Pollen metabarcoding as a tool for tracking long-distance insect migrations.

Tomasz Suchan[1†*], Gerard Talavera[2,3†], Llorenç Sáez[4], Michał Ronikier[1], Roger Vila[2]

[1] W. Szafer Institute of Botany, Polish Academy of Sciences, Lubicz 46, 31-512 Kraków, Poland

[2] Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37, ESP-08003 Barcelona, Catalonia, Spain

[3] Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA

[4] Systematics and Evolution of vascular plants, Associated unit to CSIC. Unitat de Botànica, Facultat de Biociències, Universitat Autònoma de Barcelona, E-08193, Bellaterra, Catalonia, Spain

[†] TS and GT should be considered joint first authors

[*] Corresponding author: t.suchan@botany.pl; W. Szafer Institute of Botany, Polish Academy of Sciences, Lubicz 46, 31-512 Kraków, Poland

Running title: Pollen metabarcoding for tracking insect migrations

Abstract

Insects account for the main fraction of Earth's biodiversity and are key players for ecosystems, notably as pollinators. While insect migration is suspected to represent a natural phenomenon of major importance, remarkably little is known about it, except for a few flagship species. The reason for this situation is mainly due to technical limitations in the study of insect movement. Here we propose using metabarcoding of pollen carried by insects as a method for tracking their migrations. We developed a flexible and simple protocol allowing high multiplexing and not requiring DNA extraction, one of the most time consuming part of metabarcoding protocols, and apply this method to the study of the long-distance migration of the butterfly *Vanessa cardui*, an emerging model for insect migration. We collected 47 butterfly samples along the Mediterranean coast of Spain in spring and performed metabarcoding of pollen collected from their bodies to test for potential arrivals from the African continent. In total, we detected 157 plant species from 23 orders, most of which (82.8%) were insect-pollinated. African or African-Arabian endemic taxa contributed 21.0% of our dataset, strongly supporting the hypothesis that migratory butterflies colonize southern Europe from Africa in spring. Moreover, our data suggest that a northwards trans-Saharan migration in spring is plausible for early arrivals (February) into Europe, as shown by the presence of Saharan floristic elements. Our results demonstrate the possibility of regular insect-mediated transcontinental pollination, with potential implications for ecosystem functioning, agriculture and plant phylogeography. Despite current limitations, mostly regarding the availability of plant reference sequences and distribution data, the method

proved to be useful and demonstrates great potential as plant genetic libraries and distribution datasets improve.

Keywords: insect migration, pollen metabarcoding, pollination, Sahara, Africa, Europe, *Vanessa cardui*

Introduction

Insects undergo aerial long-distance migrations (Holland, Wikielski & Wilcove, 2006; Chapman, Reynolds & Wilson, 2015) that outnumber migrations of larger organisms, such as birds, both in abundance and biomass (Hu et al., 2016). These long-range movements have important –albeit still largely unknown– implications for ecosystems and human economy (Bauer & Hoye, 2014; Chapman et al., 2015; Hu et al., 2016). Nevertheless, mostly due to the technical challenges associated with tracking small organisms (Chapman et al., 2015), our knowledge on insect migrations is extremely limited (Holland et al., 2006), especially in comparison to that on vertebrate migrations.

Tracking long-distance insect migrations involves assessing the actual path of individuals, either by mark-recapture studies, using variety of markers (reviewed in Hagler & Jackson, 2001), or by telemetry (Osborne, Loxdale & Woiwod, 2002; Kissling, Pattemore & Hagen, 2014). However, mark-recapture studies of migrating insects have low recapture success rate and are feasible only for the most emblematic species, such as monarch butterflies (Knight, Brower & Williams, 1999; Garland & Davis, 2002). On the other hand, radio telemetry is only suitable for tracking the largest insects at short distances, and is technically limited by

58 relatively high weight of the transmitters and short battery life (Kissling et al., 2014). Because
59 of that, so far only one telemetric study was performed to assess long-distance migrations in
60 insects – of the dragonfly *Anax junius*. It however involved considerable logistic challenges,
61 such as using small planes and ground teams for tracking insect movements (Wikelski et
62 al., 2006).

63 Due to the above limitations, insect migration studies are traditionally observation-based,
64 linking the place, time and, sometimes, direction of the observations with hypothetical routes
65 (e.g. Howard & Davis, 2009; Talavera & Vila, 2016). More recently, other observational
66 methods, such as vertical-looking radars (Chapman, Drake & Reynolds, 2011) or Doppler
67 weather radars (Chilson et al., 2012; Westbrook & Eyster, 2017) have been used in insect
68 migration studies, often supplemented by aerial nets or other type of traps for accurate species
69 determination (Chapman et al., 2002; Chapman et al., 2004). The new developments in this
70 technology allow insect body mass, shape, wing-beat frequency, flight altitude and heading
71 directions to be measured, often allowing species determination (Dean & Drake, 2005;
72 Chapman et al., 2011; Drake et al. 2017), but observations are usually constrained to
73 particular areas because radars have no or limited mobility.

74 Other methods of studying migrations use intrinsic markers, such as genetic markers or
75 isotope composition. By using genetic markers, populations connected by regular gene flow
76 can be identified (Lowe & Allendorf, 2010), which may suggest migratory routes and natural
77 barriers (e.g. Nagoshi, Meagher & Hay-Roe, 2012). The utility of genetic markers is
78 dependent on spatial genetic structuring. For migratory insects, structuring is expected
79 between independent migratory circuits, but might not be maintained in the case of regular

gene flow between migrating lineages (global panmixia) (Lyons et al., 2012; Lukhtanov, Pazhenkova & Novikova 2016). Because the stable isotope ratios, such as $^{2}H/^{1}H$ or $^{13}C/^{12}C$, of organic tissues are related to the site where insects developed, these can also be used, along with modelled geographic isotope patterns (isoscapes), to infer probabilistic natal origins of migrating individuals (e.g. Hobson, Wassenaar & Taylor, 1999; Brättstrom et al., 2010; Stefanescu, Soto, Talavera, Vila & Hobson, 2016). This technique does not rely on marking-recapturing specimens and it is thus also suitable for small species (Hobson, 2008). However geospatial assignments depend on limited isoscape resolution and are usually only helpful at inferring large-scale geographic patterns.

As insects visit and feed on flowers, the pollen is deposited on their bodies and can be transported across large distances (Ahmed et al., 2009). Therefore, pollen of plants endemic to certain areas could also be exploited for tracking long-distance insect migrations (Hagler & Jackson, 2001; Jones & Jones, 2001) and, indeed, it has been used as a marker in a handful of studies (Mikkola, 1971; Hendrix, Mueller, Phillips & Davis, 1987; Hendrix & Showers, 1992; Gregg, 1993; Lingren et al., 1993, 1994; Westbrook et al., 1997). However, conventional pollen identification by light or electron microscopy is time-consuming, and requires specialized taxonomic knowledge. It is therefore difficult to apply as a widely accessible tool for large-scale studies (Galimberti et al., 2014; Keller et al., 2015; Richardson et al., 2015b; Sickel et al., 2015). Moreover, taxonomical classification of pollen grains is often unachievable to the species or even genus level (Hawkins et al., 2015; Kraaijeveld et al., 2015; Richardson et al., 2015b).

The development of the next generation sequencing (NGS) technologies allowed straightforward sequencing of DNA barcodes from mixed environmental samples, termed "metabarcoding" (Taberlet et al., 2012; Deiner et al., 2017). In this study, we use a DNA metabarcoding approach to identify pollen grains carried by a long-distance migratory insect species – the painted lady butterfly *Vanessa cardui*. This is a virtually cosmopolitan species adapted to seasonally exploit a wide range of habitats and sometimes observed even at extreme latitudes or in the open ocean (Shields, 1992). The Palearctic-African migratory system involves populations that undergo yearly long-distance latitudinal migrations in a circuit between Tropical Africa (September to February) and the temperate zone (February to September) (Talavera & Vila, 2016). Such annual circuit might involve 10 generations or more, some of them performing long-distance movements of thousands of kilometres. The distances crossed by individual butterflies within a single generation are unclear. It has been recently shown, both by stable isotope and observational evidence, that the species massively immigrate and breed in autumn in sub-Saharan Africa, and that these populations are of European origin (Stefanescu, Soto, Talavera, Vila & Hobson, 2016; Talavera & Vila, 2016). These results depict a new spatiotemporal model for the migration of *V. cardui* in this part of the world, which involves migratory movements across differentiated floristic regions: central and northern Europe, the Mediterranean, the Sahara and Tropical Africa. Although the phenology of migratory movements within the western Palearctic is well known, both from observations and entomological radars (Stefanescu et al., 2013), the African locales and routes of the species from October to March are unclear (Talavera & Vila, 2016).

Using pollen metabarcoding from captured butterflies apparently migrating northwards into Europe, we test i) for the presence of DNA from pollen grains deposited on insects' bodies after a long-distance migration from Africa and, if detected, ii) whether the sequences obtained are of African endemic plant species that could be informative on the migration routes.

Despite great potential of pollen DNA metabarcoding for pollination biology or palynological studies, it has been used only in a handful of cases, mostly to investigate honey composition (Valentini, Miquel & Taberlet, 2010; Bruni et al., 2015; Hawkins et al., 2015; Prosser & Herbert, 2017), honeybee foraging (Galimberti et al., 2014; Richardson et al., 2015a,b; de Vere, 2017) and, recently, plant-pollinator interactions (Keller et al., 2015; Sickel et al., 2015; Pornon et al., 2016; Bell et al., 2017). Several DNA markers have been proposed for identifying mixed pollen loads from insects (Bell et al., 2016), the Internal Transcribed Spacer 2 (ITS2) nuclear ribosomal fragment being one of the most frequently employed (Keller et al., 2015; Richardson et al., 2015b; Sickel et al., 2015). ITS2 has several advantages for pollen metabarcoding. Although plastid markers have been successfully amplified from pollen grains (e.g. Kraaijeveld et al., 2015; Richardson et al., 2015a), the variation in the number of plastid genome copies is poorly understood (Bell et al., 2016). Because it is a nuclear marker, ITS2 should be more uniformly present in pollen. Moreover, the sequence length of this marker is short enough for amplicon sequencing using Illumina MiSeq technology. It is also sufficiently informative to discriminate most plant species (Chen et al., 2010) and the number of available reference sequences in databases is the highest among the available plant barcodes (Bell et al., 2016).

