Article

Current Biology

Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years

Highlights

- A new, chromosome-level assembly and annotation is provided for common milkweed
- Milkweed forms an enormous, panmictic population across its North American range
- Monarchs and milkweeds both had two concurrent population expansions in the past
- Neither species shows signs of recent declines in effective population size

Authors

John H. Boyle, Susan Strickler, Alex D. Twyford, ..., Georg Jander, Anurag A. Agrawal, Joshua R. Puzey

Correspondence

jrpuzey@wm.edu

In brief

Boyle et al. find correlated population histories for the monarch butterfly and its common milkweed food plant over the past 25,000 years. Using genomic data and a new, chromosome-level milkweed assembly, they find expansions in both species corresponding to the recession of glaciers and to the more recent clearing of North American forests.

Boyle et al., 2023, Current Biology *33*, 1–9
September 11, 2023
https://doi.org/10.1016/j.cub.2023.07.057

CellPress

Current Biology

CellPress

Article

Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years

John H. Boyle,[1,2] Susan Strickler,[3,4,5] Alex D. Twyford,[6,7] Angela Ricono,[1] Adrian Powell,[3] Jing Zhang,[3] Hongxing Xu,[3,8] Ronald Smith,[9] Harmony J. Dalgleish,[1] Georg Jander,[3] Anurag A. Agrawal,[10] and Joshua R. Puzey[1,11,*]

[1]Biology Department, College of William & Mary, 540 Landrum Dr., Williamsburg, VA 23185, USA
[2]Biology Department, University of Mary, 7500 University Dr., Bismarck, ND 58504, USA
[3]Boyce Thompson Institute, 533 Tower Rd., Ithaca, NY 14853, USA
[4]Chicago Botanic Garden, Plant Science and Conservation, 1000 Lake Cook Rd., Glencoe, IL 60022, USA
[5]Northwestern University, Plant Biology and Conservation Program, 2145 Sheridan Rd., Evanston, IL 60208, USA
[6]Institute of Ecology and Evolution, University of Edinburgh, Charlotte Auerbach Rd., Edinburgh EH9 3FL, UK
[7]Royal Botanic Garden Edinburgh, Edinburgh EH3 5NZ, UK
[8]College of Life Sciences, Shaanxi Normal University, South Chang'an Rd., Xi'an 710062, China
[9]Data Science Program, College of William & Mary, 540 Landrum Dr., Williamsburg, VA 23185, USA
[10]Department of Ecology and Evolutionary Biology, Cornell University, Corson Hall, Ithaca, NY 14853, USA
[11]Lead contact
*Correspondence: jrpuzey@wm.edu
https://doi.org/10.1016/j.cub.2023.07.057

SUMMARY

In intimate ecological interactions, the interdependency of species may result in correlated demographic histories. For species of conservation concern, understanding the long-term dynamics of such interactions may shed light on the drivers of population decline. Here, we address the demographic history of the monarch butterfly, *Danaus plexippus*, and its dominant host plant, the common milkweed *Asclepias syriaca* (*A. syriaca*), using broad-scale sampling and genomic inference. Because genetic resources for milkweed have lagged behind those for monarchs, we first release a chromosome-level genome assembly and annotation for common milkweed. Next, we show that despite its enormous geographic range across eastern North America, *A. syriaca* is best characterized as a single, roughly panmictic population. Using approximate Bayesian computation with random forests (ABC-RF), a machine learning method for reconstructing demographic histories, we show that both monarchs and milkweed experienced population expansion during the most recent recession of North American glaciers 10,000–20,000 years ago. Our data also identify concurrent population expansions in both species during the large-scale clearing of eastern forests (~200 years ago). Finally, we find no evidence that either species experienced a reduction in effective population size over the past 75 years. Thus, the well-documented decline of monarch abundance over the past 40 years is not visible in our genomic dataset, reflecting a possible mismatch of the overwintering census population to effective population size in this species.

INTRODUCTION

Despite the critical importance of understanding past population dynamics, especially for species of conservation concern, inferring demographic histories can be extremely challenging. Novel genomic methodologies based on sampling extant individuals and interpretation of genomic patterns of diversity have recently provided insight into the demographic histories of species ranging from protists to humans.[1,2] Over the past 25 years, conservationists have become increasingly alarmed by the decline of the monarch butterfly's overwintering population.[3–5] Despite significant academic and public energy focused on understanding and reversing this, the exact cause of this decline is still a matter of debate. Multiple factors have been proposed to underlie the monarch's decline, including a decrease in the abundance of the monarch's food source (primarily a single species of milkweed, common milkweed), reduced abundance or quality of nectar plants, climate change, and destruction of their overwintering sites.[6–9]

Here, we address correlated demographic changes of monarchs and milkweeds over three hypothesized critical events during the Holocene. Placing this recent decline in a historical context will help us begin to address fundamental questions about the relationship between milkweed, monarchs, and humans. For instance, did colonizing Europeans inadvertently increase the size of the monarch population by massively expanding common milkweed habitat through deforestation and ploughing of prairies (as suggested by Vane-Wright[10] and Brower[11])? Does the recent decline of the overwintering census population follow from an artificial high? Does it represent a

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

decline to levels lower than those seen before European colonization? And finally, are monarch and common milkweed population demographics matched, perhaps indicating that common milkweed is the limiting resource for monarch butterfly populations? Providing insight into these questions has remained intractable to date. However, recent advances in population genetic approaches and machine learning now allow us unprecedented ability to reconstruct demographic histories of populations.

To reconstruct the demographic histories of monarchs and milkweed, we use here approximate Bayesian computation with random forests (ABC-RFs).[12] Briefly, ABC modeling uses simulated datasets to estimate posterior probabilities when the likelihoods of observed data, given specific models, are difficult to calculate.[13,14] Genetic datasets are simulated under a number of different demographic models, and the simulated datasets closest to the observed data are used to estimate the posterior probabilities of individual models and distributions of parameters of interest. The RF approach described by Pudlo et al.[12] and Raynal et al.[15] implements a machine learning algorithm to do model selection and parameter estimation. The RF approach improves upon traditional ABC modeling in that ABC-RF is insensitive to the choice of summary statistics, as well as less computationally expensive. This approach has recently been employed by a number of population genetic studies on a diverse array of organisms, including insects,[16] plants,[17] chordates[18] including humans,[19] and pathogens,[1] and it has been used to reconstruct biological invasions and other demographic events happening within the past few decades or centuries.[20–22]

Accordingly, we use the ABC-RF approach to test how the last glacial retreat, the ploughing-up of the prairie and deforestation, and finally, expansion of industrial agriculture impacted monarch and milkweed populations. Specifically, we addressed the following questions: (1) have *Asclepias syriaca* (*A. syriaca*) and *Danaus plexippus* (*D. plexippus*) populations expanded in prior millennia (5–25 kya), potentially due to the retreat of the glaciers after the last glacial maximum[23]; (2) have *A. syriaca* and *D. plexippus* populations expanded in the past centuries (1751–1899), potentially due to the conversion of native forests and prairies to agriculture land, as suggested by, e.g., Brower[11]; and (3) have *A. syriaca* and *D. plexippus* populations experienced a bottleneck along with the industrialization of agriculture within past decades (1945–2015), potentially due to the increased use of herbicide in crop fields?[24,25]

To facilitate answering these questions, we assembled a new genome for *A. syriaca*. Previously existing genomic resources were limited to low coverage assemblies and transcriptomes.[26] Next, we sampled and conducted genomic analyses for 231 milkweed isolates from across the entire native range. Finally, using this dataset, we test a series of explicit hypotheses using ABC-RF to ask how these climate and anthropogenic events have impacted population change of these iconic species. We conducted these analyses in parallel on milkweed and monarchs, using previously published whole-genome sequencing data from Zhan et al. for the latter.[27] Therefore, our analysis addresses whether the demographic histories of this intimate species interaction are matched or independent.

RESULTS

Genome assembly

PacBio sequencing resulted in over 300× coverage of the expected genome size of 420 Mb. The sequence was assembled into 748 contigs with a total length of 362 Mbp and an N50 of 1.9 Mbp. Kmer analysis supports this genome size. After haplotig removal, approximately 91% of the sequence was scaffolded into 11 sequences representing pseudomolecules. The final assembly has a length of 317 Mbp and captures 96.8% of the BUSCO set.

Genome annotation of *A. syriaca*

Approximately 57% of the genome consists of repetitive sequences. A total of 42,111 genes were predicted with an average length of 2,578 bp. Approximately 93% of the BUSCO protein set was identified in the annotation. Putative functions were assigned to 99% of the gene set.

