing all positive PCR products and subsequent analysis of genotypes, amplified region 1013900-1013989 yielded a very strongly associated SNP, position 785 in the myosin gene DPOGS206617 (Supplementary Table 12 and Supplementary Table 14). This section of DPOGS206617 is very likely to house the causative variation responsible for monarch color variation because it contains SNPs perfectly associated with color in our full-genome sequence data, a SNP very strongly associated in our targeted analysis of historical

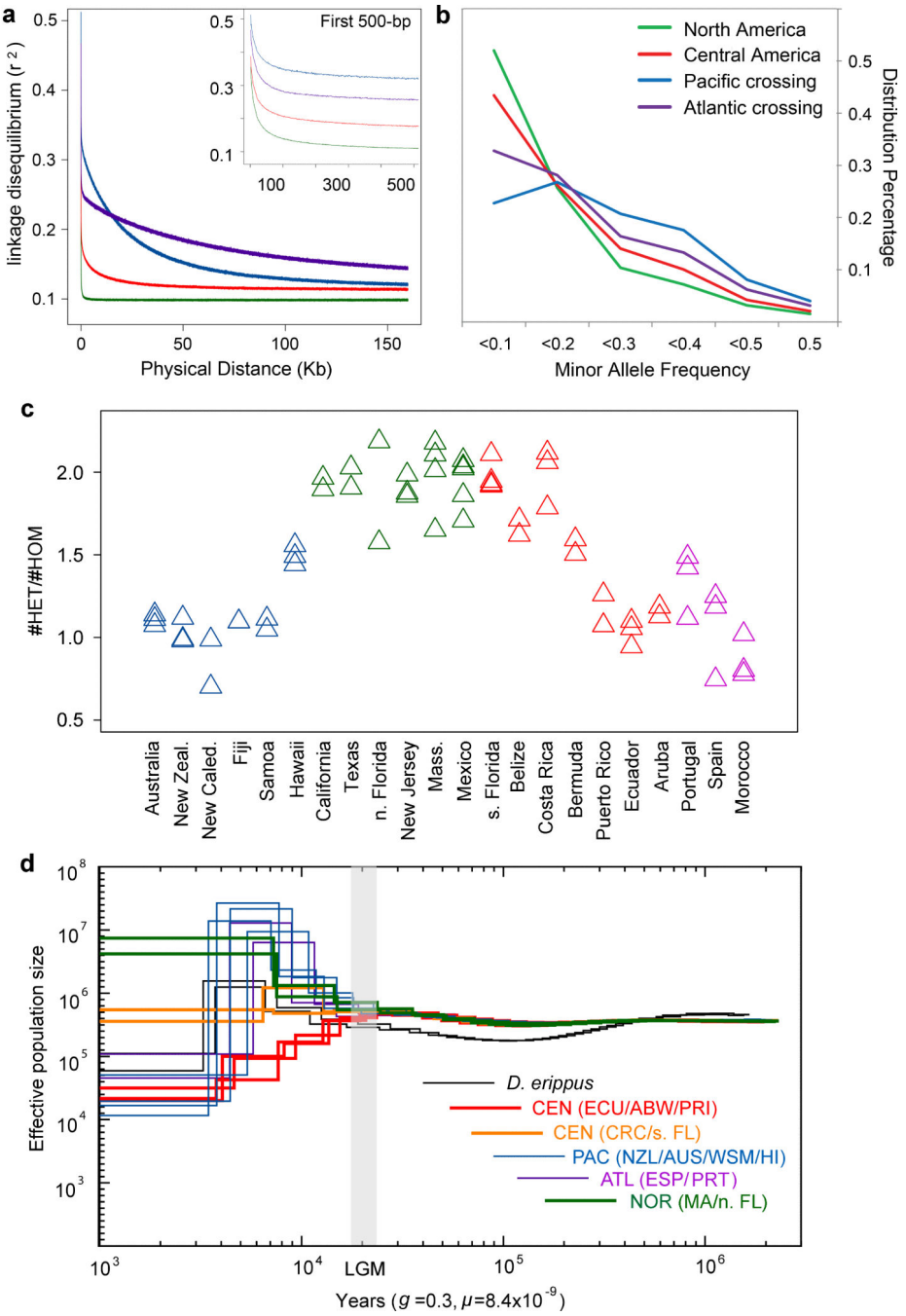
specimens, a signature of long-term purifying selection over evolutionary time, and notably, the white-associated alleles in this region are derived alleles that are found in no other monarch sample among our 101 sequencing panel.