144 Several library preparation protocols have been proposed for pollen metabarcoding.
145 Richardson et al. (2015b) performed ITS2 amplification using standard primers followed by
146 purification and NGS library preparation using commercial kits. Keller et al. (2015) and
147 Sickel et al. (2015) proposed an approach based on Kozich, Westcott, Baxter, Highlander and
148 Schloss (2013) protocol, where an ITS2 amplicon Illumina library is prepared within a single
149 PCR step. These authors used ITS2 primers tailed with appropriate technical sequences and
150 custom sequencing primers, to avoid losing sequencing cycles for ITS2 primer sequences and
151 avoid problems related to low sequence diversity of these regions. The downside of such
152 protocol is the need of replacing all the oligonucleotides when targeting other markers,
153 template-specific biases when using indexed primers (O'Donnell, Kelly, Lowell & Port,
154 2016), and the necessity to use custom sequencing primers.

155 Here, we developed a laboratory protocol with maximum flexibility, cost-effectiveness and
156 reduced workload in mind – especially by lowering the number of library preparation steps,
157 which not only require more work, but can also lead to sample cross-contamination. Our
158 method consists of two PCR steps: first, the ITS2 fragment is amplified using standard
159 primers (White, Bruns, Lee & Taylor, 1990; Chen et al., 2010), tailed with partial Illumina
160 sequences; in the second step, the fragments are double-indexed and the final library is
161 produced – an approach similar to the Illumina 16S sequencing protocol
162 (http://res.illumina.com/documents/products/appnotes/appnote_miseq_16s.pdf). Compared to
163 the protocol by Sickel et al. (2015), our approach uses one additional PCR reaction, but it is
164 more flexible and cost-effective because the sequenced DNA barcode can be changed just by
165 modifying the two primers used in the first PCR reaction. Moreover, in the second PCR

reaction we use primers compatible with standard Illumina sequencing protocol, which simplifies the sequencing step and allows sequencing the libraries along with other standard libraries. We also contribute a bioinformatic pipeline to analyse and classify the obtained reads.

Materials and methods

Sampling

We monitored seven sites along the Mediterranean coast of Spain to test for potential arrivals of *Vanessa cardui* from the African continent (Fig. 1, Tab. S1). Sampling was designed to capture specimens with high probability to be in migratory phase. To do that, we sampled sites where the species was unlikely to be found while nectaring or breeding. In particular we sampled points in the Mediterranean shorelines, usually in the sandy beaches, cliffs or inspecting the vegetation nearby the coast. Timing was also chosen within the time frame where *V. cardui* arrivals are expected to colonize the Iberian Peninsula (February-April), and when consistent wind patterns or storms from Africa occurred, that could aid insect migrations. All the samples were immediately bagged in glassine envelopes that were sealed and stored at -20°C until pollen isolation and library preparation.

NGS library construction

Pollen isolation and library construction were performed in four batches (see Tab. S2). We amplified the Internal Transcribed Region 2 (ITS2) of the ribosomal DNA by using a

combination of ITS-S2F (Chen et al., 2010) and ITS-4R (White et al., 1990) primers, tested in other pollen metabarcoding studies (Keller et al., 2015; Sickel et al., 2015). In the first step, the ITS2 fragment was amplified using the above primers tailed with technical sequences: six random nucleotides to increase sequence diversity during the first sequencing cycles, and a part of the Illumina adapter. In the second reaction, we used modified Illumina TruSeq primers (Tab. 1) in order to index the samples and produce the final library. We used index sequences from TruSeq Amplicon series, allowing the pool of 96 samples on one lane. These indexes can be supplemented by TruSeq CD index sequences (see https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-03.pdf) as in Sickel et al. (2015), for a total of 384 combinations. All the steps of the protocol, from pollen isolation up to the first PCR reaction were conducted under a laminar flow cabinet in a pre-PCR area in order to avoid external contamination. All the working spaces and equipment were cleaned with 10% bleach solution and 70% ethanol before and after work. We used only ultra-pure DEPC-treated water and a separate stock of reagents and plastics dedicated solely to metabarcoding work. Moreover, library preparation steps for all the samples and pollen isolation from most of the samples (batches 1-3) were conducted in a laboratory located far away from the sampling sites (at W. Szafer Institute of Botany, Polish Academy of Sciences, Kraków, Poland). Pollen isolation from batch 4 was performed at Institut de Biologia Evolutiva, Barcelona, Spain.

After preliminary trials (not shown), we decided not to extract DNA but to use Phire Plant Direct Polymerase (Thermo Fisher Scientific, Waltham, MT, USA) in the first PCR step,

which successfully amplified ITS2 from pollen mix without DNA extraction. Pollen was collected by vortexing butterfly bodies (with wings removed) in a 2 ml tube with 50 µl of sterile water with 0.1% SDS, centrifugation of the obtained solution and drying it under vacuum. The obtained pellet was diluted in 15 µl of Phire Plant Direct sample buffer and homogenized with five zirconium beads on the TissueLyser II machine (Qiagen, Hilden, Germany) at 30 Hz for 1 min before proceeding to the PCR. We added blank sample to each of the four extraction batches ("extraction blank samples") – a 1.5 ml tube filled with 50 µl of water, left open during the whole extraction procedure and processed like a normal sample. We processed each sample in three independent PCR reactions to avoid reaction-specific biases (Fierer, Hamady, Lauber & Knight, 2008; Sickel et al., 2015) with another four blank samples added at the PCR step ("PCR blank samples"). Each of the replicate PCR reactions consisted of 1 µl of the disrupted pollen sample, 25 µl of Phire Plant Direct Polymerase Mix, and 0.5 µM of each primer in 50 µl reaction volume with the following PCR program: 98°C for 5 min, 20 cycles of denaturation at 98°C for 40 s, annealing at 49°C for 40 s and elongation at 72°C for 40 s, final extension step at 72°C for 5 min. After the reaction, we combined the PCR triplicates and purified the product using 1x ratio of AMPure XP (Beckman Coulter, Indianapolis, IN, USA) and eluted in 10 µl of water. In the second reaction, we indexed each sample using a unique combination of primers. The second reaction, also performed in triplicates for each sample, consisted of 1 µl of the purified PCR product, 1x Q5 buffer, 0.2 U of Q5 Hot-Start polymerase, 0.5 µM of forward, and 0.5 µM of reverse indexed primer in 10 µl reaction volume. We amplified the reaction using a PCR program as follows: 30 s initial denaturation at 98°C; 12 cycles of denaturation at 98°C for 10 s and combined annealing and extension at 72°C for 30 s (shuttle PCR); final extension

at 72°C for 5 min. After the reaction, we combined the triplicates and verified the reaction success on TapeStation 4200 (Agilent, Santa Clara, CA, USA). We then pooled 10 µl of the PCR product from each sample and purified the pool using 1x ratio of AMPure XP. After that, the pooled library was quantified using the Qubit instrument (Thermo Fisher Scientific, Waltham, Massachusetts, USA), and sequenced with 15% PhiX spike-in on Illumina MiSeq (San Diego, CA, USA) using 600-cycle MiSeq Reagent Kit v3, according to the manufacturer's instructions.

Data analysis

We merged the raw paired-end reads using PEAR v0.9.8 (Zhang, Kobert, Flouri & Stamatakis, 2014) and retained only the successfully merged reads. We then trimmed the primer sequences with remove_primers script (www.biopieces.org). The obtained sequences were processed using vsearch v2.4.3 (Rognes, Flouri, Nichols, Quince & Mahé, 2016): filtered by an expected error rate (*maxe* parameter) of 0.5, minimum length of 250 nt, maximum length of 450 nt, and removing the reads with ambiguous nucleotides. Next, the reads were dereplicated and singletons removed. We classified the filtered reads with SINTAX (from the USEARCH package v10.0.240; Edgar, 2016a) using the ITS2 database of Sickel et al. (2016), modified for SINTAX by parsing with a custom script. Only reads classified to the species level with a probability of 95% or above were further considered. The classified reads were then summarized by species with a minimum threshold of 100 reads per plant species per sample. The raw sequences, the scripts used to process the data, and ITS2 database are available online (see Data Availability).

Plant distributions

Geographical distributions of all detected plants were compiled checking specific literature and occurrences available in the online databases GBIF (www.gbif.org) and the African Plant Database (www.ville-ge.ch/musinfo/bd/cjb/africa/recherche.php) (Supplementary Table S1). The plants with the geographical range not including the butterfly sampling sites were further examined, as potential marker candidates to unravel migratory paths. These mainly included Saharan-North African endemics and African species with ranges reaching southern Mediterranean (Figure S1). The extent of occurrence for each informative plant species was delimited on the maps by the minimum convex polygon containing all known sites of occurrence, according to IUCN criteria (IUCN, 2012).

Wind track

Backward wind trajectories were reconstructed for sampling sites and dates, using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) dispersion model from NOAA's Air Resources Laboratory (ARL) (Stein et al., 2015; Rolph, Stein & Stunder, 2017). Analyses were based on the Reanalysis database and computed on 48-h back trajectories arriving at sites at 12:00 h UTC and for three different altitudinal layers (500m, 1000m and 1500m asl).

