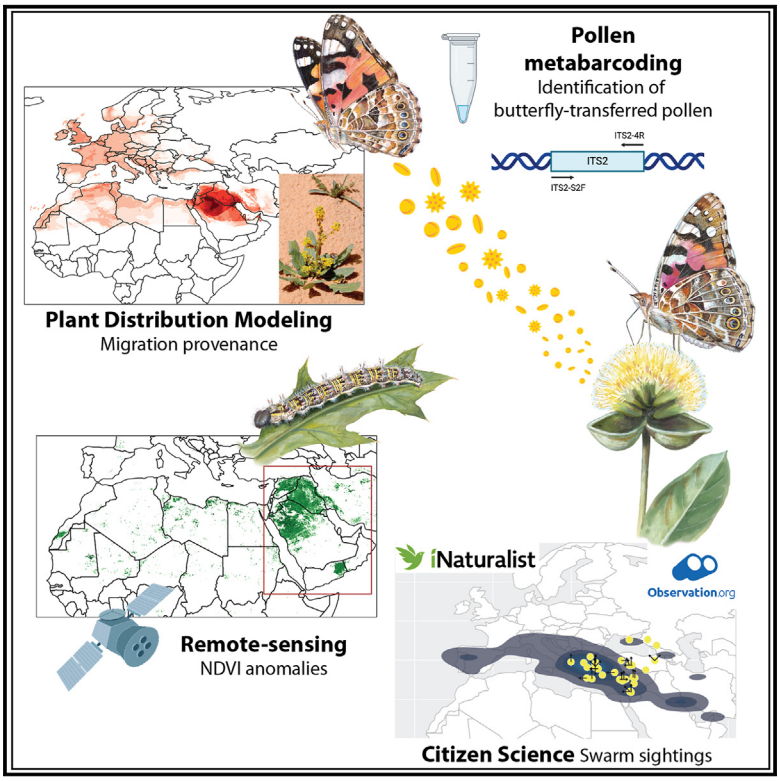
Article

Current Biology

Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak

Graphical abstract



Authors

Johanna Luise Gorki,
Roger López-Mañas, Llorenç Sáez, ...,
Roger Vila, Tomasz Suchan,
Gerard Talavera

Correspondence

gerard.talavera@csic.es

In brief

Gorki et al. trace the migrations of painted lady butterfly (*Vanessa cardui*) population explosions during 2019. Anomalous vegetation growth and citizen science pinpoint the Middle East as the outbreak origin. DNA sequencing of transported pollen grains reveals long-distance movements across generations over the Middle East, Europe, and Africa.

Highlights

- Citizen science reports explosions of the butterfly *Vanessa cardui* in 2019
- Metabarcoding of butterfly-transferred pollen identifies non-native plant species
- Plant distribution modeling infers butterfly migration and the outbreak's origin
- Middle Eastern outbreak impacts Europe and Africa demographics

Gorki et al., 2024, Current Biology *34*, 1–9
June 17, 2024
https://doi.org/10.1016/j.cub.2024.05.037

CellPress

Current Biology

CellPress

Article

Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak

Johanna Luise Gorki,[1,14] Roger López-Mañas,[1,2,14] Llorenç Sáez,[2,3] Mattia Menchetti,[4] Nazar Shapoval,[5] Anne Andersen,[6] Dubi Benyamini,[7] Steve Daniels,[8] Aurora García-Berro,[1] Megan S. Reich,[9] Stefano Scalercio,[10] Eric Toro-Delgado,[1,4] Clément P. Bataille,[9,11] Cristina Domingo-Marimon,[12] Roger Vila,[4] Tomasz Suchan,[13] and Gerard Talavera[1,15,*]

[1]Institut Botànic de Barcelona (IBB), CSIC-CMCNB, Barcelona 08038 Catalonia, Spain
[2]Departament de Biologia Animal, Biologia Vegetal i Ecologia (BABVE), Universitat Autònoma de Barcelona, ES-08193 Bellaterra, Catalonia, Spain
[3]Systematics and Evolution of Vascular Plants (UAB)—Associated Unit to CSIC (IBB), Bellaterra, Spain
[4]Institut de Biologia Evolutiva (CSIC-Univ. Pompeu Fabra), 08003 Barcelona Catalonia, Spain
[5]Department of Karyosystematics, Zoological Institute, Russian Academy of Sciences, Russia, 199034 Saint-Petersburg, Russia
[6]Entomological Society of Denmark, Zoological Museum, Universitetsparken 15, DK-2100 Copenhagen, Denmark
[7]The Israeli Lepidopterist Society, Beit Arye 7194700, Israel
[8]Independent Naturalist, Cockermouth CA13 0JJ, UK
[9]Department of Biology, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[10]Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria, Centro di Ricerca Foreste e Legno, 87036 Rende, Italy
[11]Department of Earth and Environmental Sciences, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[12]Center for Ecological Research and Forestry Applications (CREAF), Grumets Research Group, Cerdanyola del Vallès, 08193 Catalonia, Spain
[13]W. Szafer Institute of Botany, Polish Academy of Sciences, 31-512 Kraków, Poland
[14]These authors contributed equally
[15]Lead contact
*Correspondence: gerard.talavera@csic.es
https://doi.org/10.1016/j.cub.2024.05.037

SUMMARY

Migratory insects may move in large numbers, even surpassing migratory vertebrates in biomass. Long-distance migratory insects complete annual cycles through multiple generations, with each generation's reproductive success linked to the resources available at different breeding grounds. Climatic anomalies in these grounds are presumed to trigger rapid population outbreaks. Here, we infer the origin and track the multigenerational path of a remarkable outbreak of painted lady (*Vanessa cardui*) butterflies that took place at an intercontinental scale in Europe, the Middle East, and Africa from March 2019 to November 2019. Using metabarcoding, we identified pollen transported by 264 butterflies captured in 10 countries over 7 months and modeled the distribution of the 398 plants detected. The analysis showed that swarms collected in Eastern Europe in early spring originated in Arabia and the Middle East, coinciding with a positive anomaly in vegetation growth in the region from November 2018 to April 2019. From there, the swarms advanced to Northern Europe during late spring, followed by an early reversal toward southwestern Europe in summer. The pollen-based evidence matched spatiotemporal abundance peaks revealed by citizen science, which also suggested an echo effect of the outbreak in West Africa during September–November. Our results show that population outbreaks in a part of species' migratory ranges may disseminate demographic effects across multiple generations in a wide geographic area. This study represents an unprecedented effort to track a continuous multigenerational insect migration on an intercontinental scale.

INTRODUCTION

Herbivorous insect populations are inherently limited by the bottom-up pressures exerted by host plant availability.[1–3] In arid and semi-arid regions experiencing strong interannual variation in precipitation, low and highly seasonal rainfall limits plant productivity. However, when unusually high rainfall occurs, a sudden rise in the quantity and quality of host plants may trigger an outbreak in insect populations by exceeding the top-down pressures exerted by natural-enemy control.[4,5] Insect outbreak events are not fully understood mechanistically but are known to have significant impacts on ecosystem functioning, with consequences for nutrient cycles, pathogen dynamics, and plant health.[4–8]

In the case of migratory insects,[9–12] locally outbreaking populations can spread beyond their epicenter, affecting distant

Please cite this article in press as: Gorki et al., Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.05.037

ecosystems.[13,14] However, our ability to determine the full extent of the impacts of migratory insect outbreaks is hindered by logistical and technical hurdles. First, since many insect species migrate over multiple overlapping generations, studies are often limited by the difficulties of collecting observational data and samples across vast geographical areas and extended periods, which is necessary to cover each segment of a multigenerational migratory cycle. Second, observational data alone do not inform about the complex spatiotemporal linkages of insect movement, and there are few techniques available to gather empirical connectivity data. These limitations have resulted in few attempts to link observational data with evidence of provenance across succeeding generations of insects. Studies inferring provenance, such as those using stable isotopes, typically sample only a few discrete points in time and space. While valuable to understand single-leg movements, this approach does not disclose connectivity along multigenerational migrations. Only collecting samples and provenance evidence from the entire spatiotemporal expanse of the multigenerational migratory cycle can offer a complete picture of migratory connectivity.

To address these challenges in the field of movement ecology, we studied an outbreak event of the painted lady butterfly (*Vanessa cardui*), which covers the greatest distance among insects during its annual migratory cycle.[15–18] In 2019, huge densities of *V. cardui* were observed in various regions worldwide. A citizen science initiative allowed us to collect observational data and samples from large geographic (10 countries) and temporal (7 months) scales. To address the need for connectivity data, we developed a cost-effective method that combines information from metabarcoding of butterfly-transferred pollen grains[19] and spatial analyses of plant distributions to provide evidence of provenance. This approach provided an exceptional opportunity to investigate the geographical extent and ecological implications of a wide-ranging intercontinental outbreak event for a migratory species.

RESULTS AND DISCUSSION

The origin of the 2019 *Vanessa cardui* outbreak

The painted lady butterfly undergoes a cyclic multigenerational migration between tropical Africa and northern latitudes in the Palearctic in 8 to 10 generations every year.[15–18] Consequently, each generation completes only a segment of this extensive journey. In March 2019, massive swarms of this butterfly were initially observed in parts of the Middle East and Eastern Mediterranean (EMed) (Figure 1A). Their massive presence was progressively noticed in Eastern Europe, Northern Europe, and Western Europe in the following months (Figures 1A and S1; Data S1). We collected 264 butterflies from swarms across 10 countries from March to November, which were clustered into 16 spatiotemporal groups for spatial analyses (Figure 1B; Data S1). To investigate their provenance, we sequenced the pollen found on the bodies of the butterflies through metabarcoding of the internal transcribed spacer 2 (ITS2); we molecularly identified 398 plants and then modeled the species distributions of each plant (Data S1 and S2). Plants identified that were not naturally occurring in the regions where the butterflies were collected were used for provenance inferences, as their flowers could only be visited in remote locations along their migration path. Thus, only regionally absent plants could be informative about movement.