Extended Data



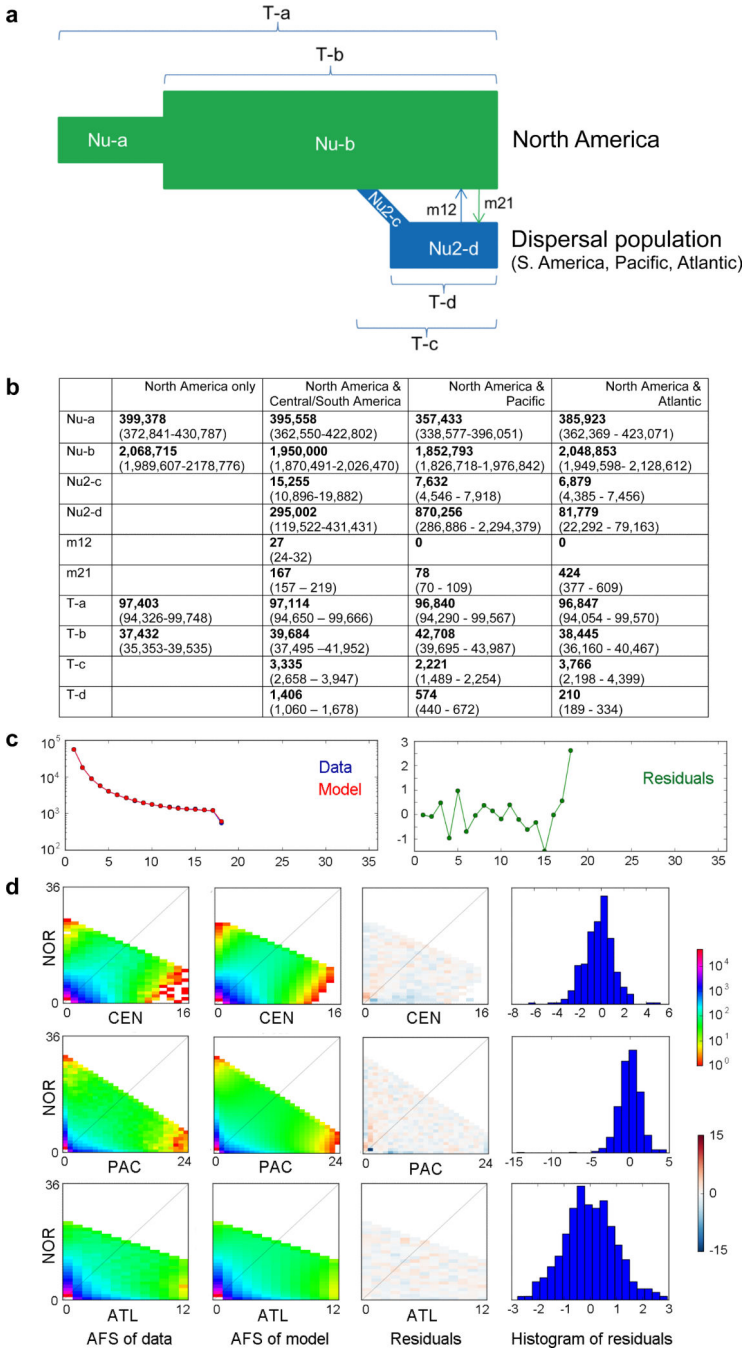
Extended Data Figure 1. Relationships among monarch populations inferred using the maximum likelihood method implemented in Treemix

Note, this is a fully resolved, bifurcating tree. The very short basal branches indicate little genetic drift in North American populations, not unresolved basal relationships. Colors correspond to those in Fig. 1. Treemix also inferred five migration events among populations: from Puerto Rico to Aruba, from Puerto Rico to Costa Rica, from New Caledonia to Fiji, from Belize or Costa Rica to Portugal, and from Belize to Puerto Rico.



Extended Data Figure 2. Demographic history of the monarch butterfly

a, Patterns of linkage-disequilibrium decay across the genome in different geographic populations. **b**, Genome-wide distribution of minor allele frequencies. **c**, Heterozygosity across populations, estimated as the ratio of heterozygous SNPs to homozygous SNPs/ individual. **d**, Demographic history inferred using PSMC. This analysis includes representative individuals of high sequencing depth for each geographic location. The period of the last glacial maximum (LGM; ~20 ky ago) is shaded in gray.



	North America only	North America & Central/South America	North America & Pacific	North America & Atlantic
Nu-a	**399,378** (372,841-430,787)	**395,558** (362,550-422,802)	**357,433** (338,577-396,051)	**385,923** (362,369 - 423,071)
Nu-b	**2,068,715** (1,989,607-2178,776)	**1,950,000** (1,870,491-2,026,470)	**1,852,793** (1,826,718-1,976,842)	**2,048,853** (1,949,598- 2,128,612)
Nu2-c		**15,255** (10,896-19,882)	**7,632** (4,546 - 7,918)	**6,879** (4,385 - 7,456)
Nu2-d		**295,002** (119,522-431,431)	**870,256** (286,886 - 2,294,379)	**81,779** (22,292 - 79,163)
m12		**27** (24-32)	**0**	**0**
m21		**167** (157 – 219)	**78** (70 - 109)	**424** (377 - 609)
T-a	**97,403** (94,326-99,748)	**97,114** (94,650 – 99,666)	**96,840** (94,290 - 99,567)	**96,847** (94,054 - 99,570)
T-b	**37,432** (35,353-39,535)	**39,684** (37,495 –41,952)	**42,708** (39,695 - 43,987)	**38,445** (36,160 - 40,467)
T-c		**3,335** (2,658 – 3,947)	**2,221** (1,489 - 2,254)	**3,766** (2,198 - 4,399)
T-d		**1,406** (1,060 – 1,678)	**574** (440 - 672)	**210** (189 - 334)