Results

Sampling

A total of 47 *V. cardui* specimens were collected in the seven sites (Fig. 1; Tab. S1) and processed for pollen metabarcoding analysis. The sampling consisted mostly of two migratory groups, according to the regions (Andalusia and Catalonia) and timing (February and April). First, a group of samples was collected in several localities of Andalusia between February 16th and 25th of 2016, a year with exceptionally early sightings of the species in the region. The individuals from this group were generally collected close to the beach, but not exactly at the moment of arrival, although a noticeable increase of individuals was observed on the 21st of February. A second group of samples were collected at the beach of Delta del Llobregat, near Barcelona, along two consecutive days (27th and 28th) in April of 2012, after a storm carrying Saharan dust and strong winds from the south. The individuals from this group were collected when landing on the beach, coming from the sea in a northwest direction. Finally, one sample was collected in June 2012 from Columbretes Islands, a small archipelago 50 km from Castelló coast.

Sequencing results and reads processing

We obtained 9,654,286 raw paired-end reads, of which only 1420 (0.0147%) were present in the blank samples. The average number of reads per sample was 205,380 (excluding blank samples; range: 7-608,640), of which 98.7% were successfully merged (Tab. S2). After quality filtering, we retained a total of 7,315,458 reads remained, of which 1,236,996 (16.9%)

were discarded as singletons. The rest of the reads were dereplicated (separately within each sample) into 229,295 unique sequences of which 35,393 (15.4%; that is 2,908,006 reads or 47.8% of all the filtered reads) were classified to the species level with probability ≥ 95%. The proportion of the filtered reads classified to higher taxonomic ranks was much higher: 99.9% to the division, 78.3% to the order, and 65.7% to the family level (Fig. S2). After filtering out the plant species represented by less than 100 reads in a butterfly sample, 2,880,443 reads were retained (47.4% of the filtered reads; Tab. S2).

Sequencing success was not even among the butterfly samples but rather followed a bimodal distribution (Fig. S1a) – 17 samples yielded less than 10,000 raw sequences (mean = 1,041; range: 7-7,702) but the remaining 30 samples had high number of reads (mean = 321,172; range: 37,420-608,640; Tab S2). We also detected variation among the four sample batches: samples from batches number 1 and 3 had a low coverage, as did five out of 11 samples from batch 2. In contrast, most of the samples of batch 4 had a high coverage (Fig. S1b; Tab. S2). Sequences assigned to the species level with high probability and coverage higher than 100 reads were found in 30 out of 47 (63.8%) butterfly samples, from all the batches except number one but only in the samples with high the number of reads higher than 10,000. No reads were retained after the filtering steps for the sample from Columbretes Islands (the only one not attributable to Andalusia or Catalonia).

The amount of PCR product in blank samples was too low to be visible in TapeStation profiles. During the sequencing, PCR blank samples yielded from 0 to 5 reads. Extraction blank samples had between 3 and 26 raw reads, except EB4 from which we sequenced 1,366 reads. After processing the latter, we detected 42 unique sequences, of which only three were

309 classified to the species level with probability ≥ 95%, belonging to *Moricandia*
310 *moricandioides* (11 reads in total) and *Prunus dulcis* (27 reads). These two plant species were
311 detected in some of the studied butterflies and are not present in the area where the libraries
312 were prepared (Kraków, Poland), pointing rather to a low level of cross-contamination than to
313 external contamination in one of the batches. As these sequences were well below the
314 threshold of 100 reads per species, they were not retained in the final dataset.

315 ## Sequence classification and plant diversity

316 In total, the filtered reads were classified to 157 species (Table S3). The sequence most
317 frequently represented in the samples (present in 21 samples) was classified as *Alternanthera*
318 sp. This appears to be an erroneous GenBank entry (JX136744.1), most likely of fungal origin
319 and it was removed from our dataset. Excluding this, the most represented species (in 12 out
320 of 30 butterfly samples for which plant sequences were detected) was *Reseda lutea* a
321 widespread species in Europe and the Mediterranean region, including both sampling areas
322 (Andalusia and Catalonia). Approximately half of the plant species (76) were present only in
323 a single butterfly sample each and another 42 species in just two butterfly samples each. On
324 average, there were 12.2 species detected per butterfly sample (SD: 6.9; range: 2-26).

325 The diversity of detected pollen included species from 23 orders of plants in total, where 13
326 orders were represented in samples from Catalonia (April) and 23 orders in Andalusia
327 (February). Asterales were the dominant order at both sites, with 31 species: 14 and 21
328 species in the two above regions, respectively. Most of the plants detected were typically
329 insect-pollinated (82.8%), while 15.3% were species not pollinated by insects. The proportion

of insect pollinated species was higher at both sites: 78.7% Catalonia and 85.6% in Andalusia (Fig. 2).

Sequences from all green plants (Viridiplantae) were present in the reference database and some sequences from our data matched green algae (Chlorophyta). Sequences from two samples were assigned to the algae genus of *Trebouxia*. Another species of green algae, *Pseudostichococcus monallantoides* was also detected in one sample.

Of the plant species detected from sequencing data, 40 were alien to the sampling sites (24 for Catalonia and 34 for Andalusia) and are thus potentially informative in estimating migratory paths (listed and illustrated in Tab. S3 and Fig. S4). From these, 33 species are present in Northern Africa and absent in Europe, and these were detected in 20 of our samples (Tab. S3). Six were present in more than two butterfly samples: *Calendula stellata* (9 butterfly samples) – occurring in Morocco, Algeria, Tunisia and Sicily, introduced in the Canary Islands; *Launaea mucronata* (6 samples) – a species of Saharo-Arabian distribution, also present in the Canary Islands; *Lotus weilleri* (4 samples) – a species endemic to northern Atlantic Morocco; *Oxalis compressa* (4 samples) – native to Southern Africa but introduced in the Mediterranean Region of Morocco and Algeria; *Raffenaldia primuloides* (4 samples) – present in Morocco and Algeria; *Farsetia aegyptia* (3 samples) – a desert species distributed throughout Northern Africa (Morocco, Algeria and Tunisia but absent from Libya and Egypt) and Asia (Tab. S3, Fig. S4).

Discussion

Pollen metabarcoding as a tool

Morphological identifications of pollen grains carried by insects have already been used to infer long-distance migratory patterns in insects (Mikkola, 1971; Hendrix et al., 1987; Hendrix & Showers, 1992; Gregg, 1993; Lingren et al., 1993, 1994; Westbrook et al., 1997). The use of this method, however, is limited because morphological identification by light microscopy is a time-consuming task, it requires specialized taxonomic expertise, and can hardly provide species-level determinations (Galimberti et al., 2014; Hawkins et al. 2015; Keller et al. 2015; Kraaijeveld et al. 2015; Richardson et al., 2015b; Sickel et al., 2015). DNA metabarcoding is a fast, high throughput method that greatly simplifies the identification process (Taberlet et al., 2012) and captures high diversity of pollen that is transported by insects. Our results show that identifying pollen grains carried by migrating insects through DNA metabarcoding is feasible to the species level and that high pollen diversity per specimen can be detected. Thanks to species-level identifications for multiple taxa, migratory paths of the insects can be traced and narrowed by additive geographic distributions of the plants. The method presented here has a wide application to all major insect orders that visit flowers (Coleoptera, Diptera, Hymenoptera and Lepidoptera; Kevan & Baker, 1983), as well as some vertebrates such as birds (Cronk & Ojeda, 2008) or bats (Fleming, Geiselman & Kress, 2009).

We show that the *V. cardui* individuals analysed here were migrating and originated in the African continent. Among the samples from which pollen was detected, all but three were

370 carrying pollen attributed to plant species alien to their collecting site, most of it
371 corresponding to African endemic plants. *Vanessa cardui* is typically colonizing the western
372 Mediterranean Europe in spring, with most arrivals usually observed in April and May
373 (Stefanescu et al., 2013). The origin of spring arrivals into the Iberian Peninsula has typically
374 been associated to large breeding sources found in Morocco in March-April (Stefanescu,
375 Alarcón, Izquierdo, Páramo & Ávila, 2007; Stefanescu, Páramo, Åkesson, Alarcón & Ávila,
376 2011). Our results partially agree with this timing and path. On the one hand, we show that
377 arrivals to Europe can occur as early as February in Andalusia. On the other hand, the pollen
378 associated to butterflies exhibits a much larger geographical range where butterflies might
379 originate and transit (Fig 2). Unlike isotopic analyses, our approach does not test for natal
380 origins of migrations, but for the most likely paths used during their migrations. During
381 migration, butterflies generally stop in the evening, feed and rest at night. In the morning,
382 they feed until it is warm enough and winds are suitable for continuing the migration (e.g.,
383 Shields, 1974). So, migratory paths would be defined by the position of stepping-stone
384 locations where they fed. In fact, it is expected that pollen from most recently visited flowers
385 is better represented on butterfly bodies than that from plants visited right after emergence.
386 Thus, we may assume a "dilution effect" of the signal to a certain degree.

387 ## Sequencing protocol

388 The flexible two-step laboratory protocol here presented can be easily adapted to other types
389 of markers, for instance standard plant chloroplast barcodes such as rbcL and matK (CBOL
390 Plant Working Group et al., 2009), just by replacing the primers in the first PCR reaction.
391 Importantly, the proposed method allows for high sample multiplexing: up to 384 samples

392 can be analysed simultaneously when combined with Illumina TruSeq CD index sequences,
393 and even more when longer index sequences are used (e.g., Fadrosh et al., 2014). Previous
394 pollen metabarcoding protocols used pollen pulverization with bullet blender and DNA
395 isolation with commercial kits (Simel, Saidak & Tuskan, 1997; Kraaijeveld et al., 2015).
396 Assessing the best pollen isolation method was beyond the scope of this study, but we found
397 that skipping the pollen isolation step and using homogenized pollen directly in the PCR
398 reaction with "direct" polymerase mix (Wong et al., 2014) is an efficient method of
399 amplifying DNA markers from pollen loads. This way, DNA isolation – the most time-
400 consuming step of pollen metabarcoding projects (Bell et al., 2017) – is avoided altogether.
401 Nevertheless, further assessment of the effectiveness of such procedure in amplifying markers
402 from all the plant species present in the pollen loads is necessary. Moreover, careful
403 contamination and cross-contamination control by using blank samples and following best
404 practices to avoid contamination are necessary in metabarcoding studies (Goldberg et al.,
405 2016; Deiner et al., 2017). In our case, we used blank samples both at the pollen isolation and
406 the library preparation steps. Both isolation and pre-PCR steps were also conducted under
407 laminar flow cabinet. These remedial steps ensured no external contamination in our samples,
408 as shown by blank samples and the virtual absence of plants native to Central Europe in our
409 dataset.

