JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12493

Research Article

Different filtering strategies of genotyping-by-sequencing data provide complementary resolutions of species boundaries and relationships in a clade of sexually deceptive orchids

Salvatore Cozzolino[1], Giovanni Scopece[1]*, Luca Roma[1], and Philipp M. Schlüter[2]

[1]Department of Biology, University Federico II of Naples, Complesso Universitario Monte Sant'Angelo, via Cinthia, Naples I-80126, Italy
[2]Institute of Botany, University of Hohenheim, Garbenstraße 30, Stuttgart D-70599, Germany
*Author for correspondence. Email: giovanni.scopece@unina.it
Received 8 October 2018; Accepted 5 March 2019; Article first published online 15 April 2019

Abstract Ongoing hybridization and retained ancestral polymorphism in rapidly radiating lineages could mask recent cladogenetic events. This presents a challenge for the application of molecular phylogenetic methods to resolve differences between closely related taxa. We reanalyzed published genotyping-by-sequencing (GBS) data to infer the phylogeny of four species within the *Ophrys sphegodes* complex, a recently radiated clade of orchids. We used different data filtering approaches to detect different signals contained in the dataset generated by GBS and estimated their effects on maximum likelihood trees, global F_{ST} and bootstrap support values. We obtained a maximum likelihood tree with high bootstrap support, separating the species by using a large dataset based on loci shared by at least 30% of accessions. Bootstrap and F_{ST} values progressively decreased when filtering for loci shared by a higher number of accessions. However, when filtering more stringently to retain homozygous and organellar loci, we identified two main clades. These clades group individuals independently from their a priori species assignment, but were associated with two organellar haplotype clusters. We infer that a less stringent filtering preferentially selects for rapidly evolving lineage-specific loci, which might better delimit lineages. In contrast, when using homozygous/organellar DNA loci the signature of a putative hybridization event in the lineage prevails over the most recent phylogenetic signal. These results show that using differing filtering strategies on GBS data could dissect the organellar and nuclear DNA phylogenetic signal and yield novel insights into relationships between closely related species.

Key words: adaptive radiation, F_{ST}, lineage divergence, ML tree, next-generation sequencing, *Ophrys*, phylogenetics, plastid and mitochondrial haplotype, speciation genomics.

1 Introduction

Understanding the evolutionary relationships in closely related, recently diverged lineages often presents a methodological challenge (Maddison & Wayne, 1997). Rapidly diverging taxa highlight the limit of the phylogenetic application of molecular markers as these lineages can be at the interface between incipient species and divergent ecotypes (Feder et al., 2012). Plastid DNA (cpDNA) has been widely applied in plant phylogenetic studies and lineage delimitation thanks to the ease of amplification and sequencing that come with its high copy number (Gielly & Taberlet, 1994). Plastid DNA markers are predominantly uniparentally inherited (including in orchids, Cafasso et al., 2004). Effective population size for such organellar markers is smaller than that of nuclear markers, thereby leading to greater genetic drift and resulting in faster coalescence times than diploid nuclear DNA (Petit et al., 2005; Hernández-León et al., 2013). However, the low evolutionary rate and the haploid nature of cpDNA severely limit its application in closely related species, particularly when introgression (sometimes leading to plastid capture) and incomplete lineage sorting are suspected (Sang et al., 1997). The use of diploid nuclear gene data is often thought to overcome the shortcomings of organellar gene genealogies, as nuclear genes have been reported to evolve up to five times faster (Wolfe et al., 1989; Ossowski et al., 2010; Schlüter et al., 2007). Nevertheless, disadvantages in the use of nuclear genes stem from their frequent occurrence in gene families (paralogy), recombination, and a general lack of available primers for non-model organisms (e.g., Doyle, 1997; Posada & Crandall, 2002) although the latter problem has been alleviated to a certain degree by the arrival of next generation sequencing (NGS) technology. The analysis of recently diverged taxa is further complicated by the frequent existence of retained ancestral polymorphism, when ancestral allelic

variants are maintained in both descendant species following neutral expectations. However, coalescent theory predicts that the noise produced by incomplete lineage sorting can be reduced by sampling multiple genes per species (Edwards & Beerli, 2000). This results in genealogies that might differ in their topologies, because unlinked nuclear genes are differently affected by introgression and intragenic recombination (Degnan & Rosenberg, 2009).