Extended Data Figure 3. ∂a∂i analysis parameter estimates

a, schematic of demographic scenario modeled in ∂a∂i labeled with parameters being estimated. Nu, effective population size (individuals); m, migration rate (individuals/year); T, time (years). **b**, inferred parameter estimates. **c**, 1D model-data comparison considering North America population only. In the left panel, the model is plotted in red and the data in blue. In the right panel, the residuals between model and data are plotted. **d**, 2D comparison for joint estimation of North America and dispersal populations (Central/South America,

Pacific, Atlantic). The left two panels are marginal spectra for data and the maximum-likelihood model, respectively. The right two panels show the residuals.

Extended Data Table 1

Inferring the monarch range expansion.

S1	S2	F_{ST}	Ψ
North America	south Florida	0.0472	−0.4219
	Caribbean	0.0960	−0.8639
	Central America	0.0473	−0.4765
	South America	0.1263	−1.0493
	Pacific	0.0929	−1.4988
	Atlantic	0.1054	−1.0268

A positive Ψ indicates that population S1 is farther away from the origin of the expansion than S2 while a negative value indicates S2 is farther from the origin of the range expansion. North America (S1) includes the United States and Mexico but excludes south Florida; Caribbean includes Bermuda and Puerto Rico; Central America includes Belize and Costa Rica; South America includes Aruba and Ecuador; Pacific includes Hawaii, Samoa, Fiji, New Caledonia, Australia and New Zealand; Atlantic includes Portugal, Spain and Morocco.

Extended Data Table 2

Top 20 migration-associated genomic regions.

Scaffold	Position (Kb)	PBS	Involved genes
DPSCF300190	186.5-207.5	0.406	– FBX045 – Transmembrane protein – Collagen alpha-1 (IV)
DPSCF300134	1-61.5	0.086	– IGF-II mRNA-binding protein – Insulin-like growth factor 2 – 2 monarch hypothetical proteins
DPSCF300190	158.5-186	0.081	– dipeptidyl-peptidase – 2 WD-40 transcription factors
DPSCF300001	3799.5-3811.5	0.080	– Acyltransferase 3
DPSCF300005	105.5-119	0.067	– Lepidopteran hypothetical protein
DPSCF300001	4632.5-4644	0.064	– Kettin – pleiotrophin-like protein
DPSCF300014	323.5-337	0.063	– RNA-binding protein – Pre-mRNA-splicing factor Cwf15 – 2 universal hypothetical protein
DPSCF300551	22.5-30.3	0.060	– butterfly hypothetical protein
DPSCF300134	135.5-172.5	0.059	– insect hypothetical protein
DPSCF300001	4697.5-4709	0.053	
DPSCF300190	213-236.5	0.052	– Collagen alpha-2(IV) – thioredoxin family Trp26 – Selenoprotein T – Tetratricopeptide-like helical
DPSCF300001	4622.5-4628.5	0.052	– kettin protein
DPSCF300134	109.5-134.5	0.052	– potassium ion transport protein
DPSCF300074	526.5-532	0.052	
DPSCF300005	137-152	0.049	– WD40 transcription factor
DPSCF300255	244-257.8	0.048	
DPSCF300014	176-202.5	0.047	– Golgin-80 (RabGTPase binding) – Phosphatidylserine synthase – WD40 protein – kismet DNA binding protein
DPSCF300083	396-431	0.047	– Zinc finger DNA-binding domain
DPSCF300001	4725–4751	0.046	– forkhead protein transcription factor – Cyclic ion channel subunit
DPSCF300005	223.5-264	0.045	– Flotillin-1 (insulin-signaling pathway)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank B. Ballister, S. Barribeau, R. Bartel, N. Chamberlain, R. Cook, A. Davis, D. Feary, D. Frey, M. Maudsley, G. Moreira, E. Osburn, R. Rarick, E. Rendon, D. Rodrigues, E. Sternberg and J. Stimson for assistance collecting or providing specimens. We also thank J. Jensen and D. Lohman for discussion. This work was supported by National Institutes of Health grant GM086794-02S1, National Science Foundation grants

IOS-0923411, DEB-0643831, DEB-1019746, and DEB-1316037, and Neubauer Funds from the University of Chicago.