410 Another source of bias are the errors occurring at the PCR and sequencing step (Coissac,
411 Riaz, & Puillandre 2012). Many metabarcoding pipelines perform clustering of similar
412 sequences in order to reduce the number of low-copy reads that are usually erroneous
413 artefacts and cluster them together with the centroid sequence (e.g., Edgar, 2010; Rognes et

al., 2016) or 'denoising' the reads in order to remove the putatively erroneous sequences (e.g., Edgar, 2016b). Studies with mock pollen samples are still needed to assess the relative performance of these methods for pollen metabarcoding.

Many factors can bias the PCR amplification of DNA templates and skew the quantitative representation of the sequenced species in the obtained reads. These factors involve PCR biases, either caused by sequence polymorphisms in priming sites (Sipos et al., 2007; Elbrecht & Leese, 2015; Piñol, Mir, Gomez-Polo & Agust, 2015), formation of chimeric reads (Bjørnsgaard Aas, Davey & Kauserud, 2017), sequence length or GC content (Krehenwinkel et al., 2017). Although some studies show some relationship between pollen abundance and the number of reads (Kraaijeveld et al., 2015; Pornon et al., 2016), other did not find such relationship (Hawkins et al., 2015; Richardson et al., 2015b). The number of reads obtained per species should be therefore treated with caution and only as a semi-quantitative method of estimating pollen abundance. Thus, following Yu et al. (2012), we used only presence-absence information when interpreting our results.

Limitations to track migrations: taxonomic assignments and species distributions

Several key factors determine the accuracy and resolution of our method for studying insect migrations. First, reference sequences from correctly determined plant species are necessary to properly classify the obtained reads. Uneven geographical coverage of sequences present in reference databases, with a bias towards better studied areas such as Europe or Northern America (Ankerbrand, Keller, Wolf, Schultz & Förster, 2015; Bell et al., 2016), is of a special concern. Long-distance migration studies would benefit largely from global plant species

435 coverage, which still remains a remote prospect. In our case, although we were able to
436 classify many reads with high probability, a high proportion (52.6%) of unclassified reads
437 remained. This is most likely caused by a number of African species/populations missing in
438 our reference database and therefore classified with low probability, as much larger
439 proportion of reads was classified into higher taxonomic ranks (Fig. S2). Representation of
440 species in databases and taxonomic errors are especially problematic with best-hit approaches
441 (e.g. BLAST; Boratyn et al., 2013), nevertheless such methods are still used in metabarcoding
442 studies (Hawkins et al., 2015; Kraaijeveld et al., 2015; Richardson et al., 2015b). Sequence
443 classifiers, such as SINTAX used here, are more robust in such cases. Nevertheless, all
444 current methods display high over-classification rates in cases when taxa are missing from
445 reference databases (Edgar, 2016a). So far, above classification methods have been mostly
446 benchmarked for bacterial sequences (e.g., Vinje, Liland, Almøy & Snipen, 2015) and more
447 studies are needed to assess their comparative performance for ITS2 metabarcoding of plant
448 sequences.

449 In this study, we did not prioritize a full assessment of plant pollen present on the migrating
450 butterflies and, in order to reduce false positives, we used a conservative 95% sequence
451 assignment threshold for the classified reads to be retained in a final dataset. Despite such
452 conservative approach, we could still detect a small number of species that are probably
453 assignment errors, i.e. with geographical distributions outside the possible migratory routes of
454 *Vanessa cardui*. This is probably because the representation of plant species belonging to
455 taxonomically complex and diverse genera is far from complete in reference databases used
456 for taxonomic assignments and could prevent the positive identification of some pollen

457 grains. Some examples of such taxa in our dataset include: *Astragalus*, with about 2500–3000
458 species, is the largest genus of flowering plants (Podlech & Zarre, 2013; Bagheri, Maassoumi,
459 Rahiminejad, Brassac & Blattner, 2017); *Artemisia* that covers approximately 600 species
460 (Richardson, Page, Bajgain, Sanderson & Udall, 2012) and genus *Thymus* that includes
461 around 400 species (Karaca, Ince, Aydin, Elmasulu & Turgut, 2015). Most of the species of
462 these genera are native to the Mediterranean region, Northern Africa, and Western Asia.
463 Moreover, molecular approaches have limitations to identify and define species in some of
464 these complex genera due to various biological phenomena, such as interspecific
465 hybridization and polyploidy, which are often correlated (Soltis & Soltis, 2009), can
466 contribute significantly to the taxonomic complexity of *Thymus* (Morales, 2010), *Artemisia*
467 (Richardson et al., 2012) and *Astragalus* (Doyle, 2012; Bagheri et al., 2017). In *Thymus*,
468 genetic polymorphism at the intraspecific level can hinder the positive identification of some
469 species (Karaca et al., 2015). Conversely, DNA sequence diversity is generally very low in
470 some *Artemisia* species (Koloren, Koloren & Eker, 2016) and also among species included
471 within several sections of *Astragalus* (Bagheri et al., 2017), which has been attributed to a
472 rapid radiation (Sanderson & Wojciechowski, 1996).

473 In order to infer the migration routes from metabarcoding sequences, detailed plant
474 distribution data are required. Evident gaps in the distribution of the plant species detected in
475 North Africa exist. In particular, presence data available for Algeria and Libya are extremely
476 poor when compared to Morocco or Tunisia. Such biases preclude more detailed analyses
477 based on actual presence records or geographical grids (see Fig. S3 for results based on a
478 presence grid). To avoid the influence of important gaps in presence data, distribution ranges

479 delimited by peripheral presence records may be used (Fig. 2). The abovementioned
480 limitations point to the importance of basic taxonomic, barcoding and floristic research, which
481 is the cornerstone for myriad of studies.

482 More research is needed on pollen retention on insects. For instance, DelScorro and Gregg
483 (2001) found that the sunflower pollen is a transient marker that is only informative of plant
484 visits that occurred during the previous two days. On the other hand, some studies, including
485 the one presented here, show support for long-distance pollen transport (Hendrix & Showers,
486 1992; Ahmed et al., 2009). In this line, it is worth noting that no informative data was
487 retrieved from a percentage of butterfly specimens (36%). In any case, pollen grains are
488 probably lost along time, and a dilution effect of the pollen load signal is to be expected, with
489 a higher representation of recently visited flowers.

490 Despite the abovementioned limitations, we have proven that pollen metabarcoding is an
491 effective and informative method for tracking insect movements, despite the current limited
492 resolution due to completeness of reference genetic libraries and plant species presence record
493 databases. The accumulation of this kind of information grows rapidly thanks to the new
494 sequencing techniques and citizen science initiatives, for example. As these two factors will
495 most likely improve in the near future, the resolution and usefulness of pollen metabarcoding
496 as a tool for tracking insect migrations can only increase.

497 ## Pollen detected and migrations of *Vanessa cardui*

498 Using our metabarcoding approach, we were able to amplify a wide range of plant DNA
499 sequences from migrating *Vanessa cardui*. We generally collected the butterflies immediately

500 or soon after they landed on the beaches of the Mediterranean (note that in several instances
501 we cannot discard that they fed on local flowers). Our study was therefore designed to test the
502 feasibility of pollen detection after a long-distance migration. In this particular case, we
503 expected to detect pollen of plant species present in Africa, including the Maghreb, the Sahara
504 and the sub-Sahara. According to our hypothesis, large proportion of reads was classified as
505 African or African-Arabian endemic plants (21.0%), or more generally, plants that were not
506 native to the butterfly sampling sites (25.5%).

507 Pollen composition may explain individual migratory histories, but it can also report
508 collective migratory histories given a time and site. Butterflies collected from the same
509 spatiotemporal migratory waves (Andalusia in February and Catalonia in April) show
510 parallelisms, but also differences, in their carrying pollen composition. Such differences
511 among specimens could be explained by variability of the visited flowers and in the retention
512 of pollen grains. In some cases, though, the plants found in different specimens are
513 geographically exclusive, which suggest that, if taxonomic attributions are correct, the
514 butterflies may have had different breeding origins and confluenced *a posteriori* during their
515 migratory paths or at destination.

516 The two waves of migrants studied (Andalusia and Catalonia) could either correspond to
517 populations originated during the winter in the Maghreb, or to populations originated in
518 tropical Africa that may replenish the temperate zone in early spring (Talavera & Vila, 2016).
519 The latter hypothesis cannot be excluded based on our data because, in addition to a marked
520 influence from Maghreb flora, several floristic elements from the Sahara and Sub-Saharan
521 Africa were detected (Fig. 2). Generally, the butterflies collected in February (Andalusia)

522 showed a higher number of plants of predominantly Saharan distribution (*Farsetia stylosa*,
523 *Launaea capitata*, *Launaea mucronata*, *Moltkioptis ciliata*, *Reseda villosa*, *Gymnocarpus*
524 *decandrus*, *Euphorbia guyoniana*), or with an important sub-Saharan representation
525 (*Pergularia tormentosa*, *Musa acuminata*). Pollen of *Musa* (banana), only detected in
526 specimen 16C413, is extensively cultivated in tropical Africa and is abundant in the Canary
527 Islands, but rarer in the Maghreb and Europe (where is generally cultivated in greenhouses).
528 In addition, this plant is apomictic and although some varieties preserve male flowers, these
529 are very rare. As the probability to find male banana flowers is higher where the plant is
530 common, a sub-Saharan origin for this sample is likely.