The use of large multilocus datasets, such as those consisting of sequence data from multiple, unrelated genomic regions, can improve phylogenetic inferences by accounting for the stochasticity in the coalescent process (Knowles & Maddison, 2002; Knowles, 2009; Carstens et al., 2013). Indeed, analyzing multiple genes and alleles per species increases the probability in approximating the underlying species tree supported by the majority of the data (Small et al., 2004). This could help overcome the typical limitations of using single/few genes to assess phylogenetic relationships and the demographic history of a species (Edwards & Beerli, 2000; Edwards, 2009; Hipp et al., 2014). Recent advances of NGS tools and multilocus analyses have been applied for successful reconstruction of phylogenetic relationships and delimitation of boundaries between closely related species within species complexes. Amongst the more common genomic methods, reduced-representation methods (reviewed in Davey & Blaxter, 2010), such as restriction-site associated DNA sequencing (RADseq; Miller et al., 2007; Baird et al., 2008; Rowe et al., 2011), or genotyping-by-sequencing (GBS; Elshire et al., 2011), identify sequence fragments of DNA that flank the recognition sites of restriction enzymes in an individual's genome (Miller et al., 2007; Baird et al., 2008) by using high-throughput sequencing technologies. This selection of DNA fragments, scattered throughout the individual genome, allows orthologous sequences to be targeted across multiple samples to identify a large number of genetic markers. These methods provide a useful tool, particularly for surveying the genome of non-model organisms (Ellegren, 2014).

Most applications of genomic reduced-representation methods have been within species (e.g., Lewis et al., 2007; Emerson et al., 2010; Hohenlohe et al., 2010; Bruneaux et al., 2013; Wang et al., 2013) or among closely related species (e.g., Stölting et al., 2013; Wagner et al., 2013). This is because the primary challenge in applying these methods to reconstructing interspecific phylogenies lies in confidently identifying and assembling orthologous loci amongst the relatively short (i.e., usually 100 to 200 bp), usually non-coding sequence fragments produced with the NGS technologies (Rubin et al., 2012). This problem stems from the fact that: (i) the number of restriction sites that are conserved among taxa is expected to decrease with increased time since divergence; (ii) the ability to compare orthologous loci is expected to decrease with phylogenetic distance due to the progressive accumulation of mutations. These caveats indicate that such genotyping data are expected to be particularly valuable for recently diverged and closely related clades (Wagner et al., 2013). Nevertheless, both simulation (Cariou et al., 2013) and experimental studies have shown that genomic reduced-representation methods can be successfully applied at the level of subfamilies or even families (Wang et al., 2017).