SNP calling

We gathered five different population genetic datasets for *D. plexippus* and *A. syriaca*. Collection sites and sample sizes for each dataset are shown in Figure 1A. The number of individuals and SNPs and the amount of missing data for each SNP dataset are shown in Table 1.

For common milkweed, the final datasets following rigorous SNP filtering were:

(1) Core Range genotyping by sequencing (GBS): the GBS approach sequenced and called approximately 900 SNPs from 87 plants.
(2) Broad Range GBS: the GBS approach sequenced and called approximately 900 SNPs from 96 plants.
(3) Broad Range whole-genome resequencing (WGR): the WGR approach identified approximately 900 SNPs from 48 plants.

For monarch butterflies:

(4) We called approximately 11,700 SNPs from 28 butterflies from Zhan et al.[27] These samples were collected between 2007 and 2009.
(5) The Talla et al. dataset we analyzed consisted of 29 individuals collected in October 2016 and 4,509 SNPs.[30]

Population genetic analysis

All three of our milkweed datasets showed little genetic structure across their ranges. Heterozygosities, both observed and expected, varied little across our populations (Table 1). Global proportion of total genetic variance partitioned among populations (F_{ST}) ranged from −0.002 (Broad Range WGR dataset) to 0.039 (Core Range GBS dataset), indicating a low amount of geographically sorted population structure. F_{ST} values between pairs of populations were similarly low, with the exception that the invasive European population was more distinct from the North American populations, with pairwise F_{ST} values around 0.08 (Table 2). We further interrogated this genetic structure using two approaches.

In the first approach, we used STRUCTURE to assign each individual ancestry to two or more subpopulations. It is important

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

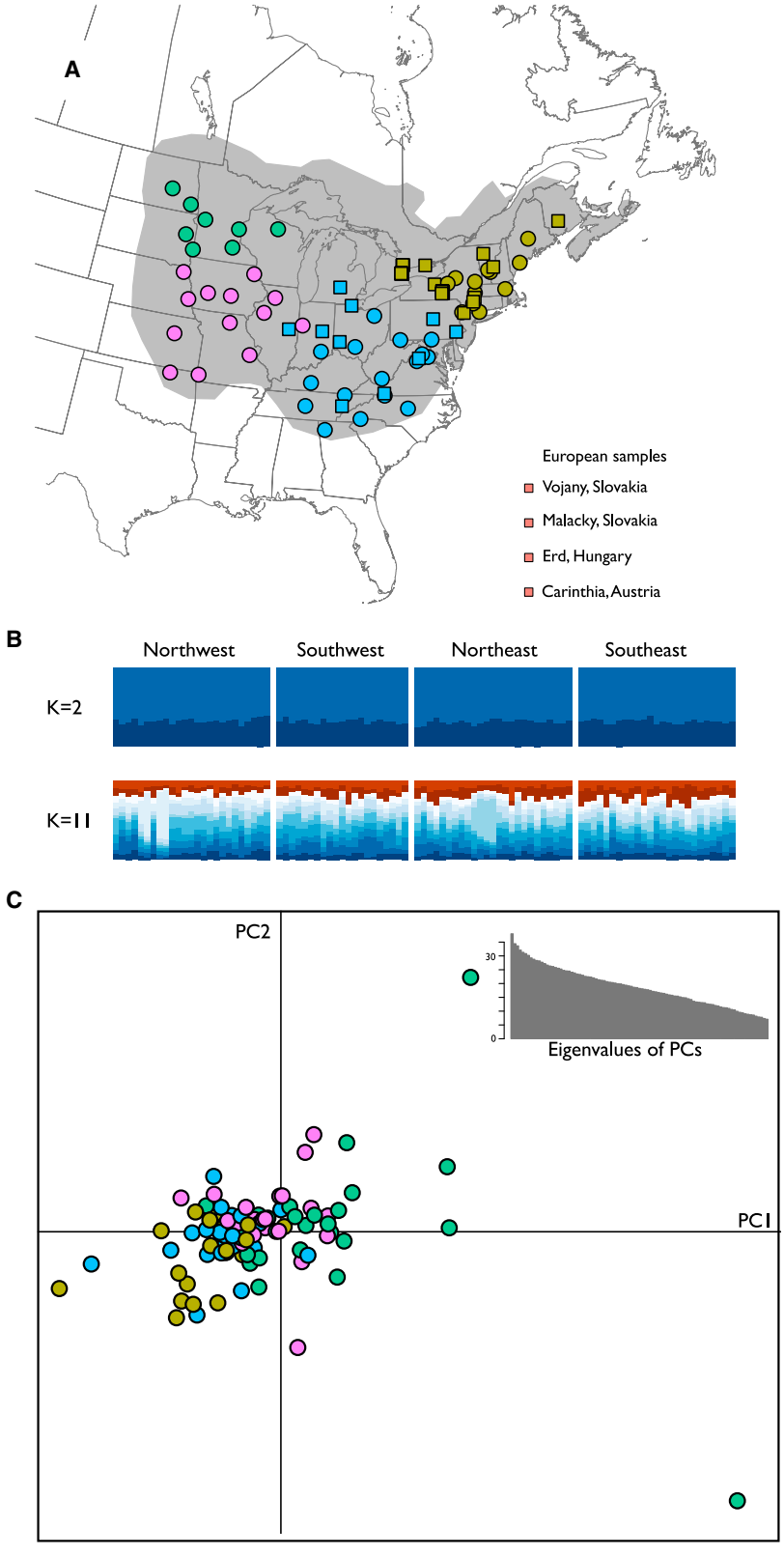

Figure 1. Population genetic structure of *A. syriaca*

(A) Our sampling scheme covers most of the North American range of *A. syriaca*. Circles represent sites sampled for the Broad Range datasets, while squares represent sites sampled for the Core Range datasets. Sites are colored according to the rough geographic zones to which we assigned them for the purposes of calculating F_{ST}. We assigned the Core Range site in Illinois to the southeastern population instead of the southwestern population, since otherwise we would have only one locality representing a population in that dataset. The gray region is an approximation of the range of *A. syriaca* based on specimen records in Global Biodiversity Information Facility.[28]

(B) STRUCTURE found no evidence of population structure among our milkweed specimens. The thin vertical bars represent individual milkweeds, and the four geographic zones are separated by thin white bars. Each bar is colored according to the cluster(s) to which it belongs. We present the results for the simplest analysis, in which STRUCTURE assumes k = 2 clusters, and the analysis chosen by the Evanno method as optimal, k = 11.[29] These results show strong genetic homogeneity across milkweed's range. These data are from the Broad Range GBS dataset; our other datasets produced similar results and are shown in the supplemental information for all k values from 2 to 20.

(C) PCA demonstrates weak geographic signal among some subsets of SNPs. Shown here are the first two principal-component (PC) axes of allele frequencies, with each point representing an individual milkweed from the Broad Range GBS dataset. Points are colored according to origin using the same color scheme as in (A). These two PC axes capture about 4% of the total variation. The inset shows the eigenvalues for each PC; these decline quite slowly, indicating that each individual PC axis explains relatively little of the variation in genotype. PC plots for additional axes, and for other datasets, show similarly weak levels of geographic signal and are given in the supplemental information. See also Figure S1.

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

Table 1. Population genetics of *A. syriaca* and *D. plexippus*

Dataset	SNPs	Missing data	Population	n	H_o	H_e	F_{IS}	F_{ST}
Broad Range GBS	891	5.2%	total	96	0.074	0.087	0.147	0.008[d]
			Northwest	25	0.076	0.089	0.107	N/A
			Southwest	21	0.077	0.087	0.085	N/A
			Northeast	25	0.071	0.085	0.120	N/A
			Southeast	25	0.073	0.087	0.103	N/A
Broad Range WGR	885[a]	3.4%	total	48	0.039	0.050	0.222	−0.002[e]
			Northwest	6	0.044	0.053	0.102	N/A
			Southwest	16	0.041	0.059	0.242	N/A
			Northeast	10	0.035	0.045	0.144	N/A
			Southeast	16	0.036	0.044	0.132	N/A
Core Range GBS	926	4.3%	total	87	0.076	0.088	0.134	0.039[d]
			Northeast	47	0.081	0.080	0.083	N/A
			Southeast	32	0.085	0.092	0.063	N/A
			Europe	8	0.062	0.080	0.176	N/A
Monarchs[27]	11,703[b]	3.6%	total	28	0.109	0.124	0.125	N/A
Monarchs[28]	4,509[c]	2.1%	total	29	0.109	0.134	0.189	N/A

n, sample size; H_o, observed heterozygosity; H_e, expected heterozygosity; F_{IS}, proportion of genetic variation in the population found in an individual; F_{ST}, proportion of total genetic variance partitioned among populations.
[a]27 loci had more than 2 alleles and were excluded from the ABC-RF analysis; a further 24 invariant SNPs were excluded from this analysis as well.
[b]1,272 loci had more than 2 alleles and were excluded from the ABC-RF analysis; a further 125 invariant SNPs were excluded from this analysis as well.
[c]566 loci had more than 2 alleles and were excluded from the ABC-RF analysis; a further 579 invariant SNPs were excluded from this analysis as well.
[d]AMOVA, $p < 1 \times 10^{-4}$
[e]AMOVA, $p = 0.47$

to note that STRUCTURE cannot be used to evaluate the fit of a single panmictic population as the optimal number of genetic clusters is determined based on the change in the log-likelihood between k values (see Janes et al.[31]). Regardless of the number of subpopulations chosen *a priori*, for every subpopulation, STRUCTURE assigned all individuals roughly the same degree of ancestry in that subpopulation, regardless of their geographic location (visualized in Figure 1B for the Broad Range GBS dataset). This was true across all three datasets; the one major exception was that in the Core Range GBS dataset, the invasive European population was quite distinct from the North American populations. STRUCTURE results for all three datasets are provided in the supplemental information.