From Arabia and the Middle East to the Eastern Mediterranean

Our results consistently point to a breeding outbreak originating in northern Arabia and the Middle East (AME). Modeled plant distributions from swarms collected in the EMed (Chios and Crete, Israel, and Syria) in March and April all produce high probabilities of provenance in AME (Figures 2A and 3A; Data S3). Depending on the stringency of the filtering criteria, we identified between 11 and 36 plant species (39.3%–41.8% of the total) that were not native to the sampling regions in the EMed (Data S1). Most of these alien plant species are primarily distributed in, and some are endemic to, AME (Figure 2A; Data S1 and S2).

In order to find a mechanistic understanding of the source of the observed outbreak, we performed a time series analysis of vegetation growth covering the period of 2001 to 2021, using the normalized difference vegetation index (NDVI), NASA's MOD13C2 remote-sensing product, at a resolution of 0.05°. The results revealed a highly anomalous vegetation growth event in the AME region from December 2018 to April 2019 (Figure 2B). This anomalous growth event covered an area of at least 151,186 km^2 (in January) and as much as 120,385 km^2 (in March). It was predominantly detected in semi-arid regions of Saudi Arabia, Iraq, Syria, and Yemen. The ecoregions most affected by this atypical positive growth (with NDVI peak in the 0.05 upper distribution) were sparse shrub vegetation zones, such as the North and Red Sea Arabian desert (peak NDVI > 0.1915, 0.0716 above expectation), the Syrian xeric grasslands and shrublands (peak NDVI > 0.4036, 0.2339 above expectation), and the Tigris-Euphrates alluvial salt marsh (peak NDVI > 0.2872, 0.1120 above expectation). This phenomenon excluded other AME ecoregions, such as sand deserts, which experienced a typical growth season (Figure S2).

Pollen metabarcoding and vegetation growth anomaly analyses provided independent but strongly consistent signals, all indicating AME as an exceptional breeding hotspot and the original population peak for the widespread outbreak of *V. cardui* in 2019 (Figure 2). We argue that the ecoregions experiencing this anomalous vegetation growth served as extensive breeding areas for a period of at least 3 months.

Parallel non-outbreaking migration in Western Europe

The pollen-based geographic signal from migrating butterflies collected in the Western Mediterranean (WMed) during March and April highlights different geographical regions compared to the specimens collected in the EMed around the same period. The WMed individuals showed an origin more consistent with Africa (Data S3), suggesting that these migrants were independent of the AME outbreak. The presence, for example, of *Tetraena simplex* pollen in samples from Andalusia (March) could indicate a trans-Saharan passage,[16,19] given that this halophytic annual plant grows in the Sahara and is especially abundant along the West Saharan coast. Moreover, the pollen-based geographic signal for samples in Catalonia (April/May) indicates strong probabilities of provenance from coastal Algeria/Morocco and southern Spain (Figure 3A; Data S3), aligned with the anticipated movements of *V. cardui* during a typical year at that stage of

Please cite this article in press as: Gorki et al., Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.05.037

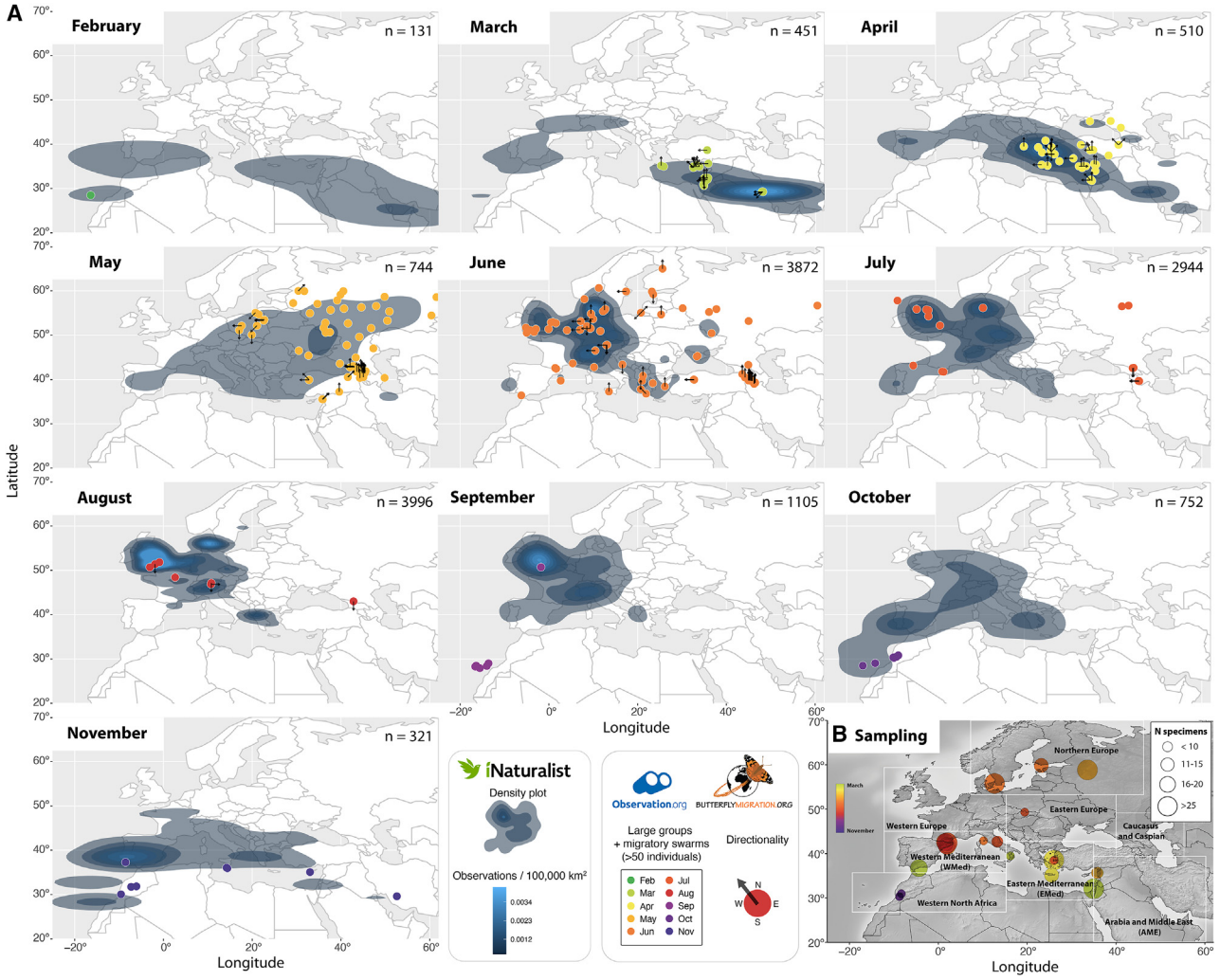


Figure 1. Observation-based patterns of *Vanessa cardui* in 2019 obtained from citizen science

(A) Monthly density plots (number of observations/100,000 km^2) indicate relative abundances throughout the studied region based on observations retrieved from iNaturalist between February and November 2019. Density plots were inferred using the *stat_density2d()* function in the R package ggplot2. Colored circles indicate observations of migratory swarms and large numbers of *V. cardui* butterflies, which were compiled from diverse sources, including citizen science programs (e.g., https://observation.org/, http://butterflymigration.org/), social networks, and field trips (Data S1). Where available, directional information of swarms is depicted in the circles using arrows. The temporal sequence of observations reveals high densities and mass reports in AME (Arabia and Middle East) starting in March, which gradually spread toward Northern Europe and extended farther to Western Europe (May to September) and southward to Africa (October and November). Collectively, these observations provide valuable insights into the migratory patterns of 2019's outbreak. See also Figures S1 and S3 and Data S1.

(B) Sampling map displaying the collection sites (colored circles) for 264 butterflies used in the pollen metabarcoding analyses. Samples were obtained from 16 spatiotemporal groups across 10 countries from March to November 2019. The relative number of specimens per group is indicated by the size of the circles, and the color of the circles indicates the collection months. Geographic regions as used in the text are delimited by white boxes. See also Data S1.

the migratory cycle.[14–19] However, many of the identified plants in these two spatiotemporal migrant groups collected from March to May have wide-ranging distributions across North Africa and AME, and probabilities from AME remained high even from WMed groups (Data S3). Samples collected in Calabria (Italy) showed a more prominent mixed signal, indicating potential origins either from AME or from unrelated migratory groups from North and/or sub-Saharan Africa. To our knowledge, no swarms comparable to those observed in the EMed and AME regions (in the order of millions of butterflies) (Data S1) were observed in southern Italy during April/March, and the numbers were more consistent with what is expected in a regular migratory season. Still, it is difficult to rule out a slight influx of specimens originating in AME into the WMed only based on demographic observations, as it is expected that swarms diffuse as they move farther away from their outbreaking grounds. However, no pollen from endemic plants from AME was identified among the WMed butterflies (including those from Calabria), as was the case with the EMed spatiotemporal groups, so we are inclined to discard a direct link.