The Mediterranean orchid genus *Ophrys* L. has not only attracted the interest of taxonomists since Darwin (e.g., Darwin, 1862; Kullenberg, 1961), but it has also become a useful system to study speciation and reproductive isolation (Scopece et al., 2007; Xu et al., 2011). Nevertheless, it also represents evidence of fast evolving clades very recalcitrant to most methods for phylogenetic analyses (Breitkopf et al., 2015). The genus can be merged or split into a large number of lineages that are at least locally and temporally reproductively isolated enough to develop some morphological differences (Bateman et al., 2011; Vereecken et al., 2011). As post-zygotic barriers are effectively absent within closely related groups, reproductive isolation in sympatry is almost exclusively based on floral isolation through specific male pollinators that are lured by the floral scent, a copy of the sexual pheromone of conspecific females, to repeatedly pseudocopulate on flowers of only a single *Ophrys* species, leading to cross-pollination (Kullenberg, 1961). An accelerated diversification rate in terminal clades has been explained by the exploitation of novel, species-rich, and diverse groups of pollinators resulting in recent and rapid radiation that is characterized by dynamic speciation processes due to repeated pollinator shifts (Breitkopf et al., 2015). Previous molecular studies in *Ophrys* (Devey et al., 2008; Breitkopf et al., 2015) support at least 10 main lineages that presumably give rise to species flocks by the adoption of pollinators from large diversified bee genera, such as *Eucera* and *Andrena*. Among these, the *Ophrys sphegodes* complex represents one of the most species-rich groups in *Ophrys*, diversified only in the last 1 million years by exploiting different *Andrena* and, to a lesser degree, *Colletes* bees as pollinators (Breitkopf et al., 2015; Delforge, 2016). Despite intensive past research, phylogenetic patterns and species diversity within this complex remain highly contentious. Both plastid and nuclear phylogenies – including the use of a dataset of multiple nuclear genes – failed to identify species relationships or to delimit species within the *O. sphegodes* complex (Soliva et al., 2001; Bateman et al., 2011; Breitkopf et al., 2015). Thus, this complex represents an ideal group for testing the application of NGS-based multilocus analyses for inference of relationships and species delimitation. The RADseq method has very recently been applied to the phylogeny of *Ophrys* at the level of the ~10 main lineages, confirming the suitability of NGS methods and approaches for phylogenetic purposes in taxonomically complex groups (Bateman et al., 2018). However, only one attempt to use multilocus NGS approaches has previously been made at the within-species-complex level at the transition zone between species and population levels. Specifically, Sedeek et al. (2014) used GBS data to present an UPGMA tree based on overall pairwise genotypic distances between individuals of the *O. sphegodes* complex. In this analysis, none of the internal nodes separating the species received any bootstrap support. Similarly, a STRUCTURE analysis, run on the same dataset, indicated a large proportion of shared polymorphism and found $K = 6$ ancestry clusters as the most probable inference, at least in the applied dataset (88 individuals and 1233 loci with one single nucleotide polymorphism (SNP) analyzed per locus). Here, we reanalyzed genome-wide sequence/SNP data collected by Sedeek et al. (2014) by using different criteria of locus

selection in order to: (i) delimit the species boundaries within a group of four sympatric southern Italian species of the O. *sphegodes* complex; (ii) infer a well-supported pattern of relationships/descendance for these species; and (iii) identify the signature of past events affecting lineage divergence in this group.

2 Material and Methods

2.1 Study system and GBS data source

Here, we investigated all four members of the *Ophrys sphegodes* Mill. species complex cogrowing in the National Park of Gargano (Apulia, Italy), that is, *O. exaltata* Ten. subsp. *archipelagi* (Gölz & H.R. Reinhard) Del Prete, *O. garganica* E. Nelson ex O. Danesch & E. Danesch, *O. incubacea* Bianca ex Tod. and *O. sphegodes*. These four species are pollinated through sexual deception by three different *Andrena* (*A. pilipes*, *A. morio*, and *A. nigroaenea*, for *O. garganica*, *O. incubacea*, and *O. sphegodes*, respectively) and a *Colletes* species (*Colletes cunicularius* for *O. exaltata*) (Paulus & Gack, 1990). The four species coflower in spring (from March to April) and occur in close proximity to each other in the study area (Xu et al., 2011; Sedeek et al., 2014).

We proceeded to reanalyze trimmed and demultiplexed GBS Illumina reads generated by Sedeek et al. (2014). From the full dataset of Sedeek et al. (2014), encompassing 127 accessions, we filtered the data according to the number of reads. To maximize the number of reads per accession, we used a more conservative approach than Sedeek et al. (2014) by selecting samples with at least 800,000 reads (a total of 54 individuals) roughly corresponding to the median value of reads per accession in the original dataset. However, we also compared the results to datasets including: (i) accessions with at least 500,000 or 300,000 reads; and (ii) accessions with a number of reads ranging from 500,000 to 2,000,000.