Second, to circumvent the inability of STRUCTURE to evaluate k = 1, we took a less-parameterized approach by performing a principal-component analysis (PCA) on the allele frequencies of the SNPs in each dataset (Figure 1C). This approach identifies groups of covarying SNPs. For all three datasets, none of the first six principal-component (PC) axes clearly separate any population from any other(s); although some PC axes show some degree of geographic structure, there is always a considerable degree of overlap between the PC values of the various populations. For instance, in the Broad Range GBS dataset (visualized in Figure 1C), PC1 largely separates several northwestern individuals from the remainder of the dataset, possibly indicating introgression from *A. speciosa*, which is known to hybridize with *A. syriaca* in the northwestern part of the *A. syriaca* range. PC2 shows a slight amount of geographic signal, with western populations tending toward positive values and eastern populations tending toward negative values, but individuals from all four regions are well mixed in PC space, indicating that this geographic signal is quite weak.

All three datasets support the conclusion that in North America, *A. syriaca* is a single large metapopulation with little geographic structure. Our results for *A. syriaca* parallel findings for the monarch butterflies, which show a lack of geographic population genetic structure in North America.[32,33]

Demographic modeling

Projecting our observed data onto the LDA axes of our simulated data indicated that our set of demographic models were realistic, as the observed data fell within or near the cloud of simulated data points along all LDA axes (Figure S3). Also, per Pudlo et al., we also confirmed that we produced enough simulations, as a preliminary analysis showed that the prior error rate decreased only slightly by the addition of the last 20% of simulations[12] (Table S1). In fact, we found a few cases in which error rates went up slightly after adding the final 20% of the data (by 0.3% or less), indicating that we are in the regime in which changes in error rate are determined by random fluctuations and confirming that adding more simulations will not further improve the accuracy of this method. Furthermore, we followed the recommendation of Pudlo et al. for determining whether we had used enough decision trees in our RF algorithm.[12] To do this, we repeated the RF algorithm several times using fewer trees, recalculating the prior error rate each time. If the error rate stays nearly flat as we approach the maximum number of trees, this means that we used an appropriate number of trees, which was indeed the case for all three datasets (Figure S3). Finally, we also confirmed that our RFs were not overfitting to their

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

Table 2. Population structure of *A. syriaca*

Dataset	Pairwise comparison	Pairwise F_{ST} (GBS)	Pairwise F_{ST} (WGR)
Broad Range	Northwest vs. Southwest	0.009	−0.021
	Northwest vs. Northeast	0.019	0.004
	Northwest vs. Southeast	0.017	−0.000
	Southwest vs. Southeast	0.009	0.002
	Southwest vs. Northeast	0.011	0.008
	Northeast vs. Southeast	0.002	−0.000
Core Range	Northeast vs. Southeast	0.009	N/A
	Northeast vs. Europe	0.082	N/A
	Southeast vs. Europe	0.081	N/A

GBS, data from genotyping by sequencing approach; WGR, data from whole genome resequencing approach.

training dataset by comparing performance on testing and training datasets. We found similar accuracies when comparing testing and training datasets for all RFs (Table S2).

Our RF results were consistent across all three milkweed datasets and between monarchs and milkweeds (Figure 2). All 20 runs for each of the 5 datasets predicted the presence of a post-glacial expansion in population size, with an average posterior probability between 0.64 and 0.85. All 20 runs for each dataset also predicted the presence of a more recent population expansion alongside 18th and 19th century agriculture, with an average posterior probability between 0.71 and 0.97.

There was more uncertainty with respect to the presence or absence of a recent bottleneck alongside the industrialization of agriculture. All 20 runs for the Broad Range GBS and Core Range GBS milkweed datasets predicted the absence of a recent bottleneck, although with less confidence: posterior probabilities were between 0.47 and 0.67. Both monarch datasets indicated the lack of a recent bottleneck, albeit with differing confidences. The monarch dataset collected from 2007 to 2009 had a posterior probability of 0.47 while the more recently collected Talla et al. monarchs had a posterior probability of 0.85. The Broad Range WGR dataset had 15 runs predicting the absence of a bottleneck (0.47 average posterior probability) and 5 runs predicting its presence (0.55 posterior probability). Model parameters estimated with the ABC-RF approach were nearly identical to their prior distributions, suggesting that our dataset does not have sufficient resolution for parameter estimation (results not shown).

Note that posterior probabilities can be relatively low even when all 20 runs produce the same results. The agreement of the different runs shows that the RF method produces similar predictions for the same observed datasets; however, it does not show how conclusively a particular dataset can rule in or out a particular demographic event; posterior probabilities are an attempt to capture the latter.

DISCUSSION

Understanding the impact of the Anthropocene on the natural world is of fundamental importance for conservation efforts. Until recently, elucidating patterns of population change in the recent past has been very difficult. In this study, we employ an ABC-RF approach to study the near-term demographic history of monarchs and milkweeds. This approach was chosen in part because it has proven useful in other systems in elucidating very recent demographic events, within decades or centuries.[21,22] In addition, this approach requires fewer simulated datasets to train the classifier than are necessary for traditional ABC, and it is much more robust to choices of summary statistics.[12,34]

We tested for changes in the effective population size of the monarch butterfly and its primary food source, common milkweed, during three events: the most recent retreat of the glaciers, European settlement in North America, and industrial agriculture. Previously, using a PSMC (pairwise sequentially Markovian coalescent) model, a method capable of testing for ancient events but less fit for resolving recent events, researchers demonstrated a population expansion of monarch butterflies after the last glaciation.[27] Using ABC-RF, we likewise detect this expansion in monarch effective population sizes and also observe an expansion of common milkweed post-glaciation; we hypothesize that both are likely owing to the large increase in ranges available to these species with the retreat of the glaciers. The low levels of population structure in common milkweed likely occur because the modern range of *A. syriaca* is a result of rapid (i.e., in the last 5–25 kya) invasion of central and eastern North America after the retreat of the glaciers. In this scenario, the rapid expansion, combined with *A. syriaca* being an obligate outcrosser with long-distance dispersal ability, has prevented the formation of extensive population structure. It is also possible that milkweed existed in a single refugium during the last glaciation, resulting in a homogenization of genetic variation.