Please cite this article in press as: Gorki et al., Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.05.037

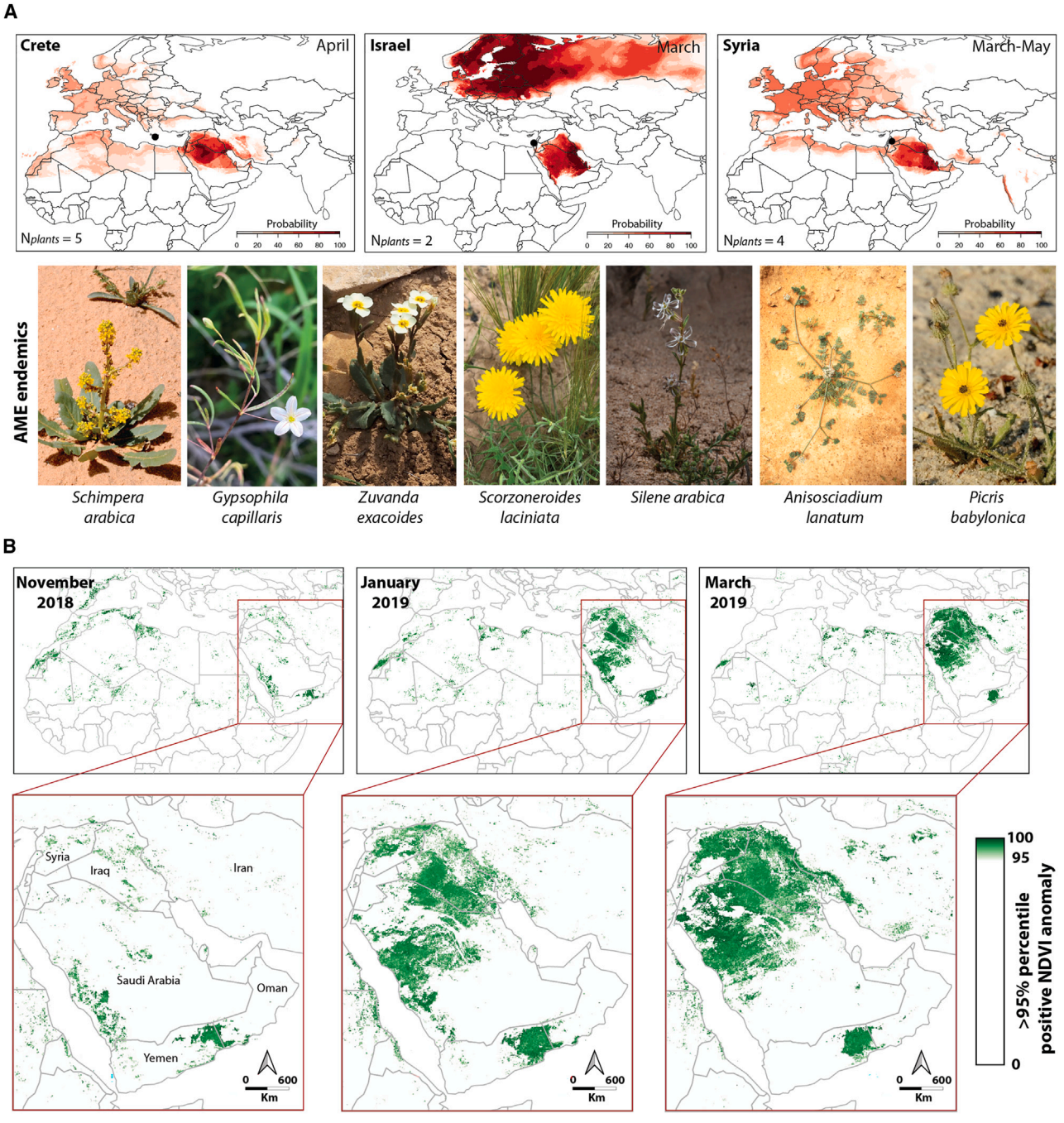


Figure 2. The outbreak source in the Arabian and Middle East region

(A) Pollen-based geolocation probabilities for three spatiotemporal groups of *V. cardui* (filtering for sequences with ≥98% taxonomy prediction confidence and ≥10 reads per match) collected in the Eastern Mediterranean (EMed) and Arabian and Middle East (AME) regions during spring. The analyses were based on identified plants absent in the collecting region/country. Only regionally absent plants can provide reliable information for inferring provenance, as their flowers can only be visited in remote locations. Images illustrate examples of endemic plants from AME identified through the pollen metabarcoding analyses, contributing to the geolocation signal from this region. See also Data S2 and S3.

(B) Monthly NDVI anomalies from a time series analysis (2001–2020) exhibit highly anomalous vegetation growth in the AME between November 2018 and April 2019. Values ≥95 can be considered extreme positive anomalies, based on the kernel density estimation (KDE), a non-parametric approach implemented in the R package npphen. Anomalous values are relative to the phenological baseline of each pixel and thus do not represent total signal of NDVI comparable between pixels. Vegetation anomalies coincide with the sightings of extraordinary *V. cardui* abundances and swarming movements in the region during the same period (Figure 1). See also Figure S2.

Please cite this article in press as: Gorki et al., Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.05.037

CellPress

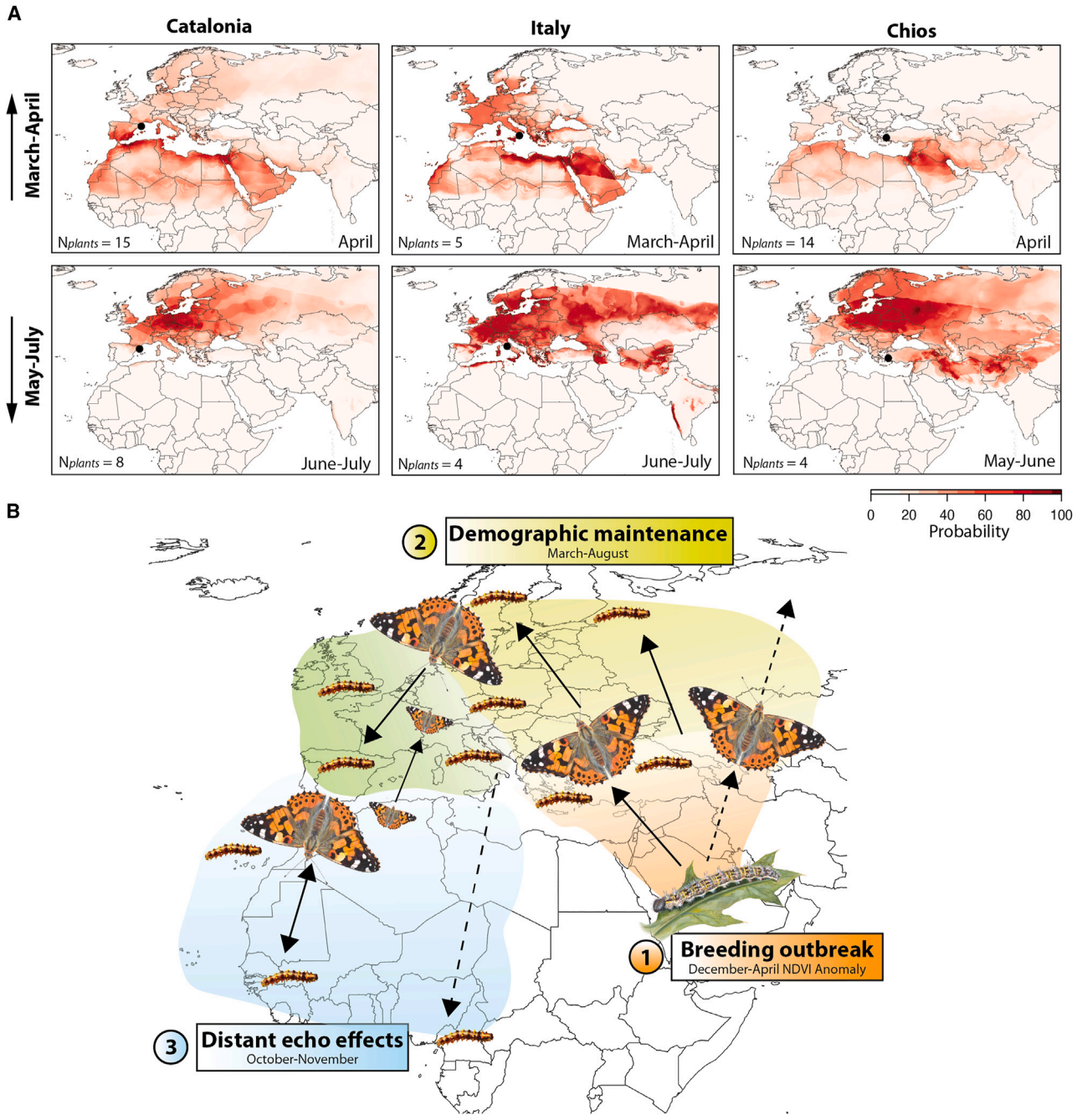


Figure 3. Provenance geolocation shows geographical shifts over generations

(A) Pollen-based geolocation probabilities for six spatiotemporal groups of *V. cardui* butterflies (98% sequence match; ≥10 reads per match). Three spatial groups in the Mediterranean (Catalonia, Italy, and Chios) at two different temporal frames (March–April and May–July) reveal seasonal contrasts in the inferred geolocation probabilities. Butterflies sampled in March–April (upper maps) showed higher probabilities of provenance from Africa, the Mediterranean basin, or AME. In contrast, butterflies sampled in late May to July (bottom maps) showed higher probabilities of provenance from farther north in Europe. See also Data S3.

(B) A summary map of the 2019 movements of *V. cardui* butterflies illustrates the transmission effect of an exceptional breeding outbreak occurring in the AME region during early spring, due to an atypical vegetation growth event confirmed by satellite. A multigenerational northward migration from AME gradually spread into Eastern Europe and Central Asia, reaching regions of Northern Europe as early as in May. A subsequent reversal of direction led to invasions in Western and Southwestern Europe, with migratory movements echoing as far as the Canary Islands and sub-Saharan Africa from September to November. The schematic map also indicates other migratory flows at relatively lower numbers that initiated in other regions different from AME, as is expected in a regular migratory season. Dashed arrows represent inferred movements based only on citizen science and monitoring observations (i.e., not studied through pollen metabarcoding).

Please cite this article in press as: Gorki et al., Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.05.037

Trans-generational transmission of the outbreak

From Eastern to Northern Europe

Starting at the end of May and peaking in June, massive swarms were observed in Northern Europe, including Russia, Scandinavia, and the Baltic countries (Figures 1A and S1; Data S1). These occurrences were consistently preceded by swarming observations in Eastern European countries. For instance, at the ornithological monitoring station in the Courish spit in Kaliningrad (Russia), the first wave of migrants was captured in a funnel trap heading north, and thus entering the Baltic Sea, between the 19th and 22nd of May. The pollen-based geographic signal for specimens collected in Northern Europe (i.e., Russia in May; Finland and Denmark in June) (Data S3) shows high probabilities of origin from European Mediterranean countries but also for Western Europe (e.g., France). However, the swarms observed in Northern Europe would not align with the relatively low abundances reported in the previous month in Western Europe. The abundances observed in Western Europe in May are consistent with a regular migratory season and likely originate from the anticipated African routes.[16–18] Therefore, we argue that the swarms observed in Northern Europe in May and June likely stemmed from the first AME immigrants' offspring. Indeed, for spatio-temporal groups in Russia, Finland, and Denmark, a substantial number of the identified plant taxa are of Central Asian distribution (13 taxa), particularly representative of the Caucasus and Caspian region. This region likely served as a corridor for the massive migration associated with the original AME outbreak, as supported by the striking abundances observed in Georgia or in the Russian Caucasus provinces in late April and May (Figures 1, S1, and S3; Data S1).