2.2 Plastid and mitochondrial haplotype network analysis

Plastid reads were identified by mapping forward and reverse GBS reads for each individual against the *Ophrys iricolor* Desf. and *O. sphegodes* plastid genomes (Roma et al., 2018) using BWA MEM version 0.7 software (Li, 2013) with the option -M that marks shorter split hits as secondary (as required by GATK software). Variant calling analysis was carried out using the Genome Analysis ToolKit version 3.5 according to the GATK Best Practices workflow (McKenna et al., 2010). After the SNP and indel recalibration, a BAM format file was generated for each sample. Finally, a VCF file was generated with GATK package HaplotypeCaller with the option -ploidy 1. Plastid haplotype network analysis was undertaken using PopART version 1.7 software (Leigh & Bryant, 2015) by only using informative SNPs.

As no *Ophrys* mitochondrial genome is available, the mitochondrial reads were identified by blasting (BLAST 2.6.0) against the Organelle Genome Resources database (https://www.ncbi.nlm.nih.gov/genome/organelle/). We used informative mitochondrial SNPs shared by all 54 individuals to reconstruct a mitochondrial haplotype network using PopART version 1.7 software.

2.3 Genotyping-by-sequencing data assembly

In contrast to Sedeek et al. (2014), we used the software pipeline PyRAD version 1.2 (Eaton, 2014) to process the GBS reads instead of Stacks (Catchen et al., 2013). We chose this approach because it allows the construction of supermatrices (i.e., sequence data from concatenated reads) with different minimum percentages of shared loci. Nucleotide base calls with a quality score below 20 were replaced with N, and sequences with more than five Ns were discarded from edited FASTA files created by PyRAD. Clustering was carried out in VSEARCH version 1.0.16 (Edgar, 2010), using the forward reads faster version without reverse complement clustering because of the low overlap between forward or reverse reads.

Only SNPs were used and their distribution per cluster checked in order to avoid markers with more SNPs potentially biasing the inference when treating each locus as one independent marker. Clusters with coverage of less than five reads per locus and more than five heterozygous sites were discarded. Consensus sequences were then clustered across accessions at 88% similarity (the PyRAD default setting) and aligned using MUSCLE version 3.8 (Edgar, 2004). We then applied a supermatrix approach in which all selected clusters were concatenated into a single alignment using PyRAD version 1.2. Missing data symbols (Ns) were inserted into the data matrix for loci without data for a given individual (Wagner et al., 2013).

2.4 Phylogenetic inference

To infer phylogenies from the GBS data, we built different supermatrices by selecting loci shared by at least 10%, 30%, 50%, 70%, and 90% of accessions and reconstructed maximum likelihood (ML) trees. Maximum likelihood analyses were carried out in RAXML version 8.2.10 software using the GTRGAMMA nucleotide substitution model (an inclusive model accounting for a large proportion of missing data; see Roure et al., 2013) and with bootstrap support estimated by 1,000 replicate searches.

Phylogenetic trees were drawn using FigTree version 1.4.3 software. To test for the effect of heterozygosity, we also reconstructed phylogenetic trees using RRHS software version 1.0.0.2 (Lischer et al., 2014) on the supermatrix with loci shared by at least 30% of accessions.

For each ML tree, we calculated a mean bootstrap support value by averaging the bootstrap values over species-level nodes. Following Sedeek et al. (2014), for each locus, we calculated "global" F_{ST} among all four species using BayeScan 2.1 (Foll & Gaggiotti, 2008), treating orchid species as four different populations. Then we plotted average bootstrap support values and F_{ST} values averaged over all loci against the percentage of shared loci among accessions.