We provide population genetic evidence that common milkweed increased in abundance during the 18th and 19th centuries. The most obvious cause for this is the clearing of forests and prairies to make way for agricultural land, a disturbance-rich environment in which *A. syriaca* thrived (at least, until the advent of herbicides). The increase observed in our data has previously been suspected, and there are two major hypotheses for how this increase affected monarch butterflies. The first hypothesis posits that *A. syriaca* has always been the most important host plant for monarchs, even before *A. syriaca*'s population boom. As *A. syriaca* increased in abundance in a newly disturbed landscape, monarchs increased in abundance alongside them. Thus, according to this hypothesis, the current size (and possible geographic extent) of the monarch migration was greater in the 18th–20th centuries than in the 17th century and prior[11]; a more radical form of this hypothesis suggests that the migratory behavior itself was absent before the 18th century.[10] However, although *A. syriaca* has increased in abundance due to disturbance, it is likely that other species of milkweeds, less tolerant of anthropogenic changes, have declined in abundance over the same period. The second hypothesis suggests that monarchs transitioned from a wider array of host plant species to become more reliant on common milkweed over this period of increase in common milkweed populations. If this occurred, then the newly increased population sizes of *A. syriaca* did not represent a net increase in food resources for monarchs, and so we would not expect the monarch abundances in the 18th–20th centuries to be higher (or lower) than previously.[11]

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

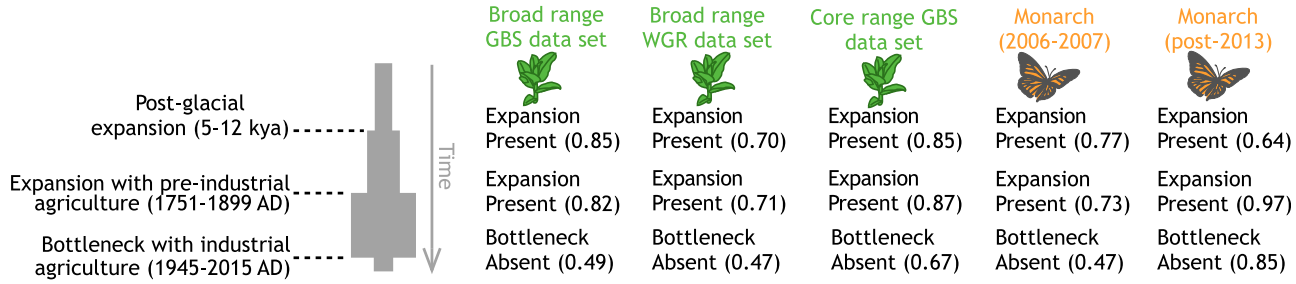

Figure 2. Population demographic modeling of *A. syriaca* and *D. plexippus*

Support for each of our hypothesized demographic events in our three milkweed and two monarch datasets. The random forest consensus on whether each event is present in the population history of that species is given, along with the estimated posterior probability of each in parentheses. The post-2013 monarch dataset was added post hoc at the suggestion of a reviewer. See also Table S3.

Regarding the fact that we found no evidence for a reduction in the effective population sizes of the monarchs or milkweed over the past 75 years, the simplest explanation for these results is that the demographic event in question did not occur. A second possibility is that the demographic event did occur, but it had an effect size that is too small to leave a signal in our dataset. Unfortunately, our dataset was not sensitive enough to estimate posterior distributions for the strength of these expansions or bottlenecks, so we are not able to quantify absolutely the minimum detectable event size. However, we can be confident that detected events are larger than undetected ones: i.e., if there was an undetected decline in monarch population size since 1945, it was less than the detected increase that occurred in the 18th and 19th centuries. A third possibility, relevant to the hypothesized bottleneck with agriculture in past decades, is that the demographic event has occurred but too recently to produce a detectable, population genetic signal. In this case, bottlenecks reduce diversity not only directly (via the elimination of the majority of lineages in the population when the bottleneck event occurs) but also indirectly after the bottleneck, as the new, smaller population size means that fixation at a particular locus is more likely, thus eliminating even more genetic diversity after the bottleneck event. It is possible that a bottleneck has occurred in the past decades, but we are unable to see it because there has not yet been enough time for alleles to be driven to fixation in the new, reduced populations. This effect is likely to be stronger in milkweeds, which have a roughly 10-fold longer generation time than do monarchs. This is one possible explanation for why the well-documented recent declines in monarch and milkweed population sizes are not reflected in our data.[3–5]

The monarchs sampled for our *D. plexippus* analysis were collected by Zhan et al. between 2007 and 2009, several years before the all-time low of the Mexican overwintering population in the winter of 2013–2014.[27,35] One possible explanation for why our population genetic data do not show clear signals of a recent decline is that our samples were collected before the lowest population sizes occurred. At the request of a reviewer, we ruled out this possibility by examining monarchs collected after the lowest point of the Mexican overwintering population. To do this, we used the sequences published by Talla et al.,[30] which were collected in 2016, and we repeated our analyses with these samples (details provided in Table S3). The results of these analyses were the same as for the monarch sequences from Zhan et al., showing that our results were not being affected by missing the tail end of the monarch decline in 2013–2014.

Our results indicate an increase in monarch populations alongside those of common milkweed in the 18th and 19th centuries. How should biologists and conservationists react to this new data? This depends largely on which hypothesis about the monarch response to this increase is correct. If the 20th century population size of the monarch was anthropogenically inflated due to elevated common milkweed abundance, this puts contemporary declines in a less worrisome light, as they may simply represent returns to pre-modern population sizes. Monarch population sizes and migratory behavior were presumably sustainable for centuries before the clearing of the forests and prairies of eastern North America. However, if monarchs responded to increased common milkweed abundance by shifting their diets without increasing the total population, then contemporary declines may well have put the monarchs at their lowest population size since the retreat of the glaciers. It is also important to note that while monarch and milkweed populations experienced correlated increases in the 18th and 19th centuries, this correlated increase does not necessarily imply that increase in milkweed populations is completely causal for driving monarch population growth. Rather, it is possible that the ecological factors that drove milkweed growth also resulted in other changes that were beneficial to the monarch. For instance, deforestation and spread of agricultural fields could result in an increase in nectar-bearing plants that would be beneficial to migrating monarchs.

The results presented here suggest that the recent decline of the monarch butterfly may be (at least in part) a return to pre-modern population sizes. That said, we encourage restraint in the interpretation of these results and encourage parallel studies to test these ideas further. Fully answering this question using population genetics will probably require improvements in our current techniques for demographic modeling and/or denser sequencing of *A. syriaca* and *D. plexippus* individuals than is currently available. However, there are other potential datasets that could shine light on this question. As a start, population genomic analyses for other important milkweed species could reveal whether or not they declined during the period of common milkweed's increase: lack of such declines would suggest that the expansion of *A. syriaca* in particular could only have

increased the monarch population. Brower suggested sampling cardenolide profiles from museum specimens of monarchs captured in the 19th and 20th centuries.[11] These profiles can indicate the specific milkweed species those individuals used as larvae and thus show whether or not monarchs experienced a shift in their host species as humans cleared forests and prairies. Shifts to more diversity in milkweed hosts might also be detectable in more recent specimens collected on the East Coast of North America, as farming has become less prevalent in this region over past decades. The presence of such recent shifts (e.g., on to *Asclepias incarnata* [*A. incarnata*]) would support the notion that changes in the availability of some hosts cause shifts in use of others, as hypothesized above.

We emphasize that our results do not directly bear on current efforts to support monarch butterfly conservation. Regardless of how many monarchs were in North America in 1600, the current monarch population brings delight to people across North America and serves as a key conservation icon that introduces many non-scientists to the importance of invertebrate conservation, pollination biology, migratory behavior, and more. Having fewer of these charismatic insects present would be a loss to humankind regardless of how many of them were present a few centuries ago.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Genome assembly and annotation
  - SNP calling
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Population genetic analysis
  - Demographic modelling

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2023.07.057.

ACKNOWLEDGMENTS

This work was supported by US National Science Foundation award 1645256 (to G.J. and A.A.A.), United States Department of Agriculture award 2020-67013-30896 (to G.J.), Triad Foundation (to G.J.), Jeffress Trust Awards Program in Interdisciplinary Research (to J.R.P.), Dominion Education Partnership (to J.R.P. and H.J.D.), and National Geographic Society GR-000000959 (to J.R.P. and H.J.D.).

AUTHOR CONTRIBUTIONS

This project was conceived by J.H.B., G.J., A.A.A., and J.R.P. The assembly and annotation of the milkweed genome were performed by S.S., A.P., J.Z., G.J., and H.X. Collection of milkweed samples, DNA extractions, and DNA library preparation were done by A.R., H.J.D., H.X., and A.D.T. Design of the ABC portion of the project was overseen by J.H.B. and J.P. J.H.B. conducted the population genetic analyses. The paper was primarily written by J.H.B. and J.R.P., with significant editing input from A.A.A. and A.D.T. All authors approved the final text. ChatGPT was used as an editor to suggest revisions for textual clarity.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 9, 2022
Revised: April 13, 2023
Accepted: July 26, 2023
Published: August 21, 2023