From Northern Europe to Western Europe

In late May and June, southward movements toward Western and Mediterranean Europe were detected by citizen science, followed by a rapid demographic increase in these regions. Simultaneously, a massive invasion persisted in Scandinavia, Russia, and Eastern Europe (Figures 1, S1, and S3; Data S1). In Western Europe, where *V. cardui* sightings had been relatively scarce, their abundance rapidly increased (Figures 1 and S1; Data S1). Interestingly, the pollen-based geographic signal for the spatio-temporal groups in the Mediterranean from late May to July (Catalonia, Pyrenees, Italy, and Chios) showed a dominant contribution of northern European plants among the identified pool of pollen grains (Figure 3A; Data S3). This pattern differed from the probabilities inferred for March and April in the same regions, when the dominant contribution was from Africa and/or AME (Figure 3A; Data S3).

On a specific day (June 8), 1,190 individuals were captured heading southwest in a funnel trap at the Courish spit (Kaliningrad, Russia), while only 31 individuals were captured in a trap heading northeast at the same site. This difference strongly suggests a significant reversal in migration directionality. Such reversals are part of the *V. cardui*'s annual migratory cycle, but in 2019 these reversals started to occur about 2 months earlier than usual. The factors driving directional turns from one generation to the next in multigenerational migrations are not fully understood, but shifting photoperiod, adverse climatic conditions, high population densities, and a lack of resources are likely to play important roles.[20–24]

As sustained northward migrations brought extraordinary numbers of butterflies to Northern Europe early in the season (as early as May), it is likely that the region was unable to sustain extensive breeding during this period. This could be attributed to suboptimal temperatures and/or a mismatch in arrival timing with host plant phenology. Although one or two generations did breed in Northern Europe later in the season between June and August, resulting in an extraordinary number of offspring for the region, a portion of the population that arrived in early May likely reversed direction early on, leading to the production of new generations all over the rest of Europe. As a result, the duration of southward movements was prolonged compared to a typical year, contributing to the high abundances observed throughout Western Europe in June, July, and August. Abundance peaks were particularly noticeable in Western Europe during that period, including in the UK, France, and Spain. Unexpected and sudden population explosions were observed in Catalonia between June 25 and July 6, in Andalusia between June 25 and July 17, and in the UK between July 28 and July 30 (Table S2). These population explosions likely represent the offspring of multiple breeding groups spreading the initial AME outbreak across all of Western Europe. However, it is probable that these groups overlapped with the offspring of unrelated migrant waves dispersing northward from Western Africa and the WMed, resulting in a mixture of northward and southward migrants.

From Western Europe to Africa

The transmission of the outbreak extended well beyond the summer, persisting throughout the fall. Between September 23 and 30, the outbreak extended to the Canary Islands with abundances far exceeding those observed in typical years (Figures 1A and S1; Data S1). In late October and early November, *V. cardui* became extraordinarily dominant in southern Morocco (Data S1). Our pollen-based geographic signal for Morocco during this period displayed a mixed probability pattern of origin, with significant probabilities from Western Europe, but even stronger probabilities originating from the Sahel (Data S3). This robust signal is primarily attributed to the presence of the Sahelian endemic plant *Guiera senegalensis*. Immigrants from both regions are expected during that time.[18,25,26] The latest generations of migrants from Europe might colonize African grounds on both the northern and southern sides of the Sahara, while the earliest European migrants could have already produced a new generation in the Sahel, blooming in September.[26] Some of these Sahelian offspring might have migrated back to North Africa, while others would continue their migration south toward more tropical latitudes, as is typical in their annual migratory cycle.[17,18] Notably, monitored abundances of *V. cardui* in the fall of 2019 were larger than usual as far south as the Cameroonian highlands.[18] Overall, these findings indicate that the impacts of the AME outbreak echoed throughout most of the annual cycle, reaching as far as sub-Saharan West Africa and spanning at least 9 months.

Implications of outbreaks for demography, ecology, and migratory cycles

Our pollen-based geolocation data across multiple sampling points uncover several details of the 2019 *V. cardui* outbreak: the outbreak originated in the AME region and produced

subsequent migratory movements of northwest (NW) tendency to Northern Europe, followed by reversed movements with a southwest (SW) tendency toward Western Europe and West Africa (Figure 3B). Although our sampling was primarily limited to Europe, observational evidence suggests that the original movements also encompassed directions toward the northeast, impacting regions throughout Central Asia and Siberia[27] (Figure S3). This suggests that migratory movements in insects do not necessarily adhere to strict directionality when considering the whole population; instead, there is a continuum of directionality with an overall latitudinal progression.

The *V. cardui* outbreak in 2019 contrasts with another massive outbreak in 2009 primarily affecting Western Europe, which likely originated in the Maghreb, also as a result of highly anomalous vegetation growth.[14] That two important outbreaks, originating in regions as different as North Africa and the Middle East, had such a large impact on *V. cardui* population sizes across Europe reinforces existing evidence that *V. cardui* has a huge migratory range in the Afro-Palearctic and should be considered a single panmictic population.[26,28]

Our results demonstrate that for migratory insects, outbreaks occurring in relatively small regions (and likely involving the offspring of relatively few individuals) can trigger a demographic uprise with a cascade effect over multiple generations and across vast geographic ranges spanning continents. Overall, when interpreting time series demographic dynamics of migratory insects, it is important to differentiate between regular fluctuations and outbreaks (or bottlenecks), the latter largely exceeding the former. Regular fluctuations in population size are expected within the cyclic multigenerational steps of seasonal migrant species, with breeding origins predictable by seasonality patterns. However, the breeding origins of massive outbreaks may lead to spatial patterns of high numbers of individuals heavily skewed toward specific areas, as observed in the 2019 *V. cardui* outbreak and other outbreaking events in Europe over the last 30 years.[14] As movements resulting from massive outbreaks are more noticeable by observers, there is a risk of overlooking regular movement dynamics in other parts of the range. Anomalous events can alter typical movement dynamics for the season and should not be used to explain the norm of migratory cycles or routes in terms of phenology and directionality.[14] For example, the early reverse in directionality observed in Northern Europe for *V. cardui* led to an atypical demographic peak in Western Europe in June, which could have been misinterpreted as a late massive inflow of southern origin.

Finally, our study makes an important empirical contribution by quantifying the ecological implications of long-range insect migration. We demonstrate that long-distance migratory insects carry substantial pollen loads, and that long-distance insect-mediated pollination may occur at both intra- and intercontinental scales. This phenomenon can have far-reaching implications for essential ecosystem services and agriculture[29] and may influence processes of hybridization and adaptation in plants. Insect migration episodes are not isolated events but represent cyclic patterns, sometimes greatly enhanced by anomalous climatic events leading to outbreaks. While we reported detailed seasonal movements for *V. cardui* in 2019, this outbreak was not unique to this species, and massive swarming of many other insects was detected in parallel movements, likely from similar sources.[30] More efforts need to be devoted to systematically monitor insect migration, and pollen metabarcoding postulates as a powerful tool to track insect movements at large scales. As knowledge of plant distribution, flowering phenology, and plant DNA barcoding libraries improves, pollen metabarcoding will become increasingly informative and holds great potential in conjunction with other expanding techniques in tracking insect migration, such as isotope geolocation.[31]

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 SUBJECT DETAILS
- METHOD DETAILS
  - Swarming observation, sample collection, and spatio-temporal groups
  - Pollen isolation and metabarcoding library preparation
  - Pollen isolation
  - ITS2 amplification
  - Library preparation
  - Bioinformatic pipeline for molecular data analyses
  - Plant taxonomy and dataset curation
  - Spatial analyses: Plant species distribution models and geolocation
  - Pollen metabarcoding and plant distributions: Signal-to-noise ratio counteraction
  - Vegetation growth anomalies
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2024.05.037.

ACKNOWLEDGMENTS

We are grateful to our colleagues who contributed samples used in this study, including Rusland Arbatov, Fulvia Benetello, Elana Erez, David Glazner, Naomi Hominer, Galina Kuftina, Lily Mondeka, Vered Naor, Liat Levinson, Shlomit Lifshitz, Talya Oron, Mudar Salimeh, Benny Shalmon, Dani Sharon, Dror Soari, Mike Taylor, Ofir Tomer, Irit Tzuk-Kovachi, and Esti Volf. We are also grateful to our colleagues who contributed important observational data, including Niclas Backström, Farid Bahleman, Magda Bou Dagher Kharrat, Jurgen Couckuyt, Sylvain Cuvelier, Vlad Dincă, Mohammad J. Ghasempoor, Martin Gibson, Fiona Hanks, Alan Hill, Mark Johnston, Khadim Kébé, Mohammad A.J. Marafi, Dulce Plasencia, Susana Saavedra, and Valery N.K. Talla. Butterfly illustrations were created by Blanca Martí. Plant image credits: Ori Fragman-Sapir, Elena Kuraeva, Ron Frumkin, Arya Aras, Alexey Sergeev, Sami Youssef, and Irvin Sabastian. This work was funded by the National Geographic Society (grant WW1-300R-18), grant LINKA20399 from the CSIC iLink program, and grant PID2020-117739GA-I00 MCIN/AEI /10.13039/501100011033 to G.T.; grant NFRE-2018-00738 from the New Frontiers in Research Fund (Government of Canada) to C.P.B. and G.T.; and project PID2022-139689NB-I00 (funded by MCIN/AEI/10.13039/501100011033 and ERDF A way of making Europe) to R.V. Financial support was provided to N.S. by the State Research Project (no. 122031100272-3) and a grant from the Ministry of Science and Higher Education of the Russian Federation (no. 075- 15-2021-1069). A.G.-B. was supported by fellowship FPU19/01593 from the Spanish Ministry of Universities. E.T.-D. was supported by fellowship FPU22/02358 from the Spanish Ministry of Science, Innovation and Universities. T.S. was supported by the statutory funds of the W. Szafer Institute of Botany of the Polish Academy of Sciences.