References

1. Dingle H, Zalucki MP, Rochester WA, Armijo-Prewitt T. Distribution of the monarch butterfly, *Danaus plexippus* (L.) (Lepidoptera : Nymphalidae), in western North America. Biol. J. Linn. Soc. 2005; 85:491–500.
2. Lyons JI, et al. Lack of genetic differentiation between monarch butterflies with divergent migration destinations. Mol. Ecol. 2012; 21:3433–3444. [PubMed: 22574833]
3. Malcolm, SB.; Zalucki, MP. Biology and Conservation of the Monarch Butterfly. Natural History Museum of LA County; 1993.
4. Oberhauser, KS.; Solensky, MJ. The Monarch Butterfly: Biology and Conservation. Cornell University Press; 2004.
5. Urquhart, FA. Found at last; the monarch's winter home. Vol. 150. National Geographic Magazine; 1976. p. 161-173.
6. Urquhart FA, Urquhart NR. 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. 1978; 56:1759–1764.
7. Wassenaar LI, Hobson KA. Natal origins of migratory monarch butterflies at wintering colonies in Mexico: new isotopic evidence. Proc. Natl. Acad. Sci. USA. 1998; 95:15436–15439. [PubMed: 9860986]
8. Froy O, Gotter AL, Casselman AL, Reppert SM. Illuminating the circadian clock in monarch butterfly migration. Science. 2003; 300:1303–1305. [PubMed: 12764200]
9. Heinze S, Reppert SM. Sun compass integration of skylight cues in migratory monarch butterflies. Neuron. 2011; 69:345–358. [PubMed: 21262471]
10. Merlin C, Gegear RJ, Reppert SM. Antennal circadian clocks coordinate sun compass orientation in migratory monarch butterflies. Science. 2009; 325:1700–1704. [PubMed: 19779201]
11. Reppert SM, Gegear RJ, Merlin C. Navigational mechanisms of migrating monarch butterflies. Trends in neurosciences. 2010; 33:399–406. [PubMed: 20627420]
12. Sauman I, et al. Connecting the navigational clock to sun compass input in monarch butterfly brain. Neuron. 2005; 46:457–467. [PubMed: 15882645]
13. Mouritsen H, Frost BJ. Virtual migration in tethered flying monarch butterflies reveals their orientation mechanisms. Proc. Natl. Acad. Sci. USA. 2002; 99:10162–10166. [PubMed: 12107283]
14. Ackery, PR.; Vane-Wright, RI. Milkweed Butterflies: Their Cladistics and Biology. British Museum; 1984.
15. Altizer S, Davis AK. Populations of monarch butterflies with different migratory behaviors show divergence in wing morphology. Evolution. 2010; 64:1018–1028. [PubMed: 20067519]
16. Dockx C. Directional and stabilizing selection on wing size and shape in migrant and resident monarch butterflies, *Danaus plexippus* (L.), in Cuba. Biol. J. Linn. Soc. 2007; 92:605–616.
17. Vane-Wright, RI. Biology and Conservation of the Monarch Butterfly. In: Malcolm, SB.; Zalucki, MP., editors. Natural History Museum of LA; 1993. p. 179-187.
18. Haeger JF, Jordano D. The Monarch butterfly *Danaus plexippus* (Linnaeus, 1758) in the Strait of Gibraltar (Lepidoptera: Danaidae). SHILAP: Revista de Lepidopterologia. 2009; 37:421–438.
19. Zhu H, Casselman A, Reppert SM. Chasing migration genes: a brain expressed sequence tag resource for summer and migratory monarch butterflies (Danaus plexippus). PloS One. 2008; 3:e1345. [PubMed: 18183285]
20. Zhan S, Merlin C, Boore JL, Reppert SM. The monarch butterfly genome yields insights into long-distance migration. Cell. 2011; 147:1171–1185. [PubMed: 22118469]
21. Kitching, IJ.; Ackery, PR.; Vane-Wright, RI. Biology and Conservation of the Monarch Butterfly. Malcolm, SB.; Zalucki, MP., editors. Natural History Museum of LA; 1993. p. 11-16.