REFERENCES

1. Schwabl, P., Boité, M.C., Bussotti, G., Jacobs, A., Andersson, B., Moreira, O., Freitas-Mesquita, A.L., Meyer-Fernandes, J.R., Telleria, E.L., Traub-Csekö, Y., et al. (2021). Colonization and genetic diversification processes of Leishmania infantum in the Americas. Commun. Biol. *4*, 139.
2. Lepers, C., Billiard, S., Porte, M., Méléard, S., and Tran, V.C. (2021). Inference with selection, varying population size, and evolving population structure: application of ABC to a forward-backward coalescent process with interactions. Heredity *126*, 335–350.
3. Thogmartin, W.E., Wiederholt, R., Oberhauser, K., Drum, R.G., Diffendorfer, J.E., Altizer, S., Taylor, O.R., Pleasants, J., Semmens, D., Semmens, B., et al. (2017). Monarch butterfly population decline in North America: identifying the threatening processes. R. Soc. Open Sci. *4*, 170760.
4. Pleasants, J.M., Zalucki, M.P., Oberhauser, K.S., Brower, L.P., Taylor, O.R., and Thogmartin, W.E. (2017). Interpreting surveys to estimate the size of the monarch butterfly population: pitfalls and prospects. PLoS One *12*, e0181245.
5. Brower, L.P., Taylor, O.R., Williams, E.H., Slayback, D.A., Zubieta, R.R., and Ramírez, M.I. (2012). Decline of monarch butterflies overwintering in Mexico: is the migratory phenomenon at risk? Insect Conserv. Divers. *5*, 95–100.
6. Haan, N.L., and Landis, D.A. (2019). The importance of shifting disturbance regimes in monarch butterfly decline and recovery. Front. Ecol. Evol. *7*. https://doi.org/10.3389/fevo.2019.00191.
7. Boyle, J.H., Dalgleish, H.J., and Puzey, J.R. (2019). Monarch butterfly and milkweed declines substantially predate the use of genetically modified crops. Proc. Natl. Acad. Sci. USA *116*, 3006–3011.
8. Inamine, H., Ellner, S.P., Springer, J.P., and Agrawal, A.A. (2016). Linking the continental migratory cycle of the monarch butterfly to understand its population decline. Oikos *125*, 1081–1091.
9. Zylstra, E.R., Ries, L., Neupane, N., Saunders, S.P., Ramírez, M.I., Rendón-Salinas, E., Oberhauser, K.S., Farr, M.T., and Zipkin, E.F. (2021). Changes in climate drive recent monarch butterfly dynamics. Nat. Ecol. Evol. *5*, 1441–1452.
10. Vane-Wright. (1993). The Columbus hypothesis: an explanation for the dramatic 19th century range expansion of the monarch butterfly. In Biology and Conservation of the Monarch Butterfly (Natural History Museum of Los Angeles County).
11. Brower, L.P. (1995). Understanding and misunderstanding the migration of the monarch butterfly (Nymphalidae) in North America: 1857–1995. J. Lepidopterists Soc.
12. Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C.P. (2016). Reliable ABC model choice via random forests. Bioinformatics *32*, 859–866.

13. Sisson, S.A., Fan, Y., and Beaumont, M.A. (2018). Overview of ABC. [2019]: (Chapman and Hall/CRC). In Handbook of Approximate Bayesian Computation (CRC Press), pp. 3–54.
14. Beaumont, M.A., Zhang, W., and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. Genetics *162*, 2025–2035.
15. Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C.P., and Estoup, A. (2019). ABC random forests for Bayesian parameter inference. Bioinformatics Oxf. Engl. *35*, 1720–1728.
16. Lombaert, E., Ciosi, M., Miller, N.J., Sappington, T.W., Blin, A., and Guillemaud, T. (2018). Colonization history of the western corn rootworm (*Diabrotica virgifera virgifera*) in North America: insights from random forest ABC using microsatellite data. Biol. Invasions *20*, 665–677. https://doi.org/10.1007/s10530-017-1566-2.
17. Nevado, B., Harris, S.A., Beaumont, M.A., and Hiscock, S.J. (2020). Rapid homoploid hybrid speciation in British gardens: the origin of Oxford Ragwort (Senecio Squalidus). Mol. Ecol. *29*, 4221–4233.
18. Smith, C.C.R., Flaxman, S.M., Scordato, E.S.C., Kane, N.C., Hund, A.K., Sheta, B.M., and Safran, R.J. (2018). Demographic inference in barn swallows using whole-genome data shows signal for bottleneck and subspecies differentiation during the Holocene. Mol. Ecol. *27*, 4200–4212. https://doi.org/10.1111/mec.14854.
19. Estoup, A., Verdu, P., Marin, J.-M., Robert, C., Dehne-Garcia, A., Cornuet, J.-M., and Pudlo, P. (2018). Application of ABC to infer the genetic history of Pygmy hunter-gatherer populations from western central Africa. In Handbook of Approximate Bayesian Computation (Chapman and Hall/CRC). https://doi.org/10.1201/9781315117195-18.
20. Boheemen, L.A. van, Lombaert, E., Nurkowski, K.A., Gauffre, B., Rieseberg, L.H., and Hodgins, K.A. (2017). Multiple introductions, admixture and bridgehead invasion characterize the introduction history of Ambrosia artemisiifolia in Europe and Australia. Mol. Ecol. *26*, 5421–5434.
21. Vallejo-Marín, M., Friedman, J., Twyford, A.D., Lepais, O., Ickert-Bond, S.M., Streisfeld, M.A., Yant, L., van Kleunen, M., Rotter, M.C., and Puzey, J.R. (2021). Population genomic and historical analysis suggests a global invasion by bridgehead processes in Mimulus guttatus. Commun. Biol. *4*, 327. https://doi.org/10.1038/s42003-021-01795-x.
22. Fraimout, A., Debat, V., Fellous, S., Hufbauer, R.A., Foucaud, J., Pudlo, P., Marin, J.M., Price, D.K., Cattel, J., Chen, X., et al. (2017). Deciphering the routes of invasion of drosophila Suzukii by means of ABC random forest. Mol. Biol. Evol. *34*, 980–996.
23. Clark, P.U., Dyke, A.S., Shakun, J.D., Carlson, A.E., Clark, J., Wohlfarth, B., Mitrovica, J.X., Hostetler, S.W., and McCabe, A.M. (2009). The last glacial maximum. Science *325*, 710–714.
24. Pleasants, J.M., and Oberhauser, K.S. (2013). Milkweed loss in agricultural fields because of herbicide use: effect on the monarch butterfly population. Insect Conserv. Divers. *6*, 135–144.
25. Pleasants, J.M., and Oberhauser, K.S. (2017). Milkweed loss in agricultural fields because of herbicide use: effect on the monarch butterfly population. Insect Conserv. Divers. *6*, 135–144. https://doi.org/10.31219/osf.io/wmj6e.
26. Weitemier, K., Straub, S.C.K., Fishbein, M., Bailey, C.D., Cronn, R.C., and Liston, A. (2019). A draft genome and transcriptome of common milkweed (Asclepias syriaca) as resources for evolutionary, ecological, and molecular studies in milkweeds and Apocynaceae. PeerJ *7*, e7649.
27. Zhan, S., Zhang, W., Niitepõld, K., Hsu, J., Haeger, J.F., Zalucki, M.P., Altizer, S., de Roode, J.C., Reppert, S.M., and Kronforst, M.R. (2014). The genetics of monarch butterfly migration and warning colouration. Nature *514*, 317–321.
28. GBIF (2021). Free and open access to biodiversity data. https://www.gbif.org.
29. Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. *14*, 2611–2620.
30. Talla, V., Pierce, A.A., Adams, K.L., de Man, T.J.B., Nallu, S., Villablanca, F.X., Kronforst, M.R., and de Roode, J.C. (2020). Genomic evidence for gene flow between monarchs with divergent migratory phenotypes and flight performance. Mol. Ecol. *29*, 2567–2582.
31. Janes, J.K., Miller, J.M., Dupuis, J.R., Malenfant, R.M., Gorrell, J.C., Cullingham, C.I., and Andrew, R.L. (2017). The K = 2 conundrum. Mol. Ecol. *26*, 3594–3602.
32. Hemstrom, W.B., Freedman, M.G., Zalucki, M.P., Ramírez, S.R., and Miller, M.R. (2022). Population genetics of a recent range expansion and subsequent loss of migration in monarch butterflies. Mol. Ecol. *31*, 4544–4557.
33. Lyons, J.I., Pierce, A.A., Barribeau, S.M., Sternberg, E.D., Mongue, A.J., and De Roode, J.C. (2012). Lack of genetic differentiation between monarch butterflies with divergent migration destinations. Mol. Ecol. *21*, 3433–3444.
34. Csilléry, K., François, O., and Blum, M.G.B. (2012). Abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. *3*, 475–479.
35. Semmens, B.X., Semmens, D.J., Thogmartin, W.E., Wiederholt, R., López-Hoffman, L., Diffendorfer, J.E., Pleasants, J.M., Oberhauser, K.S., and Taylor, O.R. (2016). Quasi-extinction risk and population targets for the Eastern, migratory population of monarch butterflies (Danaus plexippus). Sci. Rep. *6*, 23265.
36. Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770.
37. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics *33*, 2202–2204.
38. Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13*, 1050–1054.
39. Roach, M.J., Schmidt, S.A., and Borneman, A.R. (2018). Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics *19*, 460.
40. Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science *356*, 92–95.
41. Xu, G.C., Xu, T.J.X., Zhu, R., Zhang, Y., Li, S.-Q., Wang, H.-W., and Li, J.-T. (2019). LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. GigaScience *8*. https://doi.org/10.1093/gigascience/giy157.
42. Ou, S., and Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. *176*, 1410–1422.
43. Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics *9*, 18.
44. Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. *35*. W265–W268.
45. Smit, A.F.A., Hubley, R., and Green, P. (2013–2015). RepeatMasker Open-4.0. http://www.repeatmasker.org.
46. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
47. Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., et al. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. *164*, 513–524.
48. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.
49. Mapleson, D., Venturini, L., Kaithakottil, G., and Swarbreck, D. (2018). Efficient and accurate detection of splice junctions from RNA-seq with portcullis. GigaScience *7*. https://doi.org/10.1093/gigascience/giy131.