G.T., C.D.-M., and R.V. acknowledge support from the Departament de Recerca i Universitats, Generalitat de Catalunya, with grants 2021-SGR-01334, 2021-SGR-00420, and 2021-SGR-00554, respectively.

AUTHOR CONTRIBUTIONS

Conceptualization, J.L.G., R.L.-M., L.S., T.S., and G.T.; field sampling, M.M., N.S., A.A., D.B., S.D., A.G.-B., M.S.R., S.S., M.T., E.T.-D., and G.T.; data collection, J.L.G., L.S., R.L.-M., M.M., N.S., R.V., T.S., and G.T.; data analysis, J.L.G., R.L.-M., C.D.-M., T.S., and G.T.; writing – original draft, J.L.G., R.L.-M., and G.T.; writing – review & editing, J.L.G., R.L.-M., L.S., M.M., N.S., A.A., D.B., S.D., A.G.-B., M.S.R., S.S., E.T.-D., C.P.B., C.D.-M., R.V., T.S., and G.T.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 11, 2023
Revised: March 18, 2024
Accepted: May 20, 2024
Published: June 6, 2024

REFERENCES

1. White, T.C.R. (1978). The importance of a relative shortage of food in animal ecology. Oecologia *33*, 71–86.
2. Hunter, M.D., and Price, P.W. (1992). Playing chutes and ladders: heterogeneity and the relative roles of bottom-up and top-down forces in natural communities. Ecology *73*, 724–732.
3. Power, M.E. (1992). Top-down and bottom-up forces in food webs: do plants have primacy. Ecology *73*, 733–746.
4. Crawley, M. (1989). Insect herbivores and plant population dynamics. Annu. Rev. Entomol. *34*, 531–564.
5. Meserve, P.L., Kelt, D.A., Milstead, W.B., and Gutiérrez, J.R. (2003). Thirteen years of shifting top-down and bottom-up control. Bioscience *53*, 633–646.
6. Barbosa, P., and Schultz, J.C. (1987). Insect Outbreaks (Academic Press).
7. Wallner, W. (1987). Factors affecting insect population dynamics: differences between outbreak and non-outbreak species. Annu. Rev. Entomol. *32*, 317–340.
8. Berryman, A.A. (2008). Outbreaks of Insects. In Encyclopedia of Entomology, J.L. Capinera, ed. (Springer).
9. Holland, R.A., Wikelski, M., and Wilcove, D.S. (2006). How and why do insects migrate? Science *313*, 794–796.
10. Hu, G., Lim, K.S., Horvitz, N., Clark, S.J., Reynolds, D.R., Sapir, N., and Chapman, J.W. (2016). Mass seasonal bioflows of high-flying insect migrants. Science *354*, 1584–1587.
11. Satterfield, D.A., Sillett, T.S., Chapman, J.W., Altizer, S., and Marra, P.P. (2020). Seasonal insect migrations: massive, influential, and overlooked. Front. Ecol. Environ. *18*, 335–344.
12. Chapman, J.W., Reynolds, D.R., and Wilson, K. (2015). Long-range seasonal migration in insects: mechanisms, evolutionary drivers and ecological consequences. Ecol. Lett. *18*, 287–302.
13. Meynard, C.N., Gay, P.-E., Lecoq, M., Foucart, A., Piou, C., and Chapuis, M.-P. (2017). Climate-driven geographic distribution of the desert locust during recession periods: Subspecies' niche differentiation and relative risks under scenarios of climate change. Glob. Chang. Biol. *23*, 4739–4749.
14. López-Mañas, R., Pascual-Díaz, J.P., García-Berro, A., Bahleman, F., Reich, M.S., Pokorny, L., Bataille, C.P., Vila, R., Domingo-Marimon, C., and Talavera, G. (2022). Erratic spatiotemporal vegetation growth anomalies drive population outbreaks in a trans-Saharan insect migrant. Proc. Natl. Acad. Sci. USA *119*, e2121249119.
15. Talavera, G., and Vila, R. (2016). Discovery of mass migration and breeding of the painted lady butterfly *Vanessa cardui* in the Sub-Sahara: the Europe-Africa migration revisited. Biol. J. Linn. Soc. *120*, 274–285.
16. Talavera, G., Bataille, C., Benyamini, D., Gascoigne-Pees, M., and Vila, R. (2018). Round-trip across the Sahara: Afrotropical Painted Lady butterflies recolonize the Mediterranean in early spring. Biol. Lett. *14*, 20180274.
17. Menchetti, M., Guéguen, M., and Talavera, G. (2019). Spatio-temporal ecological niche modelling of multigenerational insect migrations. Proc. Biol. Sci. *286*, 20191583.
18. Talavera, G., García-Berro, A., Talla, V.N.K., Ng'iru, I., Bahleman, F., Kébé, K., Nzala, K.M., Plasencia, D., Marafi, M.A.J., Kassie, A., et al. (2023). The Afrotropical breeding grounds of the Palearctic-African migratory painted lady butterflies (*Vanessa cardui*). Proc. Natl. Acad. Sci. USA *120*, e2218280120.
19. Suchan, T., Talavera, G., Sáez, L., Ronikier, M., and Vila, R. (2019). Pollen metabarcoding as a tool for tracking long-distance insect migrations. Mol. Ecol. Resour. *19*, 149–162.
20. Bradshaw, W.E., and Holzapfel, C.M. (2007). Evolution of animal photoperiodism. Annu. Rev. Ecol. Evol. Syst. *38*, 1–25.
21. Chapuis, M.P., Loiseau, A., Michalakis, Y., Lecoq, M., Franc, A., and Estoup, A. (2009). Outbreaks, gene flow and effective population size in the migratory locust, Locusta migratoria: A regional-scale comparative survey. Mol. Ecol. *18*, 792–800.
22. Boman, J., Zhu, Y., Höök, L., Vila, R., Talavera, G., and Backström, N. (2023). Environmental stress during larval development induces head methylome profile shifts in the migratory painted lady butterfly (*Vanessa cardui*). Mol. Ecol. *32*, 3513–3523.
23. Näsvall, K., Shipilina, D., Vila, R., Talavera, G., and Backström, N. (2023). Resource availability affects activity profiles of regulatory elements in a long-distance butterfly migrant. Authorea. https://doi.org/10.22541/au.167827909.99815237/v1.
24. Nielsen, M.E., Nylin, S., Wiklund, C., and Gotthard, K. (2023). Evolution of butterfly seasonal plasticity driven by climate change varies across life stages. Ecol. Lett. *26*, 1548–1558.
25. Stefanescu, C., Soto, D.X., Talavera, G., Vila, R., and Hobson, K.A. (2016). Long-distance autumn migration across the Sahara by painted lady butterflies: exploiting resource pulses in the tropical savannah. Biol. Lett. *12*, 20160561.
26. Reich, M.S., Shipilina, D., Talla, V., Bahleman, F., Kébé, K., Berger, J.L., Backström, N., Talavera, G., and Bataille, C.P. (2023). Intercontinental panmixia despite distinct migration distances in the trans-Saharan butterfly migrant *Vanessa cardui*. Preprint at bioRxiv. https://doi.org/10.1101/2023.12.10.569105.
27. Dobronosov, V. (2019). About mass migration of Painted Lady Butterfly (*Vanessa cardui* L.) to Eurasia in 2019. Concept of Dairy and Veterinary Sciences *2*, 244–248.
28. García-Berro, A., Talla, V., Vila, R., Wai, H.K., Shipilina, D., Chan, K.G., Pierce, N.E., Backström, N., and Talavera, G. (2023). Migratory behaviour is positively associated with genetic diversity in butterflies. Mol. Ecol. *32*, 560–574.
29. Wotton, K.R., Gao, B., Menz, M.H.M., Morris, R.K.A., Ball, S.G., Lim, K.S., Reynolds, D.R., Hu, G., and Chapman, J.W. (2019). Mass seasonal migrations of hoverflies provide extensive pollination and crop protection services. Curr. Biol. *29*, 2167–2173.e5.
30. Hawkes, W.L.S., Walliker, E., Gao, B., Forster, O., Lacey, K., Doyle, T., Massy, R., Roberts, N.W., Reynolds, D.R., Özden, Ö., et al. (2022). Huge spring migrations of insects from the Middle East to Europe: quantifying the migratory assemblage and ecosystem services. Ecography *2022*, e06288.
31. Reich, M.S., Kindra, M., Dargent, F., Hu, L., Flockhart, D.T.T., Norris, D.R., Kharouba, H., Talavera, G., and Bataille, C.P. (2023). Metals and metal isotopes incorporation in insect wings: Implications for geolocation and pollution exposure. Front. Ecol. Evol. *11*, 1085903.

32. Gorki, J.L., Suchan, T., and Talavera, G. (2024). Protocol to sequence markers from pollen grains carried by long-range migratory butterflies using metabarcoding. STAR Protoc. *5*, 103012. https://doi.org/10.1016/j.xpro.2024.103012.

33. Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620.

34. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. *17*, 10–12.

35. Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ *4*, e2584.

36. Edgar, R.C. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. Preprint at bioRxiv. https://doi.org/10.1101/074161.

37. Chamberlain, S. (2017). rgbif: Interface to the Global 'Biodiversity' Information Facility 'API'. R package version 0.9.8. https://CRAN.R-project.org/package=rgbif.

38. Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B., and Anderson, R.P. (2015). spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography *38*, 541–545.

39. Thuiller, W., Lafourcade, B., Engler, R., and Araújo, M.B. (2009). BIOMOD - a platform for ensemble forecasting of species distributions. Ecography *32*, 369–373.

40. Chávez, R.O., Estay, S.R., Lastra, J.A., Riquelme, C.G., Olea, M., Aguayo, J., and Decuyper, M. (2023). npphen: an R-package for detecting and mapping extreme vegetation anomalies based on remotely sensed phenological variability. Remote Sens *15*, 73.

41. Chui, S.X. (2019). BCdatabaser - its2.viridiplantae.none.2019-12-13.zip (1.1.1) [Dataset] (Zenodo). https://doi.org/10.5281/zenodo.3575246.

42. Harris, I., Osborn, T.J., Jones, P., and Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Sci. Data *7*, 109.

43. African Plant Database (version 4.0.0). Conservatoire et Jardin botaniques de la Ville de Genève and South African National Biodiversity Institute, http://africanplantdatabase.ch

44. Euro+Med (2006). Euro+Med PlantBase. The information resource for Euro-Mediterranean plant diversity. https://ww2.bgbm.org/EuroPlusMed/.