531 Butterflies from April (Catalonia) had few strictly Saharan floristic elements, such as
532 *Launaea capitata* and *L. mucronata*, and mostly had representation of flora from the Maghreb
533 (Fig. 2). Pollen endemic to the Canary Islands was detected in some instances, both from
534 February and April. Thus, an origin in these islands cannot be discarded for some specimens.

535 Backtrack wind models agree with migratory paths coming from the African continent (Fig.
536 2), considering that *V. cardui* migrations can greatly be aided by winds (Stefanescu, Alarcón
537 & Ávila, 2007). Winds consistently came from the south-east through central Algeria in both
538 waves, at least during the previous 48h. Precise migratory sources or paths cannot be inferred
539 based on pollen data or winds alone, but both can be combined to narrow predictions.
540 Actually, the Algerian Maghreb (for the Catalonia April migratory wave) and the Algerian
541 Sahara (for the Andalusia February wave) are the areas where higher accumulative
542 probabilities of identified plants overlap. The combined evidence points to a highly probable

543 origin or pass of these two migratory waves across Algerian grounds, although some
544 individuals may have followed different paths.

545 Insect-pollinated plants prevailed in our results (82.8%), as expected in pollen obtained from
546 butterfly bodies. Worth noting, plants detected belonged primarily to the Asterales, an order
547 that includes many of *V. cardui* typical host plants and nectar sources (Nylin, Slove & Janz,
548 2014). In fact, the spectrum of plants visited by *V. cardui* is very wide, including most plant
549 orders and life forms from small plants to trees. For example, numerous *V. cardui* specimens
550 migrating northwards in February 2017 in south Morocco were observed to stop briefly to
551 feed on the flowers of *Prunus dulcis* orchards (R. Vila, pers. obs.), a tree for which we
552 detected pollen in seven butterfly specimens.

553 Most non-insect pollinated plants detected were wind-pollinated plants that belonged to the
554 Poales. This points to the accidental transport of such pollen due to physical contact, as *V.*
555 *cardui* tends to rest on the ground and grass, and only rarely on trees. Sequences from two
556 samples were assigned to the algae genus *Trebouxia*, a common and widespread photobiont in
557 lichens (Dal Grande et al., 2014). Lichens often reproduce asexually by soredia – small
558 powdery propagules containing both fungus and algae. Such small structures could possibly
559 stick to the body of butterflies accidentally and be transported on larger distances.

560 ## Implications for plant pollination

561 Here we show that *Vanessa cardui* could potentially mediate transcontinental pollination, and
562 thus gene flow, for plants species that occur in both Europe and Africa. Given the millions of
563 individuals of *V. cardui* in particular, and of insects in general, that migrate every year across

564 continents (Hu et al. 2016), the effects of this phenomenon on ecosystems and crops may be
565 not negligible.

566 Regular long-distance gene flow is a phenomenon that has rarely been acknowledged, but that
567 could potentially explain particular phylogeographic patterns in plants. For example, the Strait
568 of Gibraltar has been shown as not effective at interrupting gene flow in *Androcymbium*
569 *gramineum* (Caujapé-Castells & Jansen, 2003), several *Cistus* spp. (Fernández-Mazuecos &
570 Vargas, 2010), *Hypochaeris salzmanniana* (Ortiz, Tremetsberger, Talavera, Stuessy &
571 Garcia-Castano, 2007), and *Rosmarinus officinalis* (Mateu-Andrés et al., 2013). It is also
572 worth noting that we detected sequences belonging to crops in our dataset – *Allium sativum*,
573 *Cucumis sativus*, and *Prunus dulcis* – which points to the possibility of intercontinental
574 pollination for these economically important species. We suggest that knowledge on the
575 insect species that perform long-distance migration and the routes and temporal patterns they
576 follow may be of high importance for better understanding intercontinental plant pollination,
577 with implications extending from plant phylogeography to ecosystem services.

578 Acknowledgements

579 We thank Martin Gascoigne-Pees for important sample contribution. TS and MR are
580 supported by the statutory funds of W. Szafer Institute of Botany, Polish Academy of
581 Sciences. GT is supported by the Marie Curie Actions FP7-People-2013 IOF (project 622716)
582 and by the MINECO programme Juan de la Cierva Incorporación (IJCI-2016-29083).
583 Funding was provided by the British Ecological Society (grant LRB16/1015) to GT, by

584 project CGL2016-76322 (AEI/FEDER, UE) and by the Committee for Research and
585 Exploration of the National Geographic Society (grant WW1-300R-18) to GT and RV.

586 Data accessibility

587 Scripts used to process the data: https://github.com/TomaszSuchan/pollen-metabarcoding

588 ITS2 database for SINTAX: https://github.com/molbiodiv/meta-barcoding-dual-
589 indexing/blob/master/precomputed/viridiplantae_all_2014.sintax.fa

590 Raw sequences: European Nucleotide Archive,
591 http://www.ebi.ac.uk/ena/data/view/PRJEB26439

592 Author contributions

593 TS, GT and RV conceived the study and wrote the manuscript. GT and RV collected samples.
594 TS carried out laboratory work. TS and GT analyzed data. LS gathered plant distribution data.
595 All authors edited and approved the final version of the manuscript.

596 References

597 Ahmed, S., Compton, S. G., Butlin, R. K., & Gilmartin, P. M. (2009). Wind-borne insects
598 mediate directional pollen transfer between desert fig trees 160 kilometers
599 apart. *Proceedings of the National Academy of Sciences*, *106*(48), 20342-20347.
600 Ankenbrand, M. J., Keller, A., Wolf, M., Schultz, J., & Förster, F. (2015). ITS2 database V:
601 Twice as much. *Molecular Biology and Evolution*, *32*(11), 3030-3032.

602 Bagheri, A., Maassoumi, A. A., Rahiminejad, M. R. Brassac, J. & Blattner, F.R. (2017).
603 Molecular phylogeny and divergence times of *Astragalus* section *Hymenostegis*: An
604 analysis of a rapidly diversifying species group in Fabaceae. Scientific Reports 7,
605 14033.

606 Bauer, S. & Hoye, B. J. (2014). Migratory animals couple biodiversity and ecosystem
607 functioning worldwide. *Science, 344*, 1242552.

608 Bell, K. L., de Vere, N., Keller, A., Richardson, R. T., Gous, A., Burgess, K. S., Brosi, B. J.
609 (2016). Pollen DNA barcoding: current applications and future prospects. *Genome,*
610 *59*, 629-640.

611 Bell, K. L., Fowler, J., Burgess, K. S., Dobbs, E. K., Gruenewald, D., Lawley, B., ... & Brosi,
612 B. J. (2017). Applying pollen DNA metabarcoding to the study of plant–pollinator
613 interactions. *Applications in Plant Sciences*, *5*(6), 1600124.

614 Bjørnsgaard Aas, A., Davey, M. L., & Kauserud, H. (2017). ITS all right mama: investigating
615 the formation of chimeric sequences in the ITS2 region by DNA metabarcoding
616 analyses of fungal mock communities of different complexities. *Molecular Ecology*
617 *Resources, 17*(4), 730-741.

618 Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., ... & Raytselis,
619 Y. (2013). BLAST: a more efficient report with usability improvements. *Nucleic*
620 *Acids Research, 41*(W1), W29-W33.

621 Brattström, O., Bensch, S., Wassenaar, L. I., Hobson, K. A. & Åkesson, S. (2010).
622 Understanding the migration ecology of European red admirals *Vanessa atalanta*
623 using stable hydrogen isotopes. *Ecography, 33*, 720-729.

624 Bruni, I., Galimberti, A., Caridi, L., Scaccabarozzi, D., De Mattia, F., Casiraghi, M., & Labra,
625 M. (2015). A DNA barcoding approach to identify plant species in multiflower
626 honey. *Food Chemistry, 170*, 308-315.

627 Caujapé-Castells, J., Jansen, R.K. (2003). The influence of the Miocene Mediterranean
628 desiccation on the geographical expansion and genetic variation of *Androcymbium*
629 *gramineum* (Cav.) McBride (Colchicaceae). *Molecular Ecology, 12*, 1515–1525.

630 CBOL Plant Working Group, Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei,
631 M., Ratnasingham, S., ... & Fazekas, A. J. (2009). A DNA barcode for land
632 plants. *Proceedings of the National Academy of Sciences*, *106*(31), 12794-12797.

633 Cronk, Q., & Ojeda, I. (2008). Bird-pollinated flowers in an evolutionary and molecular
634 context. *Journal of Experimental Botany, 59*(4), 715-727.

635 Chapman, J. W., Reynolds, D. R., Smith, A. D., Riley, J. R., Pedgley, D. E., & Woiwod, I. P.
636 (2002). High-altitude migration of the diamondback moth *Plutella xylostella* to the
637 UK: a study using radar, aerial netting, and ground trapping. *Ecological*
638 *Entomology*, *27*(6), 641-650.

639 Chapman J. W., D. R. Reynolds, A. D. Smith, E. T. Smith, I. P. Woiwod, (2004) An aerial
640 netting study of insects migrating at high altitude over England. *Bulletin of*
641 *Entomological Research, 94*, 123-136.

642 Chapman, J. W., Drake, V. A., & Reynolds, D. R. (2011). Recent insights from radar studies
643 of insect flight. *Annual Review of Entomology*, *56*, 337-356.