50. Venturini, L., Caim, S., Kaithakottil, G.G., Mapleson, D.L., and Swarbreck, D. (2018). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience *7*. https://doi.org/10.1093/gigascience/giy093.

51. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped CDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637–644.

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

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

54. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). UniProtKB/Swiss-Prot. In Plant Bioinformatics: Methods and Protocols, D. Edwards, ed. (Humana Press), pp. 89–112.

55. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240.

56. Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. *35*, 543–548. https://doi.org/10.1093/molbev/msx319.

57. Fulton, T.M., Chunwongse, J., and Tanksley, S.D. (1995). Microprep protocol for extraction of DNA from tomato and other herbaceous plants. Plant Mol. Biol. Rep. *13*, 207–209. https://doi.org/10.1007/BF02670897.

58. Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One *6*, e19379.

59. Rochette, N.C., Rivera-Colón, A.G., and Catchen, J.M. (2019). Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. *28*, 4737–4754.

60. Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., and Cresko, W.A. (2013). Stacks: an analysis tool set for population genomics. Mol. Ecol. *22*, 3124–3140.

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

62. 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 Oxf. Engl. *25*, 2078–2079.

63. Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics Oxf. Engl. *24*, 1403–1405.

64. Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics Oxf. Engl. *27*, 3070–3071.

65. R Core Team (2018). R: A language and environment for statistical computing. https://www.R-project.org.

66. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

67. DIYABC (2015). script that convert vcf file into genotype data matrix to input DIYABC program. (GitHub). https://github.com/loire/vcf2DIYABC.py.

68. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

69. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

70. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics *43*, 11.10.1–11.10.33.

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

72. Goudet, J. (2005). Hierfstat, a package for r to compute and test hierarchical F-statistics. Mol. Ecol. Notes *5*, 184–186.

73. Paradis, E. (2010). Pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics Oxf. Engl. *26*, 419–420.

74. Kamvar, Z.N., Tabima, J.F., and Grünwald, N.J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ *2*, e281.

75. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.

76. Earl, D.A., and vonHoldt, B.M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. *4*, 359–361.

77. Jakobsson, M., and Rosenberg, N.A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics Oxf. Engl. *23*, 1801–1806.

78. Dray, S., and Dufour, A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. J. Stat. Software *22*, 1–20.

79. Chessel, D., Dufour, A.B., and Thioulouse, J. (2004). The ade4 package-I: one-table methods. R News *4*, 5–10.

80. Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., Marin, J.-M., and Estoup, A. (2014). DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. Bioinformatics Oxf. Engl. *30*, 1187–1189.

81. Bhowmik, P.C., and Bandeen, J.D.. The biology of Canadian weeds: 19. Asclepias syriaca L. Can. J. Plant Sci. 56, 579–589.

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

STAR★METHODS

KEY RESOURCES TABLE

Data	Accession
Milkweed GBS fastq files	SRA PRJNA975199
Milkweed WGS fastq files	SRA PRJNA975923
Code used for analyses	Dryad: https://doi.org/10.5061/dryad.k98sf7mc4
Milkweed Genome and Annotation	GenBank PRJNA787127

RESOURCE AVAILABILITY

Lead contact

Requests for additional information or resources should be directed to Joshua Puzey (jrpuzey@wm.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All original code has been deposited at Dryad and is publicly available as of the date of publication. https://doi.org/10.5061/dryad.k98sf7mc4

Raw sequencing data (GBS) used for population genetic analysis of *A. syriaca* are available on SRA (PRJNA975199).

Raw sequencing data (WGR) used for population genetic analysis of *A. syriaca* are available on SRA (PRJNA975923).

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The GBS milkweed samples used in this study were collected from wild grown plant at locations depicted in Figure 1. Plants for WGS were grown in the W&M greenhouse under ambient conditions prior to collecting a leaf for DNA extraction and sequencing.

The genome assembly and annotation of A. syriaca prsesented in this paper is available on GenBank (PRJNA787127).

METHOD DETAILS

To investigate correlated demographic histories of monarchs and milkweeds, we used five different data sets (Figure 1). In brief, the five data sets were:

(1) Core Range GBS: We used a GBS approach to sequence SNPs from 87 plants from 30 sites, primarily collected in the eastern portion of this species' range, with 1-5 plants per site. This data set includes 8 individuals collected from 4 sites in eastern Europe, where *A. syriaca* is an invasive species. Sites are mapped in Figure S1. The GBS approach was adopted to maximize the number of individuals genotyped at a subset of loci across the entire genome.

(2) Broad Range GBS: We used a Genotyping by Sequencing (GBS) approach to sequence SNPs from 96 plants from 47 sites across the North American range of this species, with 1-5 plants per site. Sites are mapped in Figure S1.

(3) Broad Range WGR: We used a skimming Whole Genome Resequencing (WGR) approach at low coverage to identify SNPs from plants collected from 48 sites across the North American range of this species, with 1 plant per site. Sites are mapped in Figure S1. The WGR approach was used to ensure that our results were not dependent on the specific SNP set produced by GBS.

We analyzed the two different GBS datasets separately as they were produced in different labs and had different sequencing coverages.

For monarch butterflies, we used:

(4) the whole genome sequences published by Zhan et al.[27], using 28 butterflies collected in 2007-2009 across the North American migratory range of this species.[27]

(5) A fifth *D. plexippus* dataset was added post-hoc at the suggestion of a reviewer. The reviewer hypothesized that the reason why no recent bottleneck (see Figure 2) was detected in the monarchs was because the Zhan et al. dataset consists of

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

monarchs collected before nadir of the monarch overwintering population in 2013-14. To address this idea, we conducted demographic analyses on a dataset of *D. plexippus* genotypes from individuals collected post-2013.[30] This dataset consisted of WGS from 29 butterflies collected from the Western North American monarch population (which were genetically indistinguishable from Eastern North American monarchs).[30]

Genome assembly and annotation

Genome sequencing and assembly of A. syriaca

Genomic DNA was prepared from one individual of *A. syriaca* from Stroglach, Austria (46.66° N, 14.47° E) and sequenced using PacBio CLR technology on six SMRT cells. Illumina sequence was generated from genomic DNA on one lane of Hi-Seq 2 x 150 bp. Kmer analysis was performed using this Illumina sequence, Jellyfish,[36] and Genomescope.[37] Hi-C libraries were prepared using the Proximo Hi-C kit for plants (Phase Genomics) and sequenced on one lane of Illumina 2 x 150 bp. *A. syriaca* PacBio sequence was assembled using Falcon v 2017.11.02-16.04 and falcon-kit 1.3.0 and the configuration file (fc_run.cfg).[38] The assembly was corrected using the Illumina sequence and Pilon v1.23. Redundancy was removed using Purge Hapolotigs.[39] Hi-C was used to scaffold the contigs using 3D-DNA v 180419[40] and gaps were filled with LR_gapcloser[41] and corrected PacBio reads.

Genome annotation of A. syriaca

For repeat identification and masking, LTR_retriever[42] was used with outputs from LTRharvest[43] and LTR_FINDER[44] to identify long terminal repeat retrotransposons (LTRs). The LTR library was then used to hard mask the genome, and RepeatModeler version: open-1.0.11[45] was used to identify additional repetitive elements in the remaining unmasked segments of the genome. Protein-coding sequences were excluded using blastx v2.7.1+[43,46] results in conjunction with the ProtExcluder.pl script from the ProtExcluder v1.2 package.[47] The libraries from RepeatModeler and LTR_retriever were then combined and used with RepeatMasker version: open-4.0.7[45] to produce the final masked version of the genome.