To detect old phylogenetic/phylogeographic signals, from the supermatrix built with loci shared by at least 70% of accessions we also selectively filtered for loci with no heterozygous state (homozygous/organellar loci). In particular, to clearly identify the organellar signal in these loci, plastid reads were identified by using the BAM file previously generated for the plastid haplotype search. Unmapped reads were retained by using SAMtools version 1.5 (Li et al., 2009) with the parameters view and -f4 and then converted in FastQ format using SAMtools Bam2fq. In contrast,

mitochondrial reads were identified using available Perl scripts (https://github.com/btmartin721/file_converters/blob/master/loci2fasta.py; https://github.com/nylander/catfasta2phyml) with minor modifications.

On two supermatrices, that is, the one with 30% of shared loci and the one filtered for retaining homozygous/organellar loci (shared by at least 70% of accessions), we also undertook analyses of population structure. First, pairwise distances between individuals based on unphased diploid SNP calls were calculated as described in Sedeek et al. (2014) by using a custom Delphi program using the biOP library (https://sourceforge.net/p/biop/). The advantage of this approach is that it avoids global threshold-based exclusion of loci from the dataset and utilizes the maximum number of data points available for any given pairwise comparison. Distance matrices were used for building neighbor joining (NJ) trees in FAMD 1.31 (Schlüter & Harris, 2006). Second, we used the Bayesian clustering approach as implemented in STRUCTURE version 2.3.4 (Pritchard et al., 2000) by using one SNP per read. Following the method described in Evanno et al. (2005), we tested K from 1 to 7 with a burn-in of 100,000 steps followed by 100,000 Markov chain Monte Carlo iterations with three replicates to confirm stabilization of the summary statistics.

3 Results

Illumina sequencing produced approximately 280 million reads with a quality score of at least 30. The number of reads recovered for each accession and the proportion of missing data are summarized in appendix S4 of Sedeek et al. (2014) and in Table 1. By selecting plastid loci from the GBS data shared by all individuals we identified six distinct haplotypes belonging to two phyletic clusters (A–D and E–F) according to the haplotype network analysis (Fig. 1). The two clusters are separated by two mutational steps. Within each cluster, the haplotypes were separated by single mutation steps (Fig. 1). By selecting four shared mitochondrial SNPs we identified six distinct haplotypes in the network analysis (Fig. S1).

Single nucleotide polymorphism distribution per locus showed that the majority of loci (76%) included a maximum of three SNPs (Fig. S2). The ML phylogenetic analysis with the supermatrices with loci shared among at least 10% and 30% of individuals (Table 1) shows species-specific clades. All *Ophrys* species are monophyletic, with the exception of *O. garganica*, which is paraphyletic. However, only in the supermatrix with loci shared among at least 30% do they all have bootstrap support above 70% (Fig. 2). Indeed, the tree built with the supermatrix with loci shared at least among 10% has higher bootstrap support for terminal clades, but the placement and monophyly of *O. garganica* was weakly supported (ML tree not shown). Analysis with RRHS software, which accounts for heterozygosity, yielded results consistent with these ML results (Fig. S3).

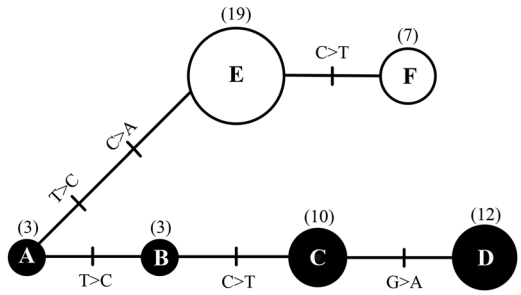

Fig. 1. Statistical parsimony haplotype network of four species within the *Ophrys sphegodes* complex based on plastid loci filtered from the genotyping-by-sequencing data and shared by all accessions. Circle size is proportional to haplotype frequency. Black and white circles indicate the two plastid haplotype lineages identified in the network analysis. Values in parentheses indicate the number of accessions.