22. Gauthreaux, SA. Avian Biology. Farner, DS.; King, JR.; Parkes, KC., editors. Vol. 4. Elsevier; 1982. p. 93-168.Ch. 2
23. Zalucki MP, Clarke AR. Monarchs across the Pacific: the Columbus hypothesis revisited. Biol. J. Linn. Soc. 2004; 82:111–121.
24. Brower LP, Oberhauser KS, Boppré M, Brower AVZ, Vane-Wright RI. Monarch sex: ancient rites, or recent wrongs? Antenna. 2007; 31
25. Pierce AA, et al. Serial founder effects and genetic differentiation during worldwide range expansion of monarch butterflies. Proc. R. Soc. B. submitted.
26. Peter BM, Slatkin M. Detecting range expansions from genetic data. Evolution. 2013; 67:3274–3289. [PubMed: 24152007]
27. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; 475:493–496. [PubMed: 21753753]
28. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009; 5:e1000695. [PubMed: 19851460]
29. Yi X, et al. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science. 2010; 329:75–78. [PubMed: 20595611]
30. Green RE, et al. A Draft Sequence of the Neandertal Genome. Science. 2010; 328:710–722. [PubMed: 20448178]
31. Schnorrer F, et al. Systematic genetic analysis of muscle morphogenesis and function in *Drosophila*. Nature. 2010; 464:287–291. [PubMed: 20220848]
32. Kelemen-Valkony I, et al. *Drosophila* basement membrane collagen col4a1 mutations cause severe myopathy. Matrix Biol. 2012; 31:29–37. [PubMed: 22037604]
33. Plaisier E, et al. COL4A1 mutations and hereditary angiopathy, nephropathy, aneurysms, and muscle cramps. New Engl. J. Med. 2007; 357:2687–2695. [PubMed: 18160688]
34. Hakeda S, Endo S, Saigo K. Requirements of Kettin, a giant muscle protein highly conserved in overall structure in evolution, for normal muscle function, viability, and flight activity of *Drosophila*. J. Cell Biol. 2000; 148:101–114. [PubMed: 10629221]
35. Gibo DL, Pallett MJ. Soaring flight of monarch butterflies, *Danaus plexippus* (Lepidoptera: Danaidae), during the late summer migration in southern Ontario. Can. J. Zool. 1979; 57:1393–1401.
36. Niitepold K, et al. Flight metabolic rate and Pgi genotype influence butterfly dispersal rate in the field. Ecology. 2009; 90:2223–2232. [PubMed: 19739384]
37. Mitikka V, Hanski I. Pgi genotype influences flight metabolism at the expanding range margin of the European map butterfly. Ann. Zool. Fenn. 2010; 47:1–14.
38. Niitepold K, Mattila ALK, Harrison PJ, Hanski I. Flight metabolic rate has contrasting effects on dispersal in the two sexes of the Glanville fritillary butterfly. Oecologia. 2011; 165:847–854. [PubMed: 21190042]
39. Reichstein T, von Euw J, Parsons JA, Rothschild M. Heart poisons in the monarch butterfly. Science. 1968; 161:861–866. [PubMed: 4875496]
40. Ritland DB, Brower LP. The viceroy butterfly is not a batesian mimic. Nature. 1991; 350:497–498.
41. Stimson J, Kasuya M. Decline in the frequency of the white morph of the monarch butterfly (*Danaus plexippus* plexippus L., Nymphalidae) on Oahu, Hawaii. J. Lep. Soc. 2000; 54:29–32.
42. Stimson JS, Meyers L. Inheritance and frequency of a color polymorphism in *Danaus plexippus* (Lepidoptera: Danaidae) on Oahu, Hawaii. J. Res. Lep. 1984; 23:153–160.
43. Nijhout, HF. The Development and Evolution of Butterfly Wing Patterns. Smithsonian Press; 1991.
44. Mercer JA, Seperack PK, Strobel MC, Copeland NG, Jenkins NA. Novel myosin heavy chain encoded by murine dilute coat colour locus. Nature. 1991; 349:709–713. [PubMed: 1996138]
45. Fukuda M, Kuroda TS. Missense mutations in the globular tail of myosin-Va in dilute mice partially impair binding of Slac2-a/melanophilin. J. Cell Sci. 2004; 117:583–591. [PubMed: 14730011]

46. Rendón-Salinas, E.; Tavera-Alonso, G. Forest surface occupied by monarch butterfly hibernation colonies in December 2013. World Wildlife Fund-México; 2014.
47. Haber, WA. Biology and Conservation of the Monarch Butterfly. Malcolm, SB.; Zalucki, MP., editors. Natural History Museum of LA County; 1993. p. 201-207.
48. James, DG. Biology and Conservation of the Monarch Butterfly. Malcolm, SB.; Zalucki, MP., editors. Natural History Museum of LA County; 1993. p. 189-200.
49. Smith DAS, Owen DF. Colour genes as markers for migratory activity: The butterfly *Danaus chrysippus* in Africa. Oikos. 1997; 78:127–135.
50. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011; 21:936–939. [PubMed: 20980556]
51. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 2011; 43:491–498. [PubMed: 21478889]
52. Zhan S, Reppert SM. MonarchBase: the monarch butterfly genome database. Nucleic acids research. 2013; 41:D758–763. doi:10.1093/nar/gks1057. [PubMed: 23143105]
53. Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 2012; 487:94–98. [PubMed: 22722851]
54. International Silkworm Genome Consortium. The genome of a lepidopteran model insect, the silkworm Bombyx mori. Insect Biochem. Mol Biol. 2008; 38:1036–1045. [PubMed: 19121390]
55. She R, Chu JS, Wang K, Pei J, Chen N. GenBlastA: enabling BLAST to identify homologous gene sequences. Genome Res. 2009; 19:143–149. [PubMed: 18838612]
56. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004; 14:988–995. [PubMed: 15123596]
57. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56
58. Guindon S, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 2010; 59:307–321. [PubMed: 20525638]
59. Xia Q, et al. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). Science. 2009; 326:433–436. [PubMed: 19713493]
60. PHYLIP (phylogeny inference package) v. 3.6. 2005
61. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012; 8:e1002967. [PubMed: 23166502]
62. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: Analytical and study design considerations. Genet. Epidemiol. 2005; 28:289–301. [PubMed: 15712363]
63. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]
64. Knight A, Brower LP. The influence of eastern North American autumnal migrant monarch butterflies (*Danaus plexippus* L.) on continuously breeding resident monarch populations in southern Florida. J. Chem. Ecol. 2009; 35:816–823. [PubMed: 19579046]
65. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21:263–265. [PubMed: 15297300]
66. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
67. Haag-Liautard C, et al. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. Nature. 2007; 445:82–85. [PubMed: 17203060]
68. Keightley PD, et al. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. 2009; 19:1195–1201. [PubMed: 19439516]
69. Zalucki MP, Clarke AR. Monarchs across the Pacific: the Columbus hypothesis revisited. Biol. J. Linn. Soc. 2004; 82:111–121.
70. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA. 1979; 76:5269–5273. [PubMed: 291943]
71. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123:585–595. [PubMed: 2513255]