Libraries with an insert size of 350 bp were prepared from leaf RNA and sequenced on one lane of 2 x 100 bp Illumina Hi-Seq. RNA-seq reads were mapped to the genome with HISAT2 v2.2.0.[48] Portcullis v 1.1.2[49] and Mikado v 1.2.2[50] were used to process and filter the resulting bam files. Augustus v 3.2.0[51] and Snap v 2006-07-28[52] were trained and implemented through the Maker v 2.31.10 pipeline,[53] with proteins from Swiss-Prot[54] and processed RNA-seq added as evidence. Gene models were filtered with the following criteria: 1) at least one match found in the Trembl database (4-17-19)[54] with an E-value less than 1e-20, 2) InterProScan matches to repeats were removed, 3) genes with an AED score of 1 and no InterPro domain were removed, and 4) single-exon genes with no InterPro domain were removed. Functional annotation and classification were performed using BLASTx v2.7.1+[46] and InterProScan v5.36-75.0.[55] Both genome and annotation completeness were assessed by BUSCO v3.1.0[56] using the embryophyta lineage.

SNP calling

Genotyping by sequencing (GBS) of the A. syriaca *Core Range data set*

Common milkweed plants collected from different places around US and Europe were germinated and cultivated in our greenhouse. Fresh collected tissue was flash frozen in liquid nitrogen. The DNA was extracted from the leaf of individuals using a CTAB (cetyltrimethyl ammonium bromide)-based extraction protocol (adapted from Fulton et al.[57]). The DNA was quantified using a CFX384 C1000 Real-Time thermal cycler (BioRad, Hercules, CA) and normalized to 30–100 ng/ul using a GBFit Arise Pipetting System (Pacgen Inc., Irvine, CA). Quality checks were performed by agarose gel observation of 300 ng of undigested and *Hind*III digested DNA per sample. Genotyping was performed following the GBS protocol,[58] using *Ape*KI as the restriction enzyme. The libraries were sequenced on a HiSeq 2500 system (Illumina Inc., USA) with the single-end mode and read length of 101 bp.

Genotyping by sequencing (GBS) of the A. syriaca *Broad Range data set*

DNA was extracted from flash-frozen leaf samples using the Qiagen DNeasy Plant extraction kit. 100ng of sample DNA was used for GBS library preparation using the ApeKI restriction enzyme, as above. 95 samples and a water control (blank) were pooled per multiplex and sequenced using 100bp single-end mode on the HiSeq 2500 at the University of Rochester Medical Center.

Whole genome resequencing (WGR) of the A. syriaca *Broad Range data set*

DNA was extracted from *A. syriaca* using Qiagen DNeasy kit, libraries prepared using Illumina library DNA kit, and sequenced using Illumina HiSeq 2x150.

SNP calling of the A. syriaca *Core and Broad Range GBS data sets*

Genotyping By Sequencing reads were demultiplexed using Stacks 2.2.[59,60] Reads from each individual were then mapped against the *A. syriaca* genome using Bowtie2 2.3.2,[61] using end-to-end alignment and the "--very-sensitive" alignment settings. Reads with a mapping quality lower than 5 were discarded using samtools 1.5.[62] We then used Stacks in combination with custom scripts to call SNPs and to filter low-quality individuals and loci from our data set. The scripts are available on our Dryad repository. Briefly, several individuals in our data set had been identified as possible *A. speciosa* or *A. syriaca* x *A. speciosa* hybrids. Since *A. syriaca* and *A. speciosa* can be difficult to distinguish when they are not in flower, we did an initial clustering of our data using the find.clusters function implemented in adegenet 2.1.1[63,64] in R 3.5.2.[65] This identified several more putative *A. speciosa* individuals, which were removed.

Since *A. syriaca* can reproduce asexually, we also screened our data set for clones; i.e., different ramets of the same genet. To do so, we considered all pairs of individuals, calculating what percentage of their loci had identical SNP calls. Across all pairs of

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

individuals, this distribution was bimodal. The vast majority of pairs were normally distributed around a sequence identity of 0.898, with a small number of comparisons clearly outside of this distribution, clustered near 1. Accordingly, we considered all pairs of individuals with a sequence identity greater than 0.96 to be clones. Where clones were found at the same site, we randomly selected a single exemplar, discarding all its clones from the data set. A few pairs of clones were found in different sites; in this case we discarded both members of the pair.

Combining the Broad Range and Core Range GBS Data Sets in subsequent analyses produced strong batch effects between the two data sets (see below), likely because they were sequenced on different machines, at different times, to different read depths. We therefore performed the following analyses separately for the two data sets.

After discarding *A. speciosa*, clones, and individuals for which relatively few loci (i.e., less than 80% of the total number of loci) had been sequenced, we then randomly downsampled the Core Range data set to include a maximum of 5 individuals per site, to homogenize sampling effort across the sites. Finally, we used Stacks to filter SNPs across these individuals, including SNPs with observed heterozygosity less than or equal to 0.6 and present in at least 80% of individuals. Where multiple SNPs were found at the same GBS locus, we randomly excluded all but one. To reduce linkage disequilibrium, we filtered SNPs so that each was at least 50 kb from its nearest neighbor.

We also used this data set, after excluding invasive individuals collected from Europe using vcftools 0.1.15,[66] for demographic modelling. This data set was converted to DIYABC format using vcf2diyabc.py.[67]

Identifying batch effects in GBS data sets

We identified SNPs from the combined Cornell and W&M datasets using the same stacks pipeline described above. This resulted in 872 SNP markers from 181 *A. syriaca* individuals. These markers were then used in a STRUCTURE analysis identical to that described below, with the exception that we only analyzed possible numbers of clusters between K = 2 and K = 10. STRUCTURE results were processed and visualized using the same pipeline described below.

For many values of K, the differences between the STRUCTURE results for the Cornell data set and the W&M data set were subtle: for instance, for K = 2, Cornell individuals had approximately 25-35% ancestry from Cluster 1, while W&M individuals had around 35-45% ancestry from the same cluster. We therefore also used a second clustering method implemented in the adegenet 2.1.2 package[63,64] in R, which uses a K-means approach to assign individuals to one of K clusters, with the appropriate K chosen based on the Bayesian Information Criterion.

Runs with K = 2 and K = 3 produced the two lowest BICs, which were nearly equal. Both runs produced similar results, with the cluster assignments almost exactly mirroring membership in the Cornell or W&M datasets (Table S4). The difference between the two is that at K = 3, some European individuals from the Cornell data set were split off from the remainder of the Cornell individuals.

SNP Calling of the A. syriaca *Broad Range WGR data set*

We called SNPs using the Genome Analysis Toolkit (GATK) pipeline.[68–70] Reads from each individual were mapped against the *A. syriaca* genome using Bowtie2 2.3.2, with an expected range of inter-mate-pair distances of 100-2000 and the "--very-sensitive-local" alignment settings. Indicies of the genome were first built using both bowtie2 and samtools, and a sequence dictionary created using Picard 2.18.15 from the Genome Analysis Toolkit.[68–70]

We further used Picard to fix mate pair information, mark and remove duplicate reads, and replace read group names; we then used samtools to index the alignments for each resequenced individual. We then called polymorphisms for each individual with the HaplotypeCaller tool, then combined the outputs from each scaffold using GenomicsDBImport. We then used GenotypeGVCFs to do joint genotyping on all individuals simultaneously. Indels were removed with the SelectVariants tool, and the remaining SNPs were filtered using the VariantFiltration tool, discarding SNPs for which any of the following were true: quality by depth (QD) less than 2; phred-scaled p-value of Fisher's Exact Test for strand bias (FS) greater than 60; root mean square of the mapping quality (MQ) less than 35; mapping quality rank sum test (MQRankSum) less than -12.5; read position rank sum test (ReadPosRankSum) less than -8. We also filtered out loci with greater than 5% missing data or a minimum read depth of less than 5, as well as removing individual genotypes with a minimum quality 5 or less. Finally, SNPs were thinned to be 50 kb apart or more, so as to match the amount of thinning done for the GBS data set.

SNP Calling of the D. plexippus *WGR data set*

We used the whole genome sequencing data of Zhan et al. to gather genomic data from 29 monarch butterflies collected in North America (which individual specimens we used are given in Table S5; we chose migratory individuals from the continental United States and Mexico, excluding non-migratory individuals from South Florida).[27] We called SNPs using the pipeline described above, aligning reads from each individual to the *D. plexippus* genome of Zhan et al., GenBank accession GCA_000235995.2.[71] SNPs were filtered using the same criteria as for the *A. syriaca* WGR data, except that SNPs were thinned to be one per contig of the *D. plexippus* genome in order to produce a roughly similar number of SNPs to those found in the *A. syriaca* data sets. Average read depth at genotyped SNPs was calculated for each of our datasets and are as follows: Broad Range GBS: 300; Core-range GBS: 217; WGR: 17; *Danaus* from Zhan et al.[27]: 10; *Danaus* from Talla et al.[30]: 12.