We observed similar phylogenetic relationships but a decay in the bootstrap support when using datasets with fewer reads (at least 500,000 or 300,000 reads) per sample (76 and 93 individuals, respectively) and when using a dataset with a number of reads per individual ranging from 500,000 to 2,000,000 (60 individuals) (ML trees not shown). Thus, all following analyses were undertaken with the dataset including 54 accessions with at least 800,000 reads.

In a small supermatrix (with loci shared by at least 50% of accessions), resolution of the four species clades slightly decreases, as does bootstrap support, for the placement of *O. garganica* as sister species of remaining taxa (Table 1; Fig. S4). By progressively reducing the number of loci (shared by at least 70% and 90% of accessions) we observe a further progressive decay of bootstrap support across clades in the trees (Table 1, Fig. 3A). In these last analyses, individuals of the same species do not form monophyletic clades (Figs. S5, S6). Like bootstrap support, F_{ST} values also decrease progressively as the number of loci increases and the dimensions of the supermatrices are reduced (Table 1; Fig. 3B).

Table 1 Number of informative single nucleotide polymorphisms (SNPs), average bootstrap value, global F_{ST} value, and number of plastid and mitochondrial SNPs in the different supermatrices built by using different percentages of shared loci in four species of orchid within the *Ophrys sphegodes* complex

Minimum percentage of shared loci	Informative SNPs	Average bootstrap value	Global F_{ST} value	Plastid SNPs	Mitochondrial SNPs
10	123,080	74.50	0.220	93	132
30	59,435	90.00	0.152	35	53
50	31,272	68.00	0.110	21	35
70	16,710 (253†)	53.00	0.087	6	20
90	6,210	16.18	0.076	3	8

†Number of informative loci after selectively filtering for homozygous/organellar loci.