45. JACQ consortium (2004). Virtual Herbaria. https://www.jacq.org/consultedon2022-07-01.

46. POWO (2023). Plants of the World Online (Royal Botanic Gardens). http://www.plantsoftheworldonline.org/.

47. Pfeiler, E., and Markow, T.A. (2017). Population connectivity and genetic diversity in long-distance migrating insects: divergent patterns in representative butterflies and dragonflies. Biol. J. Linn. Soc. *122*, 479–486.

48. Liao, C., Downie, S.R., Li, Q., Yu, Y., He, X., and Zhou, B. (2013). New insights into the phylogeny of angelica and its allies (Apiaceae) with emphasis on East Asian species, inferred from nrDNA, cpDNA, and morphological evidence. Syst. Bot. *38*, 266–281.

49. Weng, L., Jiang, Y., Wang, Y., Zhang, X., Zhou, P., Wu, M., Li, H., Sun, H., and Chen, S. (2023). Chloroplast genome characteristics and phylogeny of the Sinodielsia clade (apiaceae: apioideae). BMC Plant Biol. *23*, 284.

50. Stull, G.W., Pham, K.K., Soltis, P.S., and Soltis, D.E. (2023). Deep reticulation: the long legacy of hybridization in vascular plant evolution. Plant J. *114*, 743–766.

51. Tessarolo, G., Lobo, J.M., Rangel, T.F., and Hortal, J. (2021). High uncertainty in the effects of data characteristics on the performance of species distribution models. Ecol. Indic. *121*, 107147.

52. Pearson, R.G., Raxworthy, C.J., Nakamura, M., and Townsend Peterson, A. (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. J. Biogeogr. *34*, 102–117.

53. van Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J., and Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. Ecography *39*, 542–552.

54. Velazco, S.J.E., Ribeiro, B.R., Laureto, L.M.O., and De Marco Júnior, P. (2020). Overprediction of species distribution models in conservation planning: a still neglected issue with strong effects. Biol. Conserv. *252*, 108822.

55. Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., and Araújo, M.B. (2011). Ecological Niches and Geographic Distributions, Monographs in Population Biology (Princeton University Press).

56. Briscoe, N.J., Elith, J., Salguero-Gómez, R., Lahoz-Monfort, J.J., Camac, J.S., Giljohann, K.M., Holden, M.H., Hradsky, B.A., Kearney, M.R., McMahon, S.M., et al. (2019). Forecasting species range dynamics with process-explicit models: matching methods to applications. Ecol. Lett. *22*, 1940–1956.

57. Jia, H., Liu, Y., Li, X., Li, H., Pan, Y., Hu, C., Zhou, X., Wyckhuys, K.A.G., and Wu, K. (2022). Windborne migration amplifies insect-mediated pollination services. eLife *11*, e76230.

58. Didan, K. (2015). MOD13C2 MODIS/Terra Vegetation Indices Monthly L3 Global 0.05Deg CMG V006. https://doi.org/10.5067/MODIS/MOD13C2.006.

59. Motohka, T., Nasahara, K.N., Murakami, K., and Nagai, S. (2011). Evaluation of sub-pixel cloud noises on MODIS daily spectral indices based on in situ measurements. Remote Sens *3*, 1644–1662.

60. Bounouh, O., Tarquis, A.M., Essid, H., and Farah, I.R. (2020). Normality of NDVI time series under scope: case study of various plant types of Tunisia, a Mediterranean country. In 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS) (IEEE), pp. 355–358.

61. Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., et al. (2017). An ecoregion-based approach to protecting half the terrestrial realm. Bioscience *67*, 534–545.

Please cite this article in press as: Gorki et al., Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.05.037

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Pollen grains isolated from samples of *Vanessa cardui* butterflies	This study	N/A
Chemicals, peptides, and recombinant proteins		
Beckman Coulter Agencourt AMPure XP Beads	Beckman Coulter GmbH	Cat#A63881
dNTP Mix, 10 mM each	Thermo Fisher Scientific	Cat#R0192
Phusion High-Fidelity DNA-Polymerase (2 U/μL)	Thermo Fisher Scientific	Cat#F530S
Ethanol ≥99.5%, Ph. Eur., extra pure	Carl Roth	Cat#5054
Deionized ultrapure water, UV treated	Millipore	N/A
Optionally: 1M Tris buffer.	Sigma	Cat#648314
Critical commercial assays		
Phire Plant Direct PCR MasterMix	Thermo Fisher Scientific	Cat#F106S or F106L
Qubit dsDNA HS Assay Kit	Thermo Fisher Scientific	Cat#Q32854
Deposited data		
ITS2 Metabarcoding data	This study	ENA: PRJEB61277
Oligonucleotides		
ITS-S2F: ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCTNNN NNN ATG CGA TAC TTG GTG TGA AT	Suchan et al.,[11] Gorki et al.[32]	N/A
ITS-4R7: GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATCTNN NNN NTC CTC CGC TTA TTG ATA TGC	Suchan et al.,[11] Gorki et al.[32]	N/A
i7 indexing primer: CAAGCAGAAGACGGCAT ACGAGAT [i7] GTGACTGGAGTTCAGACGTGTGC	Suchan et al.,[11] Gorki et al.[32]	N/A
i5 indexing primer: AATGATACGGCGAC CACCGAGATCTACAC [i5] ACACTCTTTCCCTACACGACGC	Suchan et al.,[11] Gorki et al.[32]	N/A
Software and algorithms		
PEAR	Zhang et al.[33]	https://www.h-its.org/software/pear-paired-end-read-merger/
CUTADAPT	Martin[34]	https://github.com/marcelm/cutadapt/
VSEARCH	Rognes et al.[35]	https://github.com/torognes/vsearch
SINTAX	Edgar[36]	https://www.drive5.com/usearch/manual/sintax_downloads.html
rgbif	Chamberlain[37]	https://CRAN.R-project.org/package=rgbif
spThin	Aiello-Lammens[38]	https://cran.r-project.org/web/packages/spThin/
Biomod2	Thuiller et al.[39]	https://cran.r-project.org/web/packages/biomod2/
npphen	Chávez et al.[40]	https://cran.r-project.org/web/packages/npphen/
Other		
Screw cap free-standing micro tube, sterile, 2 mL	Sarstedt	Cat#72.694.005
Screw caps suitable for screw cap micro tubes	Sarstedt	Cat#65.716.725
Borosilicate Glass beads 3 mm	Merck KGaA	Cat#Z143928
Brand PCR strip tubes and caps	Merck KGaA	Cat#BR781327
TissueLyzer LT	Qiagen	Cat#85600
Savant SpeedVac	Thermo Scientific	Cat#SPD111V-230
CL-1000 UV crosslinker or similar (optional)	Analytik Jena	Cat#95-0174-02

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Qubit or Qubit Flex Fluorometer (optional)	Thermo Fisher	Cat#Q33238 or Q33327
2100 Bioanalyzer or 5300 Fragment Analyzer system (optional)	Agilent	Cat#G2991BA or M5311AA
BCdatabaser	Chui[41]	https://doi.org/10.5281/zenodo.3575246
MOD13C2 V006	NASA	https://doi.org/10.5067/MODIS/MOD13C2.006
African Plant Database (version 4.0.0)	African Plant Database	http://africanplantdatabase.ch
Euro+Med PlantBase	Euro+Med PlantBase	https://ww2.bgbm.org/EuroPlusMed
GBIF, Global Biodiversity Information Facility: plant occurrences	GBIF	https://www.gbif.org/
Virtual Herbaria: plant occurrences	JACQ consortium	https://www.jacq.org/
POWO: plant occurrences	Plants of the World Online	http://www.plantsoftheworldonline.org/
CRU TS v4 (Climatic Research Unit): monthly high-resolution gridded multivariate climate dataset	Harris et al.[42]	https://crudata.uea.ac.uk/cru/data/hrg/cru_ts_4.04/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Gerard Talavera (gerard.talavera@csic.es)

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The obtained sequence data supporting the findings of this study have been deposited in the European Nucleotide Archive project (project PRJEB61277): https://www.ebi.ac.uk/ena/data/view/PRJEB61277.
- All original code used to process the data is publicly available at https://github.com/GTlabIBB/Pollen2019 (https://zenodo.org/doi/10.5281/zenodo.10802533).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Painted lady (*Vanessa cardui*) adult butterflies were collected across 10 countries over 7 months in 2019. Collected specimens were stored in sealed glassine envelopes after collection, and frozen at −20°C before pollen analyses. For analytical purposes, samples were clustered in 16 spatio-temporal groups according to site and date of collection. See method details section for more information.

METHOD DETAILS

Swarming observation, sample collection, and spatio-temporal groups

Field observations of large groups and migratory swarms of *Vanessa cardui* butterflies were obtained from the authors' own surveys, citizen science platforms, mainly the Butterfly Migration Monitoring Scheme (http://www.butterflymigration.org/), iNaturalist (https://www.inaturalist.org) and https://observation.org/, records shared on social networks (e.g., Twitter), press notes, and personal communications (Data S1). When available, data on directionality and estimated numbers were gathered. Butterflies for pollen metabarcoding were collected across 10 countries over 7 months (from March to November 2019). Specimens were stored in sealed glassine envelopes after collection, thus minimizing potential environmental contamination of pollen, and frozen at −20°C before further pollen analyses. For downstream spatial analyses, the collected butterflies were classified in 16 spatio-temporal groups, clustering identical or close collecting sites and dates, and including a minimum of four samples per group (Data S1).

Pollen isolation and metabarcoding library preparation

Pollen isolation and metabarcoding library construction were conducted for a total of 264 butterflies, with a maximum of 23 samples per batch.[32] Library construction comprised two essential steps: first, the amplification of the Internal Transcribed Region 2 (ITS2),

followed by a second reaction, where the ITS2 fragments were indexed for sequencing. To ensure contamination-free procedures, pollen isolation and set up of the initial polymerase chain reaction (PCR) were performed under a laminar flow cabinet. The working space and equipment were thoroughly cleaned with 10% bleach solution and 70% ethanol, and further decontaminated under UV light for 15 min. To prevent any cross-contamination during butterfly processing, tweezers were additionally sterilized by flaming between each sample. To monitor possible contamination, we included one blank sample for pollen isolation and one blank for the first PCR reaction for each batch of butterflies.