72. Durand EY, Patterson N, Reich D, Slatkin M. Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. 2011; 28:2239–2252. [PubMed: 21325092]
73. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 2009; 4:44–57. [PubMed: 19131956]
74. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 2010; 10:564–567. [PubMed: 21565059]
75. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol. Biol. Evol. 2013; 30:2725–2729. [PubMed: 24132122]
76. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25:1451–1452. [PubMed: 19346325]
77. Hudson RR, Kreitman M, Aguade M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987; 116:153–159. [PubMed: 3110004]
78. Murrell B, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol. Biol. Evol. 2013; 30:1196–1205. [PubMed: 23420840]
79. Murrell B, et al. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012; 8:e1002764. [PubMed: 22807683]
80. Pond SLK, Frost SDW. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 2005; 22:1208–1222. [PubMed: 15703242]
81. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
82. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotech. 2010; 28:511–515.
83. Bartholomew GA, Vleck D, Vleck CM. Instantaneous measurements of oxygen consumption during pre-flight warm-up and post-flight cooling in sphingid and saturniid moths. J Exp Biol. 1981; 90:17–32.
84. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 2012; 9:357–359. [PubMed: 22388286]
85. Purcell S, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007; 81:559–575. [PubMed: 17701901]

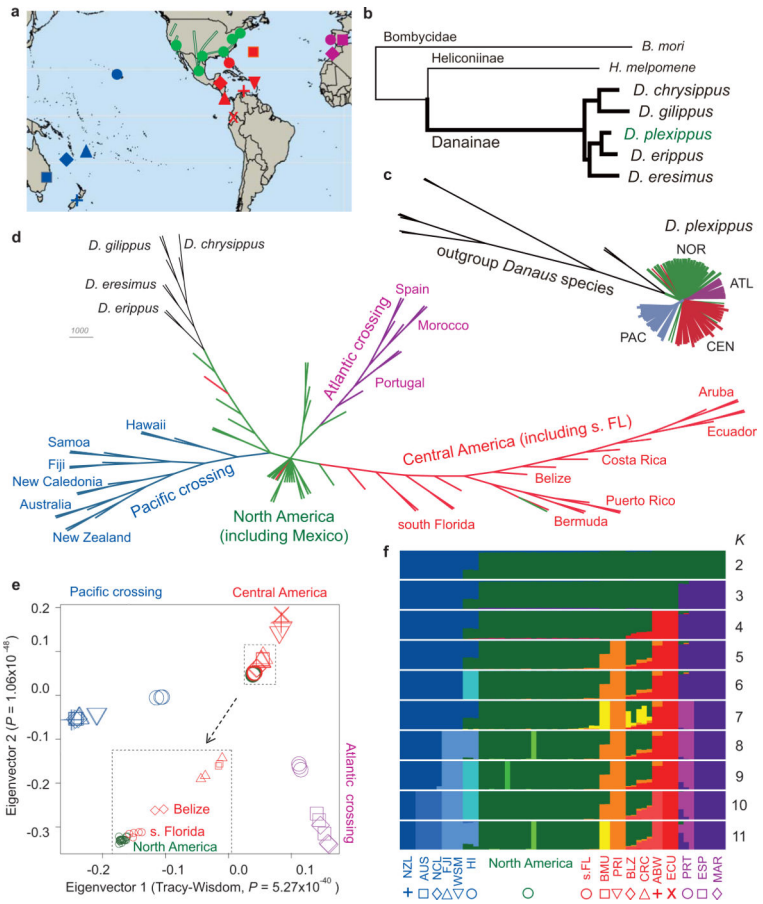


Figure 1. Global dispersal of the monarch butterfly

a, Monarch butterfly sampling locations. **b**, Inferred phylogeny among *Danaus* species based on maximum likelihood analysis of 3,714 single-copy genes. **c**, Neighbor-joining phylogeny of all *D. plexippus* individuals, based on genome-wide SNP data. **d**, NJ consensus tree based on 1,000 bootstrap replicates. **e**, PCA plots based on the first two principal components; inset shows separation between North America and south Florida. **f**, Genetic structure and individual ancestry; colors in each column represent ancestry proportion over range of population sizes K = 2-11.