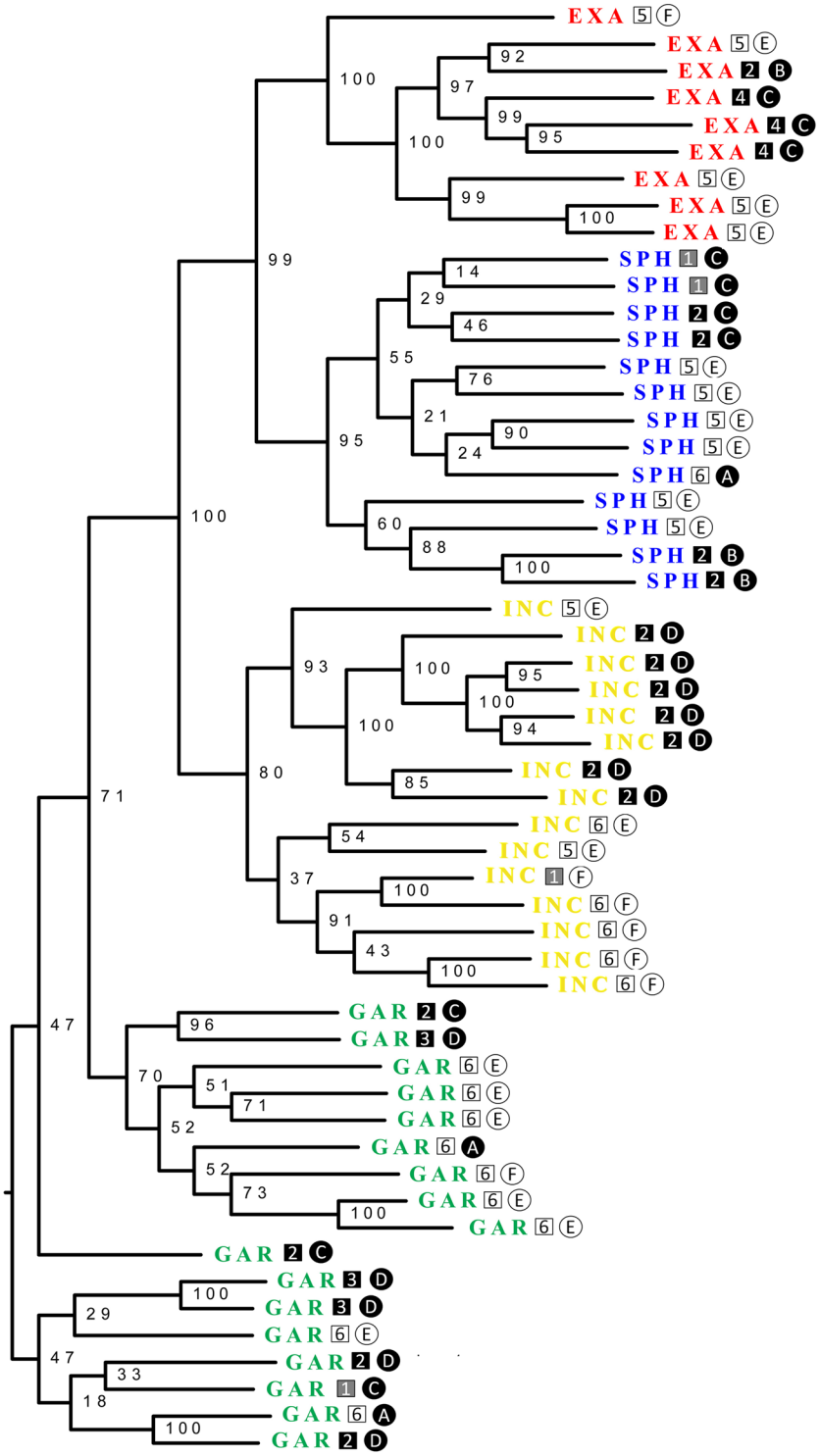

Fig. 2. RaxML tree of four species within the *Ophrys sphegodes* complex obtained by using the supermatrix with loci shared by at least 30% of accessions. Support values are derived from 1,000 bootstrap replicates. Letters in circles represent plastid haplotypes; numbers in squares represent mitochondrial haplotypes. EXA, *Ophrys exaltata*; GAR, *O. garganica*; INC, *O. incubacea*; SPH, *O. sphegodes*.

The phylogenetic analysis using the small supermatrix (loci shared by at least 70% of accessions) selectively filtered for homozygous/organellar loci (185,019 base pairs in width, 253 informative SNP) again produced a tree topology with high bootstrap support, but only for the main basal nodes (Fig. 4). With this supermatrix, we identified main lineages (bootstrap support above 90%) that group accessions independently from their species assignment, but according to the plastid and mitochondrial clusters identified in the haplotype network analyses instead (Figs. 1, S1). When using this reduced dataset (i.e., homozygous/organellar loci shared by at least 70% of accessions), the NJ trees based upon pairwise