Pollen isolation

To isolate pollen from butterfly bodies,[32] we vortexed the bodies in a 2 mL tube filled with 500 μL of 100% sterile EtOH. After vortexing, the tube was centrifuged for 15 min at 10,000 rpm. The bodies were then removed from the tube, and the remaining ethanol was evaporated at 50°C for a minimum of 30 min using a speed vacuum centrifuge. For the blank sample, 500 μL of 100% EtOH was used, and the tube was left open for 20s in the laminar flow cabinet to mimic the opening and closing of sample tubes during pollen isolation. The obtained pollen pellets were stored at −20°C until further usage.

ITS2 amplification

The amplification of the ITS2 region of the ribosomal DNA was carried out using the ITS-S2F and ITS4R primer combination (see key resources table), which were tailed with six random nucleotides to increase sequence diversity and also included a part of the Illumina adapter.[32] Based on successful results from a previous study[19] that used Phire Plant Direct Polymerase without prior DNA extraction, we directly performed the PCR with the Phire Plant Direct MasterMix Kit (ThermoFisher Scientific #F160L, Waltham, MT, USA). To prepare pollen pellets for PCR, we diluted them in 25 μL of 0.5x Phire Plant Direct Dilution Buffer and homogenized them using three 3 mm glass beads per sample on a TissueLyser LT machine (Qiagen, Hildesheim, Germany) at 30 Hz for 3 min. After homogenization, the samples were centrifuged at 12,000 rpm for 20 s to separate the supernatant for PCR. Three independent PCR reactions were performed for each sample to avoid reaction-specific bias. Each reaction consisted of 2 μL of the pollen sample disrupted in buffer, 25 μL of 2x Phire Plant Direct PCR MasterMix and 2.5 μL of each primer (10 μM) in a total volume of 50 μL. For blank samples, 2 μL of nuclease-free water were added instead of the pollen sample. The PCR program included an initial denaturation step at 98°C for 5 min, followed by 20 cycles of denaturation at 98°C for 40 s, annealing at 49°C for 40 s, elongation at 72°C for 40 s, and a final extension at 72°C for 5 min. PCR triplicates were combined before purifying the product using AMPure XP beads (Beckman Coulter, Indianapolis, IN, USA) at 1x ratio. The beads were washed 3 times with 70% EtOH before eluting the purified PCR product in 10 μL of nuclease-free water. The purified PCR product was stored at −20°C until further usage.

Library preparation

To prepare the final library,[32] we used index sequences from the TruSeq Amplicon series and supplemented them with TruSeq CD index sequences, resulting in a total of 16 different forward primers and 24 different reverse primers. Similar to previous steps, the PCR for the final library was conducted in triplicates for each sample. Each PCR reaction included 0.2 μL dNTPs (10 mM), 0.5 μL of each forward and reverse primer (10 μM), 0.2 μL Phusion High-Fidelity Polymerase (ThermoFisher Scientific #F530S, Waltham, MT, USA), 2 μL 5x HF Buffer, and 1 μL of purified product of the first PCR, resulting in a 10 μL reaction volume. The PCR program consisted of an initial denaturation step at 98°C for 30 s, 12 cycles of denaturation at 98°C for 10 s and simultaneous annealing and elongation at 72°C for 30 s (shuttle PCR), and a final extension at 72°C for 5 min. The PCR triplicates were combined after the reaction. Blank samples were processed alike, but using 1 μL of water instead of sample. Finally, 3μL of the PCR product from each sample belonging to the same lane were pooled and purified as described above. Pooled libraries were sequenced with 15% PhiX spike-in on an Illumina MiSeq platform, with 2x250 bp paired-end reads across two lanes.

Bioinformatic pipeline for molecular data analyses

Paired-end reads were merged using PEAR v0.9.11[33] and successfully merged reads were trimmed of primer sequences with cutadapt v3.2.[34] The subsequent processing steps were performed using VSEARCH v2.20.0[35]: first, reads were filtered by an expected error rate of 1 and a minimal length of 250nt and reads with ambiguous nucleotides were removed. Then, for each sample the remaining reads were dereplicated *(-derep_fulllength* option), the obtained sequences of all samples combined and dereplicated a second time. Next, sequences were clustered with 98% similarity and the centroids sorted by cluster size. Centroids with a cluster size of 1 and detected chimeras (*-uchime_denovo*) were removed. The remaining centroids were classified with SINTAX *(-sintax)*[36] using the ITS2 database by Chui et al.,[41] and the abundance of sequences in each sample was determined (*-usearch_global* and *-id 0.98*) (Data S1). Only centroid sequences classified to species level with a probability larger or equal to 95%, 98% and 100% respectively were considered and the reads were then summarized by plant species resulting in 3 different datasets. The code used for read processing is available at https://github.com/GTlabIBB/Pollen2019.

To exclude potential biases caused by contamination, each batch of samples were manually analyzed. For each plant species, reads of the batch samples (S_S) and the two batch blanks (S_B) were added respectively. If S_B was larger than S_S the sample reads were discarded. In the opposite case, sample reads were discarded when $S_B/S_S \geq 0.5$. However, if S_B was higher than 100 although $S_B/S_S < 0.5$ the reads were removed as we considered the high number of reads in the blanks compared to the ratio too high.

Please cite this article in press as: Gorki et al., Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.05.037

To rule out possible false positives, we further applied two thresholds for each classification dataset: for each sample, classified species with less than 3 and 10 reads respectively were removed. In total, we ended up with 6 different datasets (95% match - $\geq$ 3 reads, 95% match - $\geq$ 10 reads, 98% match - $\geq$ 3 reads, 98% match - $\geq$ 10 reads, 100% match - $\geq$ 3 reads, 100 match - $\geq$ 10 reads), where 100% identification probability and minimum of 10 reads per species in a sample represents the strictest conditions.

Plant taxonomy and dataset curation

A total of 398 plant occurrence datasets were retrieved from the Global Biodiversity Information Facility (GBIF) using the R package rGBIF,[37] corresponding to each identified plant with the less stringent set of criteria (95% match - $\geq$ 3 reads). The filtering options ''capitals'', ''centroids'', ''equal'', ''gbif'', ''institutions'', ''seas'', ''zeros'' and ''countries'' were used in rGBIF to flag potentially erroneous occurrences, which were further removed from each individual species set or coordinates. The R package spThin[38] was used to spatially thin each species occurrence data by a minimum separation distance of 10 km between occurrences, to minimize potential spatial biases in subsequent species distribution modelling.

All resulting occurrence datasets were manually curated by visually inspecting plotted distribution maps for obvious outliers, which were removed from datasets when detected in very few cases. Potential taxonomic issues affecting the retrieved set of occurrences were also considered when validating final datasets for species distribution modelling. For 43 plant species (ca. 11% of the total), additional records were compiled from published studies and/or the following online resources: the African Plant Database,[43] the Euro+Med PlantBase[44] and digital images from virtual herbaria.[45,46] This approach was particularly valuable for AME endemic plants, which had limited available occurrences in GBIF. The resulting curated occurrence datasets, along with the code used for processing, are available on the GitHub repository https://github.com/GTlabIBB/Pollen2019.

Spatial analyses: Plant species distribution models and geolocation

The study area was set between -30° to 90° longitude and -5° to 72° latitude, thus including Europe and most of Africa and Asia. This large region encompasses the known extent of the Afro-Palearctic migratory range of *Vanessa cardui* in latitude, and an extended region in longitude covering the Indian subcontinent. Extending the study area towards East Asia seems unnecessary as winter breeding groups are unlikely to be related to the spring-to-autumn migratory circuit between Africa, Middle East and Europe, although subpopulations are likely to exchange genes in the northern part of the range.[28,47]

The predictor variables used for plants species distribution modelling (SDM) were retrieved from the Climate Research Unit (CRU) database (version 4.05)[42] at a pixel size of 0.5°, and these were cropped for the extent of the study area. Since CRU variables are provided as monthly time-series spanning from 1901 to 2021, we calculated the average for each variable to create a single layer, focusing on the period from 1970 to 2020. This choice was made to avoid incorporating noise from major climatic changes expected between the earlier and later years of the whole time-series, and because most of the gathered occurrences fell within the selected period. A subset of variables was selected after excluding highly correlated variables with Pearson's coefficients higher than 0.8. The final subset of variables included: vapour pressure (vap), monthly average daily mean temperature (tmp), precipitation (pre), cloud cover (cld), frost day frequency (frs), diurnal temperature range (dtr) and potential evapotranspiration (pet).

We used Biomod2[39] to model the distribution of each individual plant, using an ensemble approach. The modelling strategies used were: Generalized Linear Models (GLM), Generalized Additive Models (GAM), Multiple Adaptive Regression Splines (MARS), Classification Tree Analysis (CTA), Flexible Discriminant Analysis (FDA), and Random Forests (RF). Occurrences were split in a proportion of 80%–20% for calibrating and evaluating the accuracy of the models, respectively. Five runs of each model were performed, using a random split of the data in each. A total of 1000 pseudoabsences were randomly set, including five replicates of each pseudoabsence dataset. Ensemble models were obtained through committee averaging, and parameters for model ensemble were set to only account for runs with high evaluation scores of the True Skill Statistic (TSS) higher than 0.7. The ensembled models were projected and plotted into probability rasters.

To infer probabilities of provenance for each butterfly spatio-temporal group, we calculated the joint spatial probabilities for all plants found within each group. First, the obtained SDM probability rasters for each plant were rescaled from 0 to 1, thus making the contribution of every plant to the joint spatio-temporal group equally important. To extract a joint probability raster accounting for all plants in the group, all rasters were aggregated through averaging them, and posteriorly rescaled again from 0 to 1 to make probabilities comparable between spatio-temporal groups. We discriminated between present and absent plants at each of the butterfly collecting regions/countries, based on the compiled occurrence data for each species (Data S1 andS2), and joint probability rasters were computed independently for the identified regionally absent plants, regionally present plants, and both together (all plants found in the group) (Data S3). This process was fundamental for recovering provenance signal, which only emerged when the analysis included exclusively regionally absent plants (Data S3).