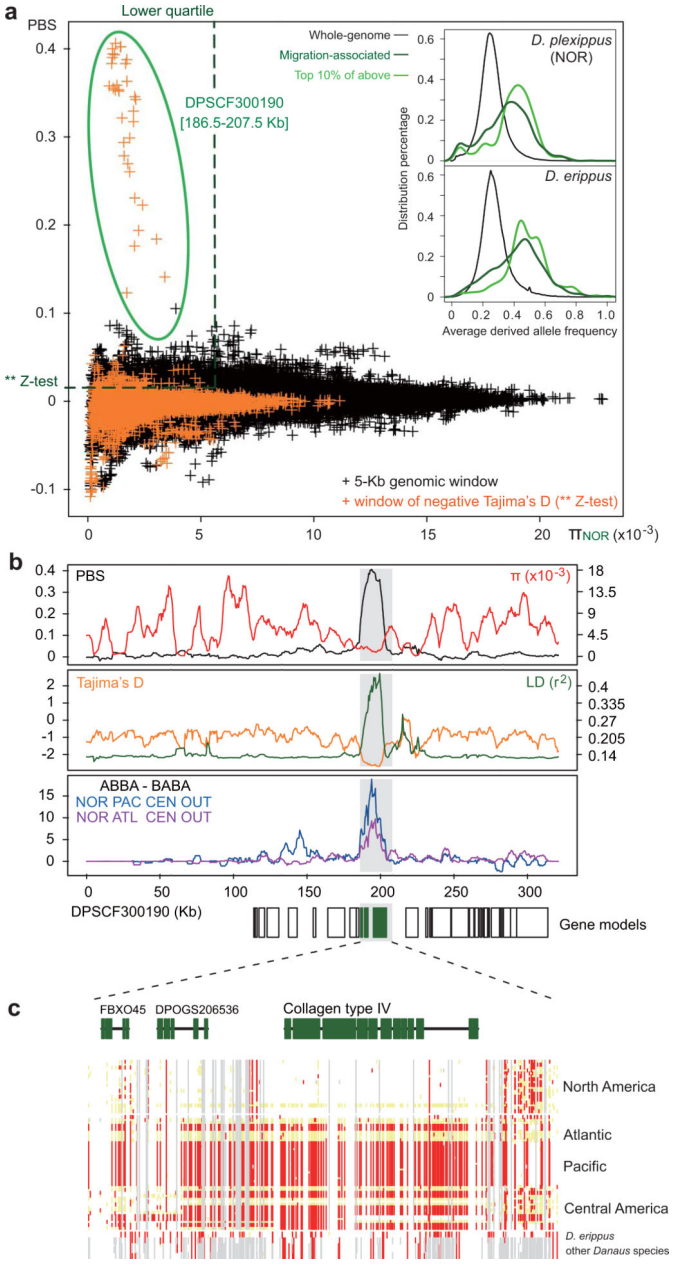


Figure 2. A selective sweep associated with migration

a, Distribution of PBS and polymorphism in North America (π_{NOR}), calculated in 5 kb sliding windows. Migration-associated genomic regions were identified as the points above the dashed line ($P < 0.01$) and to the left of the vertical green dashed line (lower quartile). Circled points consist of a single 21 kb region. **b**, Population genetic statistics were plotted across DPSCF300190 in 5 kb sliding windows. **c**, Gene models and SNP allele: white represents homozygous for the reference allele; red, homozygous for alternative allele; yellow, heterozygous; gray, masked site.