Filtering of genotypes from the Talla et al. D. plexippus *dataset*

We used the final set of SNP genotypes used by Talla et al.,[30] available at https://github.com/venta380/Monarch_genomics. From this data set, we chose the 29 Western North American monarch individuals. SNPs were filtered using the same parameters as used for the Zhan et al. monarch data set.[27]

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

QUANTIFICATION AND STATISTICAL ANALYSIS

Population genetic analysis

F_{ST} analysis and basic population genetic statistics

Using all three *A. syriaca* data sets, and the two *D. plexippus* data sets, we estimated several population genetic statistics in R, using the adegenet and hierfstat packages.[63,64,72] We assigned each individual to one of five broad geographic populations based on its location (Figure 1A). Population assignments are shown in Figure 1A. We tested whether this arrangement captured significant genetic structuring using an AMOVA test, using the pegas method[73] as implemented in poppr 2.8.2[74] with 10,000 permutations.

Population genetic statistics for each of the populations are shown in Tables 1 and 2 of the main text. The genetic differentiation of the subpopulations was low, but statistically significant for the GBS data sets (F_{ST} = 0.008 for Broad Range; 0.052 for Core Range; AMOVA $p < 1*10^{-4}$ for both). For the Broad Range WGR data set, genetic differentiation was even lower, and not significant (Fst = -0.002, or effectively zero, AMOVA $p = 0.47$), possibly due to the smaller number of individuals in each population. In the Core Range GBS data set, the greatest pairwise F_{ST} was between the invasive European population and native populations; pairwise F_{ST} was lower between the northeast and southeast populations by a factor of 10. In the Broad Range GBS data set, the greatest pairwise F_{ST} was between the Northwest population and the two eastern populations, although even this was relatively low, at 0.02. Within each dataset, heterozygosity was relatively constant among populations, with the exception that both observed and expected heterozygosity were lower in Europe than in the other populations in the Core Range data set, showing reduced genetic diversity in the invasive range of *A. syriaca*. The *A. syriaca* specimen chosen for genome sequencing was an invasive, European milkweed, on the logic that the invasion process had likely led to more inbreeding than is usual in other *A. syriaca* populations, and the reduced heterozygosity of this population suggests that this was indeed the case. The reduced heterozygosity is beneficial for genome assembly.

STRUCTURE analysis

To examine clustering and admixture within the *A. syriaca* populations, we used STRUCTURE 2.3.4.[75] We analyzed all three data sets using an admixture model within STRUCTURE and all possible values for the number of clusters (*k*) between 1 and 20; running 10 replicates for each *k* value. For each run we did 1 million iterations beginning after an initial burn-in period of 100,000 iterations. We chose the best number of clusters using the Evanno method[29] as implemented in Structure Harvester 0.6.94.[76] We also used Structure Harvester to convert STRUCTURE output files for use with CLUMPP 1.1.2.[77] We used CLUMPP to assign consistent cluster identities across our multiple replicates for each *k* value above 1, using the LargeKGreedy algorithm with 1000 random input orders and the G' matrix similarity statistic.

PCA analysis

To complement our STRUCTURE analysis, we also performed a PCA analysis to examine geographic distribution of genetic structure in a less parameterized way using the ade4[78,79] and adegenet[63,64] packages in R. We first scaled each genotype using the scaleGen() function, replacing missing data with the mean allele frequency for that SNP, and then performed a Principle Components Analysis on these scaled allele frequencies.

Applying the Evanno method to our STRUCTURE results resulted in an optimal number of $k = 5$ (Figure S2) for the Core Range Data Set. Examination of the STRUCTURE results shows a very similar pattern for all values between $k = 2$ and $k = 5$: a single cluster dominates all individuals from North America, and a second cluster is found in a number of invasive *A. syriaca* collected from Europe (Figure S2). Other clusters, when present, account for very little of the ancestry of any *A. syriaca* specimens. For the Broad Range data sets, the Evanno method selected $k = 11$ for the GBS data set and $k = 2$ for the WGR data set (Figure S2). However, the Evanno method is unable to consider $k = 1$ as the best cluster, since it uses changes in the likelihood of the data between $k = x$ and $k = x-1$. Visualizing the cluster results showed patterns in which each genetic cluster was found in every individual to a similar extent, which suggests that there is minimal geographic structuring within the Broad Range data set (Figures S2).

Demographic modelling

We next used all five data sets (3 milkweed and 2 monarch) to estimate the recent demographic history of the two species. To investigate the recent demographic history of monarchs and common milkweed, we used an ABC-RF algorithm for model selection and parameter estimation.

As our observed data, we used the five monarch and milkweed data sets described above. Guided by the results of our STRUCTURE analysis, we treated *A. syriaca* as a single population. We simulated data sets using DIYABC 2.1.0[80] to test the following hypotheses (visualized in Figure 2):

1. Have *A. syriaca* populations experienced a bottleneck within past decades, potentially due to the increased use of herbicide in crop fields as described by, e.g., Pleasants?[4,25]
2. Have *A. syriaca* populations expanded in the past centuries, potentially due to the conversion of native forests and prairies to agriculture land, as suggest by, e.g., Brower?[11]
3. Have *A. syriaca* populations expanded in prior millennia, potentially due to the retreat of the glaciers after the last glacial maximum?[23]

Considering every possible combination of the three hypotheses produced 8 demographic scenarios (visualized in Figure 2). We used DIYABC to simulate 80,000 data sets across all 8 demographic scenarios. For each scenario, population sizes were selected

Please cite this article in press as: Boyle et al., Temporal matches between monarch butterfly and milkweed population changes over the past 25,000 years, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.07.057

from uninformative prior distributions, while event times were chosen from uniform distributions. We chose event times to correspond to 1945-2015 for the recent bottleneck, 1751-1899 for the recent expansion, and 5-25 thousand years ago for the ancient expansion. *A. syriaca* plants flower in their second growing season,[81] so we assumed a 2 year generation time for this species. *D. plexippus* has 4-5 generations per year, so we assumed a 0.2-0.25 year generation time for that species. We outputted all 4 summary statistics calculated by DIYABC, which would be used for ABC-RF model selection, alongside the linear discriminant axes that were the combinations of those summary statistics that best distinguished the demographic models (one variable, ''Proportion of zero values'', was invariant across our simulations since only variable SNPs were used; this variable was not used in the following analyses). We repeated this process 20 additional times, producing a total of 105 simulation sets, 21 for each of our three milkweed and two monarch data sets.

Following Pudlo et al.,[12] and using the abcrf package in R, we performed a number of validations of our ABC-RF approach: we first tested the compatibility of our models with our observed data by projecting our observed data, along with the simulations, along the linear discriminant (LD) axes that best distinguished the 8 models given the set of summary statistics (Figure S2).[12,15] We then constructed a random forest of 1000 decision trees, each of which provided a prediction of which demographic model produced a given set of summary statistics. To test whether we had produced a sufficient number of simulations, we compared the error rate of this random forest to that of a second random forest constructed using only 80% of the 80,000 simulations. Finally, to test whether 1000 decision trees was a sufficient number, we calculated the prior error rate using forests of different size, from 40-1000 (Table S1).

Preliminary analyses showed that using the default settings for constructing the random forest produced substantial overfitting, so based on these analyses we reduced the maximum depth of each tree in the forest to 8 (for random forests to determine the overall model) or 16 (for random forests to determine the presence of a single demographic event) to minimize overfitting (results not shown).

For each of our three milkweed and the two monarch data sets, we then produced 20 different random forests using 20 different simulation sets. For each random forest, we then measured its accuracy in predicting the training data set used to produce the random forest. We also measured its accuracy in predicting the 21st data set, which was our testing data set, to ensure that training and testing accuracy were similar (i.e., the model was not overfitting our data) (Table S2).

We then fed our observed data set into these 20 random forests in order to estimate the best model and approximate its posterior probability. Because the posterior probability of any single model was low, we followed the same procedure to produce separate random forests to approximate posterior probabilities for each of the three hypotheses listed above, i.e., by grouping together all models that had a recent bottleneck vs all models that did not, etc.

We then used the approach of Raynal et al., employing the ABC-RF approach to estimate parameter values.[15] We first used DIYABC to simulate 10,000 data sets for the single best demographic scenario. We then used this simulation set to estimate posterior medians and quantiles of a number of demographic parameters using ABC-RF with a maximum tree depth of 8.