644 Chapman J. W., Reynolds, D. R., Wilson, K. (2015). Long-range seasonal migration in
645 insects: mechanisms, evolutionary drivers and ecological consequences. *Ecology*
646 *Letters, 18*, 287-302.

647 Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., ... & Luo, K. (2010). Validation of the
648 ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS*
649 *ONE, 5*(1), e8613.

650 Chilson, P. B., Frick, W. F., Kelly, J. F., Howard, K. W., Larkin, R. P., Diehl, R. H., ... &
651 Kunz, T. H. (2012). Partly cloudy with a chance of migration: weather, radars, and
652 aeroecology. *Bulletin of the American Meteorological Society, 93*(5), 669-686.

653 Coissac, E., Riaz, T., & Puillandre, N. (2012). Bioinformatic challenges for DNA
654 metabarcoding of plants and animals. *Molecular Ecology, 21*(8), 1834-1847.

655 Dal Grande, F., Alors, D., Divakar, P. K., Bálint, M., Crespo, A., & Schmitt, I. (2014).
656 Insights into intrathalline genetic diversity of the cosmopolitan lichen symbiotic
657 green alga *Trebouxia decolorans* Ahmadjian using microsatellite markers. *Molecular*
658 *Phylogenetics and Evolution, 72*, 54-60.

659 De Vere, N., Jones, L. E., Gilmore, T., Moscrop, J., Lowe, A., Smith, D., ... & Ford, C. R.
660 (2017). Using DNA metabarcoding to investigate honey bee foraging reveals limited
661 flower use despite high floral availability. *Scientific Reports, 7*, 42838.

662 Dean, T. J., & Drake, V. A. (2005). Monitoring insect migration with radar: the ventral-aspect
663 polarization pattern and its potential for target identification. *International Journal of*
664 *Remote Sensing, 26*(18), 3957-3974.

665 Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ...
666 & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how
667 we survey animal and plant communities. *Molecular Ecology, 26*(21), 5872-5895.

668 Doyle, J. J. (2012). Polyploidy in Legumes. In Soltis, P. S., & Soltis, D. E. (eds.), *Polyploidy*
669 *and genome evolution* (pp. 147-180). Berlin: Springer.

670 Del Socorro, A. P., & Gregg, P. C. (2001). Sunflower (*Helianthus annuus* L.) pollen as a
671 marker for studies of local movement in *Helicoverpa armigera* (Hübner)
672 (Lepidoptera: Noctuidae). *Austral Entomology*, *40*(3), 257-263.

673 Drake, V. A., Chapman, J. W., Lim, K. S., Reynolds, D. R., Riley, J. R., & Smith, A. D.
674 (2017). Ventral-aspect radar cross sections and polarization patterns of insects at X
675 band and their relation to size and form. *International Journal of Remote*
676 *Sensing*, *38*(18), 5022-5044.

677 Edgar, R. C. (2010). Search and clustering orders of magnitude faster than
678 BLAST. *Bioinformatics*, *26*(19), 2460-2461.

679 Edgar, R. C. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS
680 sequences. *BioRxiv*, 074161.

681 Edgar, R. C. (2016b). UNOISE2: improved error-correction for Illumina 16S and ITS
682 amplicon sequencing. *BioRxiv,* 081257.

683 Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species
684 abundance? Testing primer bias and biomass—sequence relationships with an
685 innovative metabarcoding protocol. *PLoS ONE*, *10*(7), e0130324.

686 Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., & Ravel, J.
687 (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene
688 sequencing on the Illumina MiSeq platform. *Microbiome*, *2*(1), 6.

689 Fernández-Mazuecos, M., Vargas, P. (2010). Ecological rather than geographical isolation
690 dominates Quaternary formation of Mediterranean *Cistus*
691 species. *Molecular Ecology, 19*, 1381–1395.

692 Fierer, N., Hamady, M., Lauber, C. L., & Knight, R. (2008). The influence of sex,
693 handedness, and washing on the diversity of hand surface bacteria. *Proceedings of*
694 *the National Academy of Sciences, 105*(46), 17994-17999.

695 Fleming, T. H., Geiselman, C., & Kress, W. J. (2009). The evolution of bat pollination: a
696 phylogenetic perspective. *Annals of Botany, 104*(6), 1017-1043.

697 Galimberti, A., De Mattia, F., Bruni, I., Scaccabarozzi, D., Sandionigi, A., Barbuto, M., ... &
698 Labra, M. (2014). A DNA barcoding approach to characterize pollen collected by
699 honeybees. *PLoS ONE, 9*, 109363.

700 Garland, M. S., & Davis, A. K. (2002). An examination of monarch butterfly (*Danaus*
701 *plexippus*) autumn migration in coastal Virginia. *The American Midland*
702 *Naturalist, 147*(1), 170-174.

703 Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., …
704 Cornman, R. S. (2016). Critical considerations for the application of environmental
705 DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7,* 1299–
706 1307.

707 Gregg, P. (1993). Pollen as a marker for migration of *Helicoverpa armigera* and *H.*
708 *punctigera* (Lepidoptera: Noctuidae) from western Queensland. *Australian Journal*
709 *of Ecology, 18*, 209–219.

710 Hagler, J. R., & Jackson, C. G. (2001). Methods for marking insects: current techniques and
711 future prospects. *Annual Review of Entomology, 46*(1), 511-543.

712 Hawkins, J., de Vere, N., Griffith, A., Ford, C. R., Allainguillaume, J., Hegarty, M. J., ... &
713 Adams-Groom, B. (2015). Using DNA metabarcoding to identify the floral
714 composition of honey: a new tool for investigating honey bee foraging preferences.
715 *PLoS ONE, 10*(8), e0134735.

716 Hendrix, W. H., & Showers, W. B. (1992). Tracing black cutworm and armyworm
717 (Lepidoptera: Noctuidae) northward migration using *Pithocellobium* and *Calliandra*
718 pollen. *Environmental Entomology, 21*, 1092–1096.

719 Hendrix, W. H., Mueller, T. F., Phillips, J. R., & Davis, O. K. (1987). Pollen as an indicator
720 of long-range movement of *Heliothis zea* (Lepidoptera: Noctuidae). *Environmental*
721 *Entomology*, *16*, 1148–1151.

722 Hobson, K. A. (2008). Applying isotopic methods to tracking animal movements. *Terrestrial*
723 *Ecology, 2*, 45-78.

724 Hobson, K. A., Wassenaar, L. I., & Taylor, O. R. (1999). Stable isotopes (δD and δ13C) are
725 geographic indicators of natal origins of monarch butterflies in eastern North
726 America. *Oecologia, 120*(3), 397-404.

727 Holland, R. A., Wikelski, M., & Wilcove D. S. (2006). How and why do insects migrate?
728 *Science, 313*, 794-796.

729 Howard, E., & Davis, A. K. (2009). The fall migration flyways of monarch butterflies in
730 eastern North America revealed by citizen scientists. *Journal of Insect*
731 *Conservation, 13*(3), 279-286.

732 Hu, G., Lim, K. A., Horvitz, N., Clark, S. J., Reynolds, D. R., Sapir, N., Chapman, J. W.
733 (2016). Mass seasonal bioflows of high-flying insect migrants. *Science, 354*, 1584-
734 1587.

735 IUCN (2012). *IUCN Red List Categories and Criteria: Version 3.1*. Gland, Switzerland.

736 Jones, G. D., & Jones, S. D. (2001). The uses of pollen and its implication for
737 entomology. *Neotropical Entomology*, *30*(3), 314-349.

738 Karaca, M., Ince, A. G., Aydin, A., Elmasulu, S. Y. & Turgut, K. (2015). Microsatellites for
739 genetic and taxonomic research on thyme (*Thymus* L.). *Turkish Journal of Biology,*
740 *39*, 147-159.

741 Keller, A., Danner, N., Grimmer, G., Ankenbrand, M., von der Ohe, K., von der Ohe, W., ...
742 & Steffan-Dewenter, I. (2015). Evaluating multiplexed next-generation sequencing
743 as a method in palynology for mixed pollen samples. *Plant Biology, 17*, 558–566.

744 Kevan, P. G., & Baker, H. G. (1983). Insects as flower visitors and pollinators. *Annual*
745 *Review of Entomology, 28*(1), 407-453.

746 Kissling, D. W., Pattemore, D. E., & Hagen, M. (2014). Challenges and prospects in the
747 telemetry of insects. *Biological Reviews, 89*(3), 511-530.

748 Knight, A., Brower, L. P., & Williams, E. H. (1999). Spring remigration of the monarch
749 butterfly, *Danaus plexippus* (Lepidoptera: Nymphalidae) in north-central Florida:
750 estimating population parameters using mark-recapture. *Biological Journal of the*
751 *Linnean Society, 68*(4), 531-556.

752 Koloren, O., Koloren, Z. & Eker, S. (2016). Molecular phylogeny of *Artemisia* species based
753 on the internal transcribed spacer (ITS) of 18S-26S rDNA in Ordu Province of
754 Turkey. *Biotechnology & Biotechnological Equipment, 30*(5): 929-934.

755 Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., Schloss, P. D. (2013).
756 Development of a dual-index sequencing strategy and curation pipeline for analyzing

757 amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and*
758 *Environmental Microbiology, 79*, 5112–5120.

759 Kraaijeveld, K., de Weger, L. A., Ventayol García, M., Buermans, H., Frank, J., Hiemstra, P.
760 S., den Dunnen, J. T. (2015). Efficient and sensitive identification and quantification
761 of airborne pollen using next-generation DNA sequencing. *Molecular Ecology*
762 *Resources, 15*, 8–16.

763 Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G.
764 (2017). Estimating and mitigating amplification bias in qualitative and quantitative
765 arthropod metabarcoding. *Scientific Reports, 7*(1), 17668.