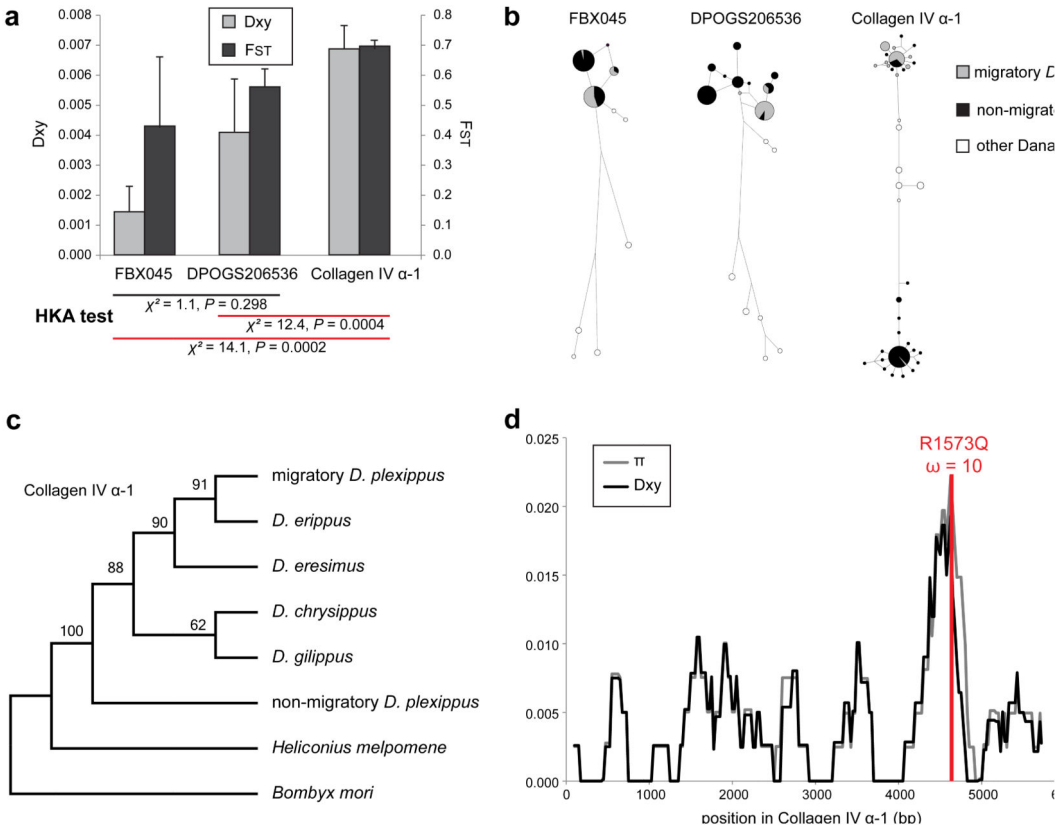


Figure 3. Divergent selection on collagen IV α-1

a, Collagen IV α-1 shows elevated sequence divergence (Dxy) and differentiation (F_{ST}) between migratory and non-migratory monarchs (mean ± s.e.m), an excess of polymorphism (HKA test), and **b**, haplotype divergence. **c**, A maximum-likelihood tree shows the non-migratory haplotype predates species-level divergence within *Danaus* while the migratory haplotype is similar to *D. erippus*. **d**, A subsection of high polymorphism and divergence in collagen IV α-1 coincides with an amino acid experiencing positive selection, including a R1573Q substitution on the migratory haplotype. **e**, expression of collagen IV α-1 and α-2 differ between migratory and non-migratory populations in flight muscle tissue. **f**, flight metabolic rates differ more than resting metabolic rates between migratory and non-migratory populations (mean ± s.e.m.).

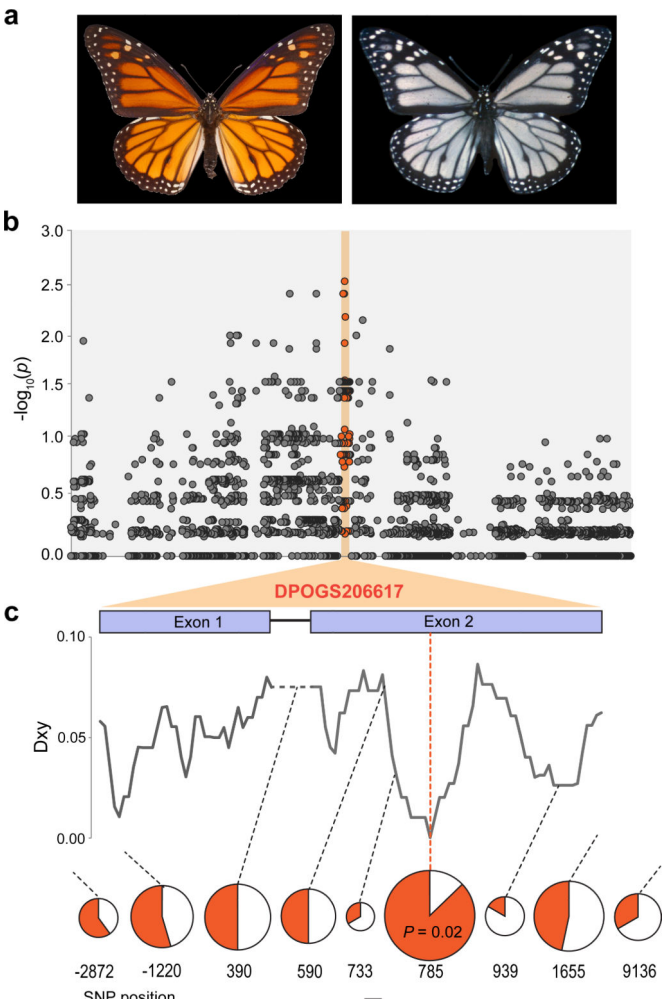


Figure 4. The genetic basis of warning coloration

a, While generally bright orange, the *nivosus* morph lacks orange pigmentation. **b**, A comparison of 12 Hawaiian monarch genome sequences (5 wild-type, 5 *nivosus*, and 2 F1 hybrids) reveals perfect SNP associations in one gene, the myosin gene DPOGS206617. **c**, Comparison of DNA sequence divergence (Dxy) between *D. plexippus* and *D. chrysippus* shows strong purifying selection in exon 2, coinciding with SNP associations in modern samples, crosses, and field collections from the 1980's. SNP position 785 is associated in 17/20 samples from the 1980's (P = 0.02, one-tailed Fisher's exact test).