Pollen metabarcoding and plant distributions: Signal-to-noise ratio counteraction

Our approach is not exempt of synergistic sources of potential error that may increase uncertainty in our predictions. Uncertainty can be primarily introduced at two scales: molecular identifications and spatial distributions. Possible sources of error at the molecular level include: wrong molecular identification (e.g., due to limited sequences at reference databases or hybridization), pollen or DNA contamination (e.g., environmental or across samples), or amplification errors. For spatial distributions, sources of noise include poor, unbalanced or misidentified occurrence observations leading to poorly performed SDM predictions, or the presence of ornamental plants outside their original distributions.

Please cite this article in press as: Gorki et al., Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.05.037

Although a strict set of measures were applied through the molecular protocol and in filtering molecular data to minimize noise, our individual-level results still showed obvious cases of error, likely due to limitations in the classification of molecular reads. For an objective evaluation of our spatial modelling approach, we used a blind set of criteria within our study range, where we did not pre-select between likely or unlikely informative plants. However, we discarded plants for which records were exclusively outside the study range, mainly including American and Eastern Asian species. We believe that this is a reasonable assumption considering the dimensions of the study range and the *a priori* butterfly occurrence observation-based evidence. Including molecularly identified OTUs classified to taxa outside the study range would exclusively incorporate noise to the joint probability rasters. A total of 29 plant taxa were excluded for this reason, representing 7.3% of the plants identified among the total 265 analyzed butterflies, which indicates that misidentification is not negligible. Therefore, we can expect that a similar percentage of problematic matches might apply to the dataset used for spatial distribution analyses within the studied range. A few OTUs matched to Chlorophyta and Bryophyta taxa (17 OTUs, 4.3%), which were also excluded because of the difficulty to track their distribution and biological meaning.

In this context of signal-to-noise ratio (SNR), we obtained a variety of classified plant species with intriguing geographic distributions. For example, we identified *Vicatia thibetica* in Finland, Denmark and the Pyrenees (June) under the most stringent filtering criteria (100% match, >10 reads). *Vicatia thibetica* is an Asian species from the Himalayan-Tibetan region and appears to be unrealistically identified. It belongs to a small genus comprising 5 species[46] endemic to Central Asia. However, its closely related genus, *Angelica*, is a rich clade of ~60 species widely distributed in the boreal and alpine zone of the Holarctic region, and for which taxonomy and phylogenetic relationships are not well resolved.[48,49] Misidentifications within species' rich genera and/or recent radiations are even more likely to occur, due to genetic incomplete lineage sorting or frequent hybridization.[50] We identified several potential cases of misidentification within various genera, such as *Puccinellia*, *Convolvulus* or *Senecio*. One major problem contributing to potential misidentifications may occur when reference sequence libraries are unbalanced between regions, because classifications of taxa from underrepresented regions may tend to be classified as their closest relatives from better represented regions. This may occur to some extent in our dataset, given that European plants are overrepresented in reference libraries compared to AME and African plants.

Among the applied sequence classification thresholds, we generally observe that thresholds < 98% sequence probability match and < 10 reads often lead to inaccurate identifications, ultimately introducing excessive noise in the datasets. Therefore, we recommend interpreting results from classified taxa with a probability match of $\geq$ 98% and a minimum of $\geq$ 10 reads. While applying the strictest thresholds may improve accuracy, it will also reduce the number of plants in the datasets, leading to geolocation inferences relying on a limited sample. This is the case of our dataset for Israel (March), where the strictest conditions remove likely misidentified Western European taxa but only retain a single AME plant. Another notable case is the spatio-temporal group of Chios in May/June, which show both plants with Northern Europe/Caucasus distribution and endemics from southern Caucasus, even with the strictest thresholds (Data S1 and S3). However, this group consists of only 4 specimens, with three collected on the 17th of June and one on the 25th of May. Indeed, pollen metabarcoding for these two dates shows different unique matches (Data S1), indicating independent origins. This temporal separation (the only one of its kind among all established spatio-temporal groups) may account for the dual provenance signal, and both provenance options are considered plausible.

At the spatial scale, SNR also applies. For example, models from widespread species contribute a higher proportional area of uncertain geolocation than restricted endemics with small distributions, as it is not realistic to assume that the insects have migrated across their whole predicted range. Therefore, geographically restricted plants should be more informative for geolocation and SDM performance have proved better for species of small geographic range size,[51] although on the other hand are also more prone to poor model predictions if occurrences are very low.[51–53] Thus, a similar representation of widespread and restricted species may play against the signal coming, for example, from endemism-rich regions. Another possible SNR effect could be due to imbalances between unique plant representatives from different floristic regions and/or geographic imbalances in occurrence records for widespread species. For example, for Russia, Finland and Denmark we find: 13 taxa of Central-Asian distribution, 15 taxa that are strictly distributed West of approximately 30° of longitude, and 15 taxa are distributed in both sides of this longitude. However, occurrences for the widespread taxa are largely overrepresented for the West and these West-East unbalances are translated in the predicted models (Data S2).

Accounting for SDM overprediction is also important in distributional modelling and spatial conservation prioritization.[54] As SDM may overpredict broader geographic distribution ranges where species typically occur[54–56] and this could blur our interpretation for insect movement, we masked the resulting probability rasters for each detected plant by applying relaxed convex hulls based on the existing gathered occurrences. These convex hulls allowed for a buffered zone of at least 12 pixels (corresponding to ca. 222 km^2) from the nearest record and a maximum of 83 pixels (ca. 1535 km^2) if contiguous high suitability spanned from the nearest record (Data S2). In a minority of cases showing obvious large-scale disjunct distributions, two independent convex hulls were applied. This exercise was expected to optimize the signal from species' occupied distribution area over the invadable distribution area.

Thus, accounting for a signal-to-noise ratio is essential to interpret results when using our approach. It is expected that the predicted geolocation signal progressively improves while increasing the number of insect samples and the number of plants identified within each spatio-temporal group, being both likely correlated but not necessarily, as we obtain imbalanced numbers of pollen reads within spatio-temporal butterfly groups (Data S1). Larger sample sizes (for butterflies and plants) should counteract noise introduced by molecular misidentifications and distribution models of lower quality, being a consequence of poorly represented genetic libraries

Please cite this article in press as: Gorki et al., Pollen metabarcoding reveals the origin and multigenerational migratory pathway of an intercontinental-scale butterfly outbreak, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.05.037

and insufficient/biased plant distributional data. In addition, a larger number of spatio-temporal groups will also improve the interpretation of the resulting models at a continental scale, as extracting conclusions from a single group could be artefactual if noise is not sufficiently counteracted within the group.

While the pollen-based geolocation method retains some caveats, our approach successfully detects the general movement patterns of the multiple generations of the studied *V. cardui* outbreak. Noise is expected to be greater when investigating migration across small ranges or across contiguous land (and thus through plant distribution gradients), such as within Europe. Signal from pollen metabarcoding is less ambiguous when studying migrations that connect different biomes.[19,57] This also applies to our study, provenancing abilities are stronger when comparing different floristic regions, such as AME in respect to Europe, but are lower in contiguous floristic regions within Europe.

Vegetation growth anomalies

Given the trophic relationship between *V. cardui* larvae and vegetation growth activity, we investigated the possible geographic epicenter of the outbreak by calculating vegetation growth anomalies through the remote sensing derived Normalized Difference Vegetation Index (NDVI). NDVI takes advantage of the low reflectance in the red wavelength (~650nm) and the high reflectance in the near infrared wavelength (~820nm) of plant photosynthesis process to proxy the vegetation activity. We used MODIS satellite MOD13C2 product from NASA, which provides global coverage of NDVI at 0.05° of pixel size and monthly temporal resolution.[58] We gathered a time series of NDVI images for the winter distributional area of *V. cardui* (from 20° to 45° latitude and from −20° to 60° longitude), covering the period of 2001 to 2021. Considering massive occurrences observed in the Middle East in early spring, together with the provenance information obtained from pollen metabarcoding, we paid particular attention to anomalies in this region that could mechanistically explain the source of the observed outbreak.

To ensure data accuracy, the raster time series were filtered using the pixel reliability index provided within the quality layers of the original source product. This step is critical for reducing signal noise caused by atmospheric effects such as clouds or dust plumes, which could otherwise interfere with the identification of anomalous vegetation growth events.[59]

Subsequently, the Phenological Baseline was computed at pixel scale to establish a reference database. To avoid assuming normality in the temporal distribution of NDVI,[60] given the irregular nature of vegetation's annual phenological cycle, we used a non-parametric method - the Kernel Density Estimation (KDE). This method, implemented through the PhenAnoMap function in the R package *npphen*,[40] estimates the most expected value of a vegetation index throughout the growing season of the whole time series based on a probability density function. It identifies the most frequent values of VI along the phenological cycle (the reference annual phenology for the area) and its frequency distribution (phenological variability along the years). Once the Phenological Baseline was established, anomalies were calculated as the difference between the MODIS observed values and the most frequent or expected ones extracted from the Phenological Baseline.

Together with the anomalies, the KDE function provides an indicator the anomaly's extremity (Ae) during a specific season, ranging from 0 to 100. We selected NDVI observations with values at Ae $\geq$ 95 (representing the 95th to 100th percentile of the NDVI interval for each pixel), to identify the most extreme positive NDVI anomalies that could be considered vegetation outbreaks (Figure 2B). Anomalies were also calculated for 18 delimited ecoregions within AME, as delimited in,[61] using averaged NDVI values within each ecoregion (Data S3).

QUANTIFICATION AND STATISTICAL ANALYSIS

Details on the clustering process and filtering of molecular reads can be found in the "bioinformatic pipeline for molecular data analyses" section. Sequences were clustered with 98% similarity and the centroids were sorted by cluster size. Centroids classified to species level with probabilities larger or equal than 95%, 98% and 100% were retrieved for downstream analyses.

Details on species distribution models and spatial analyses are described in the "spatial analyses: plant species distribution models and geolocation" section. Plant SDMs were inferred using an ensemble approach of 6 modelling strategies in Biomod2. A True Skill Statistic (TSS) higher than 0.7 was used to evaluate and retain models.