766 Lingren, P. D., Bryant, V. M., Raulston, J. R., Pendelton, M., Westbrook, J., Jones, G.
767 D. (1993). Adult feeding host range and migratory activities of corn earworm,
768 cabbage looper and celery looper (Lepidoptera: Noctuidae) moths as evidenced by
769 attached pollen. *Journal of Economic Entomology, 86*, 1429–1433.

770 Lingren, P.D., Westbrook, J.K., Bryant, V.M., Raulston, J.R., Esquivel, J.F., Jones,
771 G.D. (1994). Origin of corn earworm (Lepidoptera: Noctuidae) migrants as
772 determined by citrus pollen markers and weather systems. *Environmental*
773 *Entomology, 23*, 562–570.

774 Lowe, W. H., Allendorf, F. W. (2010). What can genetics tell us about population
775 connectivity? *Molecular Ecology, 19*, 3038-3051.

776 Lukhtanov, V. A., Pazhenkova, E. A., & Novikova, A. V. (2016). Mitochondrial chromosome
777 as a marker of animal migratory routes: DNA barcoding revealed Asian (non-
778 African) origin of a tropical migrant butterfly *Junonia orithya* in south
779 Israel. *Comparative cytogenetics*, *10*(4), 671.

780 Lyons, J. I., Pierce, A. A., Barribeau, S. M., Sternberg, E. D., Mongue, A. J., Roode, D., &
781 Jacobus, C. (2012). Lack of genetic differentiation between monarch butterflies with
782 divergent migration destinations. *Molecular Ecology*, *21*(14), 3433-3444.

783 Mateu-Andrés, I., Aguilella, A., Boisset, F., Currás, R., Guara, M., Laguna, E., ... & Pedrola,
784 J., (2013). Geographical patterns of genetic variation in rosemary (*Rosmarinus*
785 *officinalis*) in the Mediterranean basin. *Botanical Journal of the Linnean Society,*
786 *171*, 700–712.

787 Mikkola K. (1971). Pollen analysis as a means of studying the migrations of
788 Lepidoptera. *Annales Entomologici Fennici, 37*, 136–139.

789 Morales, R. (2010). *Thymus* L. In Morales, R., Quintanar, A., Cabezas, F., Pujadas, A.J. &
790 Cirujano, S. (Eds.), *Flora iberica. Vol. XII. Verbenaceae-Labiatae-Callitrichaceae*
791 (pp. 349-409). Madrid: Real Jardín Botánico de Madrid (C.S.I.C.).

792 Nagoshi, R. N., Meagher, R. L. & Hay-Roe, M. (2012). Inferring the annual migration
793 patterns of fall armyworm (Lepidoptera: Noctuidae) in the United States from
794 mitochondrial haplotypes. *Ecology and Evolution, 2*, 1458–1467.

795 Nylin, S., Slove, J., & Janz, N. (2014). Host plant utilization, host range oscillations and
796 diversification in nymphalid butterflies: a phylogenetic investigation. *Evolution*,
797 *68*(1), 105-124.

798 O'Donnell, J. L., Kelly, R. P., Lowell, N. C., & Port, J. A. (2016). Indexed PCR primers
799 induce template-specific bias in large-scale DNA sequencing studies. *PLoS ONE,*
800 *11*(3), e0148698.

801 Ortiz, M. A., Tremetsberger, K., Talavera, S., Stuessy, T., García-Castano, J. L. (2007).
802 Population structure of *Hypochaeris salzmanniana* DC (Asteraceae), an endemic
803 species to the Atlantic coast on both sides of the Strait of Gibraltar, in relation to
804 Quaternary sea level changes. *Molecular Ecology, 16*, 541–552.

805 Osborne, J. L., Loxdale, H. D., & Woiwod, I. P. (2002). Monitoring insect dispersal: methods
806 and approaches. In: Bullock, J. M., Kenward, R. E., Hails, R. S. (Eds.), *Dispersal*
807 *Ecology: the 42nd Symposium of the British Ecological Society* (pp. 24-49). Oxford:
808 Blackwell Science.

809 Piñol, J., Mir, G., Gomez-Polo, P., & Agust, N. (2015). Universal and blocking primer
810 mismatches limit the use of high-throughput DNA sequencing for the quantitative
811 metabarcoding of arthropods. *Molecular Ecology Resources, 15*, 819–830.

812 Podlech, D. & Zarre, S. (2013). *A Taxonomic Revision of the Genus* Astragalus *L.*
813 *(Leguminosae) in the Old World, vols. 1–3*. Wien: Naturhistorisches Museum.

814 Pornon, A., Escaravage, N., Burrus, M., Holota, H., Khimoun, A., Mariette, J., ... & Vidal, M.
815 (2016). Using metabarcoding to reveal and quantify plant-pollinator
816 interactions. *Scientific Reports*, *6*, 27282.

817 Prosser, S. W., & Hebert, P. D. (2017). Rapid identification of the botanical and
818 entomological sources of honey using DNA metabarcoding. *Food Chemistry, 214*,
819 183-191.

820 Richardson, B. A., Page, J. T., Bajgain, P., Sanderson, S. C. & Udall, J.A. (2012). Deep
821 sequencing of amplicons reveals widespread intraspecific hybridization and multiple
822 origins of polyploidy in big sagebrush (*Artemisia tridentata*; Asteraceae). *American*
823 *Journal of Botany, 99*(12), 1962-1975.

824 Richardson, R.T., Lin, C.-H., Quijia, J.O., Riusech, N.S., Goodell, K., Johnson, R.M. (2015a).
825 Rank-based characterization of pollen assemblages collected by honey bees using a
826 multi-locus metabarcoding approach. *Applications in Plant Sciences, 3*(11),
827 1500043.

828 Richardson, R.T., Lin, C.-H., Sponsler, D.B., Quijia, J.O., Goodell, K., Johnson, R.M.
829 (2015b) Application of ITS2 metabarcoding to determine the provenance of pollen
830 collected by honey bees in an agroecosystem. *Applications in Plant Sciences, 3*,
831 1400066.

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

834 Rolph, G., Stein, A., & Stunder, B. (2017). Real-time environmental applications and display
835 system: Ready. *Environmental Modelling & Software*, *95*, 210-228.

836 Sanderson, M. J. & Wojciechowski, M. F. (1996). Diversification rates in a temperate legume
837 clade: Are there "so many species" of *Astragalus* (Fabaceae)? *American Journal of*
838 *Botany, 83*, 1488–1502.

839 Shields, O. (1974). Toward a theory of butterfly migration. *Journal of Research on the*
840 *Lepidoptera*, *13*(4), 217-238.

841 Shields, O. (1992). World distribution of the *Vanessa cardui* group (Nymphalidae). *Journal*
842 *of the Lepidopterists' Society, 46*, 235–238.

843 Sickel, W., Ankenbrand, M. J., Grimmer, G., Holzschuh, A., Härtel, S., Lanzen, J., & Keller,
844 A. (2015). Increased efficiency in identifying mixed pollen samples by meta-
845 barcoding with a dual-indexing approach. *BMC ecology*, *15*(1), 20.

846 Simel, E. J., Saidak, L. R., & Tuskan, G. A. (1997). Method of extracting genomic DNA from
847 non-germinated gymnosperm and angiosperm pollen. *Bio Techniques, 22*, 390–392,
848 394.

849 Sipos, R., Székely, A. J., Palatinszky, M., Révész, S., Márialigeti, K., & Nikolausz, M.
850 (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on
851 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiology*
852 *Ecology*, *60*(2), 341-350.

853 Soltis, P. S., & Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual*
854 *Review of Plant Biology, 60*, 561–588.

855 Stefanescu, C., Alarcón, M., & Àvila, A. (2007). Migration of the painted lady butterfly,
856 *Vanessa cardui*, to north-eastern Spain is aided by African wind currents. *Journal*
857 *of Animal Ecology, 76*(5), 888-898.

858 Stefanescu, C., Alarcón, M., Izquierdo, R., Páramo, F., & Ávila, A. (2011). Moroccan source
859 areas of the painted lady butterfly *Vanessa cardui* (Nymphalidae: Nymphalinae)
860 migrating into Europe in spring. *Journal of the Lepidopterists' Society, 65*, 15-26.

861 Stefanescu, C., Páramo, F., Åkesson, S., Alarcón, M., Ávila, A., Brereton, T., ... & Hill, J. K.
862 (2013). Multi-generational long-distance migration of insects: studying the painted
863 lady butterfly in the Western Palaearctic. *Ecography, 36*(4), 474-486.

864 Stefanescu, C., Soto, D. X., Talavera, G., Vila, R., & Hobson, K. A. (2016). Long-distance
865 autumn migration across the Sahara by painted lady butterflies: exploiting resource
866 pulses in the tropical savannah. *Biology Letters, 12*(10), 20160561.

867 Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J., Cohen, M. D., & Ngan, F. (2015).
868 NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bulletin*
869 *of the American Meteorological Society*, *96*(12), 2059-2077.

870 Talavera, G., & Vila, R. (2016). Discovery of mass migration and breeding of the painted
871 lady butterfly *Vanessa cardui* in the Sub-Sahara: the Europe–Africa migration
872 revisited. *Biological Journal of the Linnean Society*, *120*(2), 274-285.

873 Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards
874 next-generation biodiversity assessment using DNA metabarcoding. *Molecular*
875 *Ecology, 21*(8), 2045-2050.

876 Valentini, A., Miquel, C., & Taberlet, P. (2010). DNA barcoding for honey
877 biodiversity. *Diversity, 2*(4), 610-617.

878 Vinje, H., Liland, K. H., Almøy, T., & Snipen, L. (2015). Comparing K-mer based methods
879 for improved classification of 16S sequences. *BMC bioinformatics*, *16*(1), 205.

880 Westbrook, J. K., Wolf, W. W., Lingren, P. D., Raulston, J. R., Lopez Jr, J. D., Matis, J. H., ...
881 & Schleider, P.G. (1997). Early-season migratory flights of corn earworm
882 (Lepidoptera: Noctuidae). *Environmental Entomology*, *26*, 12–20.

883 Westbrook, J. K., & Eyster, R. S. (2017). Doppler weather radar detects emigratory flights of
884 noctuids during a major pest outbreak. *Remote Sensing Applications: Society and*
885 *Environment, 8*, 64-70.

886 White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). Amplification and direct sequencing of
887 fungal ribosomal RNA genes for phylogenetics. In Innis, M. A., Gelfand, D. H.,
888 Sninsky, J. J., White, T. J. (Eds.), *PCR protocols: a guide to methods and*
889 *applications* (pp. 315–322). New York: Academic Press.

890 Wikelski, M., Moskowitz, D., Adelman, J. S., Cochran, J., Wilcove, D. S., & May, M. L.
891 (2006). Simple rules guide dragonfly migration. *Biology letters*, *2*(3), 325-329.

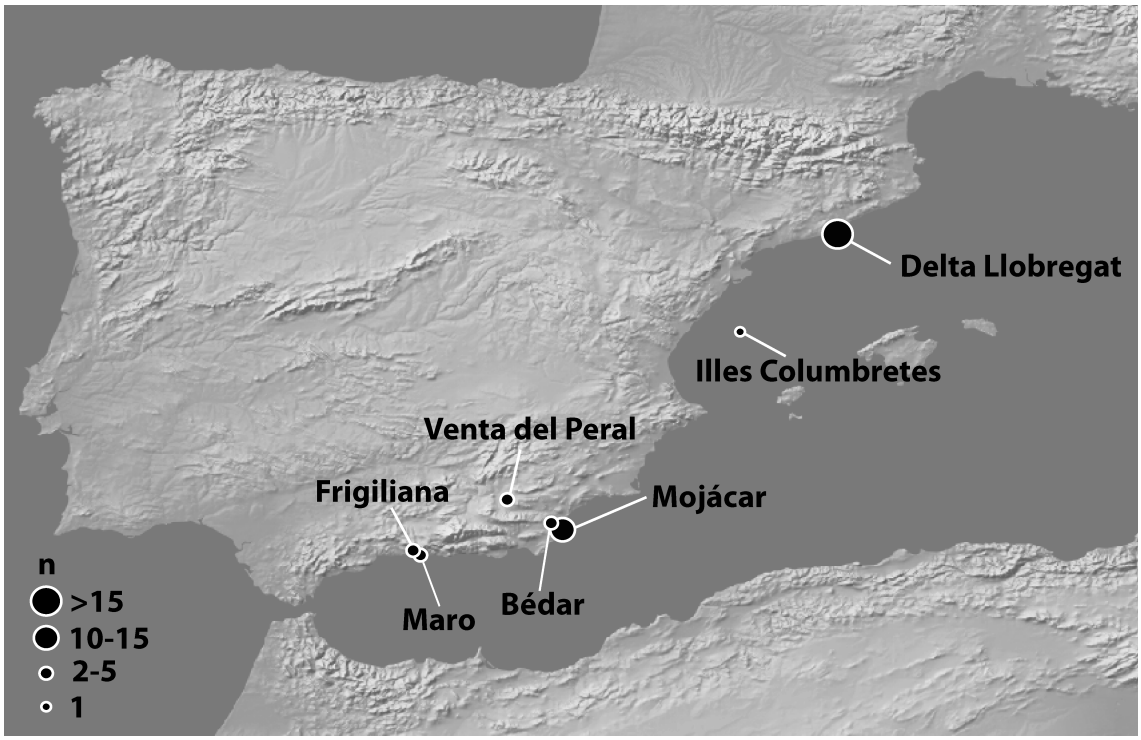
892 Wong, W. H., Tay, Y. C., Puniamoorthy, J., Balke, M., Cranston, P. S., & Meier, R. (2014).
893 'Direct PCR'optimization yields a rapid, cost-effective, nondestructive and efficient
894 method for obtaining DNA barcodes without DNA extraction. *Molecular Ecology*
895 *Resources*, *14*(6), 1271-1280.

896 Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012).
897 Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment
898 and biomonitoring. *Methods in Ecology and Evolution, 3*(4), 613-623.

899 Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2013). PEAR: a fast and accurate Illumina
900 Paired-End reAd mergeR. *Bioinformatics, 30*(5), 614-620.

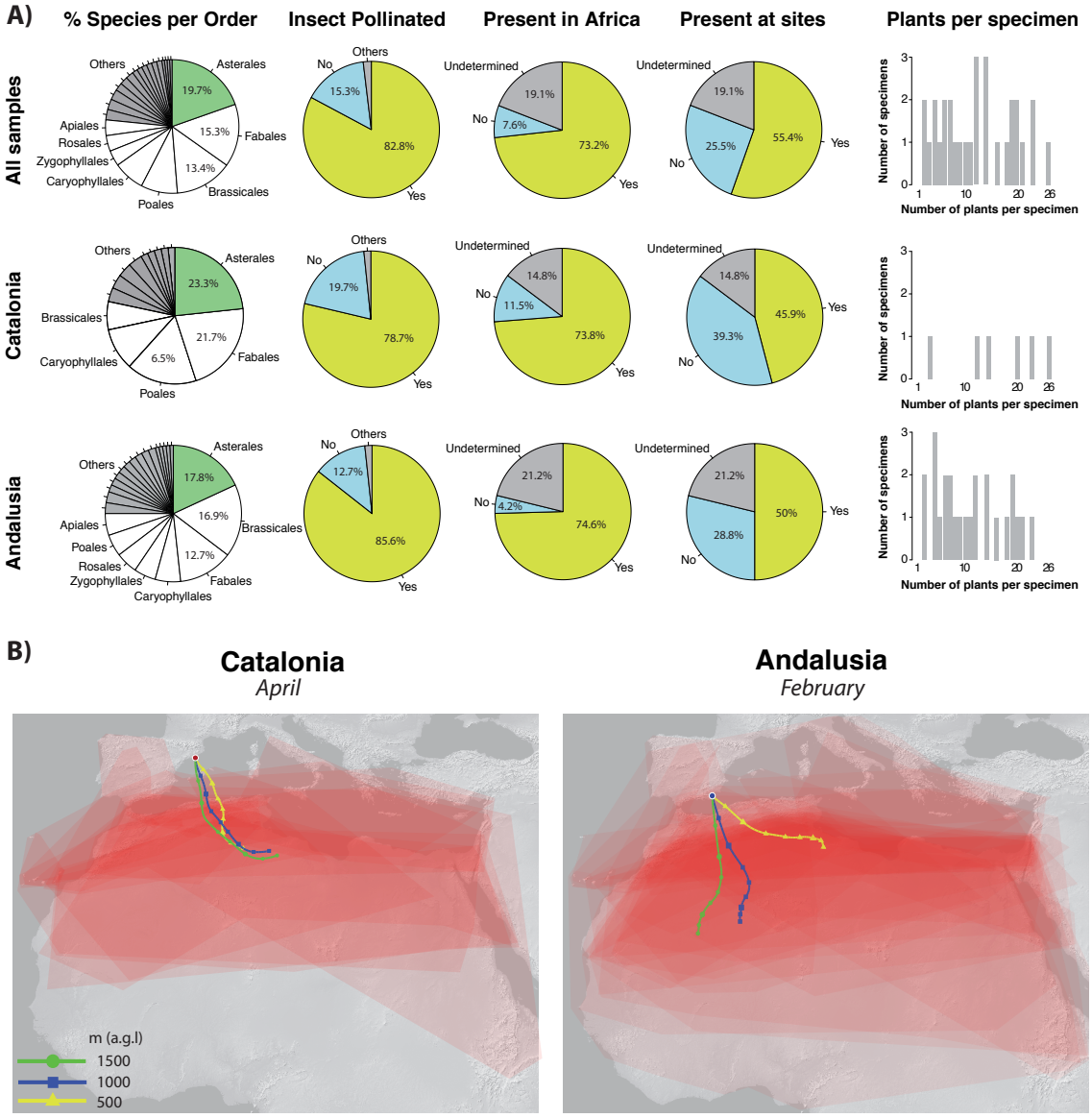
901

902 # Tables and figures



903

904 Fig. 1. Map of the sampled specimens. The size of the points indicates the number of samples
905 collected in each locality.



906

907 Fig. 2. Detected plant species that are informative of migration: a) Pies showing classified
908 percentages of plants per order, insect pollinated and presence in Africa and at sampling sites.
909 Barplots show number of informative plants per specimen. Pies and barplots are shown for all
910 samples together and for two migratory waves (Catalonia and Andalusia) independently. b)
911 Additive extent of occurrences for informative plants detected in Catalonia and Andalusia
912 migrants. Higher red intensity indicates higher probability of plant overlap). 48h backward
913 wind trajectories at three altitudes (500m, 1000m and 1500 m agl) are shown (colour lines)
914 for the specific dates of observed peaks of migration (April 27th – Catalonia, February 21st –
915 Andalusia).

916 Table 1. Sequences of the primers used: ITS2-S2F and ITS2-4R primers were used in the first
917 PCR reaction; PCR_F_D50x and PCR_R_D7xx are the indexing primers used in the second
918 PCR reaction; XXX XXX XX are 8 nt-long index sequences.

919

920 ITS2-S2F: ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN NNN ATG

921 CGA TAC TTG GTG TGA AT

922 ITS2-4R: GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TNN NNN NTC

923 CTC CGC TTA TTG ATA TGC.

924 PCR_F_D50x: AAT GAT ACG GCG ACC ACC GAG ATC TAC ACX XXX XXX XAC

925 ACT CTT TCC CTA CAC GAC GC

926 PCR_R_D7xx: CAA GCA GAA GAC GGC ATA CGA GAT XXX XXX XXG TGA CTG

927 GAG TTC AGA CGT GTG C