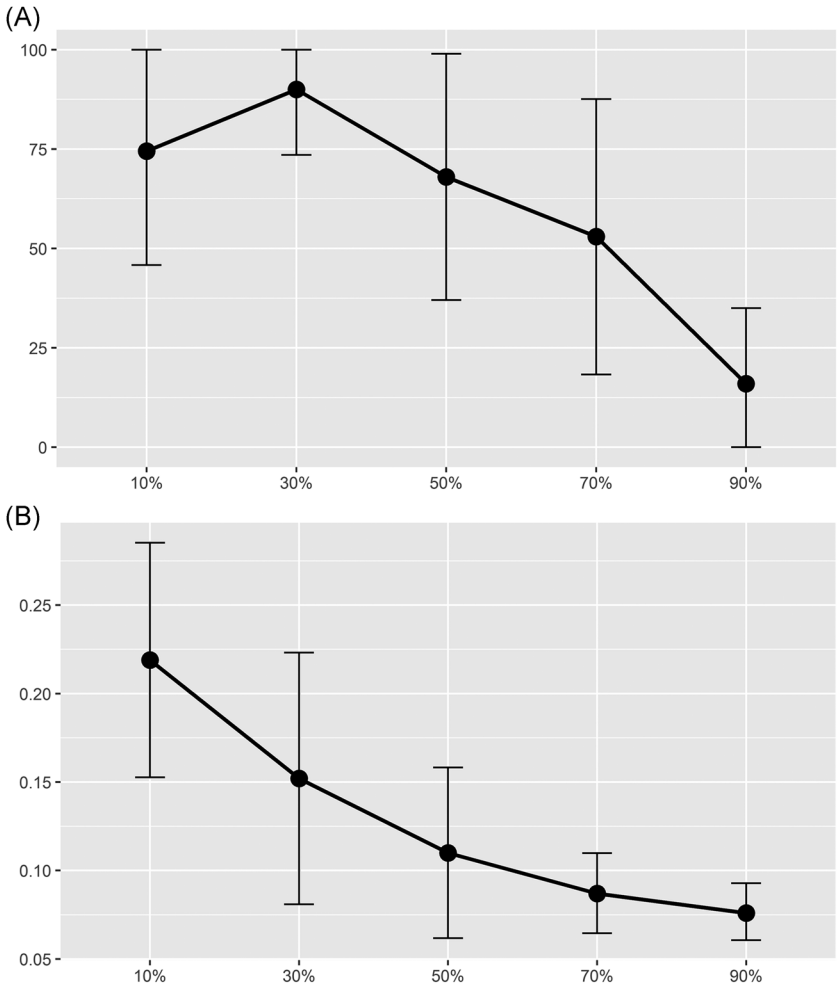

Fig. 3. A, Average bootstrap support values in a RaxML tree of four species within the *Ophrys sphegodes* complex obtained by using the supermatrices with loci shared by at least 10%, 30%, 50%, 70%, and 90% of accessions. **B,** Global F_{ST} values among the four *Ophrys* species by using the supermatrices with loci shared by at least 10%, 30%, 50%, 70%, and 90% of accessions.

SNP distances identified two main lineages corresponding to the two haplotype clusters (cf. Fig. 1) (Fig. S7A). Accordingly, Bayesian analysis on this dataset identified K = 2 (Fig. S7A) as the most probable.

Instead, when using the large dataset including loci shared by at least 30% of accessions, the resulting NJ tree confirmed the ML tree topology: a clear delimitation of the four *Ophrys* species is evident (Fig. S7B). However, Bayesian analysis on this dataset identified K = 3 as the most probable K, mostly corresponding to species assignment, but with *O. incubacea* and *O. sphegodes* combined (Fig. S7B). After we removed homozygous/organellar loci from this dataset, the resulting Bayesian analysis identified K = 5 as the most probable K, corresponding to the four species assignment recognized by the corresponding ML and NJ trees, but with *O. incubacea* divided in two groups (Fig. S7C).

4 Discussion

Despite the great deal of attention the phylogeny of the Mediterranean orchid genus *Ophrys* has received over the past 20 years, relationships among closely related species are still unresolved when using traditional phylogenetic nuclear and plastid markers (Bateman et al., 2011; Soliva et al., 2001; Devey et al., 2008; Breitkopf et al., 2015). Here we show that multilocus GBS data, when properly filtered, can provide a useful tool to assess the degree of genetic separateness/togetherness and patterns of relationships among four incipient species of the *O. sphegodes* complex that are treated as separate species, subspecies, varieties, or populations depending on contrasting taxonomic treatments (Bateman et al., 2011; Vereecken et al., 2011; Delforge, 2016). Previous studies used plastid and/or nuclear genes to infer phylogenetic relationships in *Ophrys* and included in their analysis multiple accessions from the *O. sphegodes* complex (Soliva et al., 2001; Devey et al., 2008; Breitkopf et al., 2015). Results of these studies supported the monophyly of the *O. sphegodes* complex, but patterns of relationships within the species complex were largely unresolved. The application of high-throughput sequencing generating a large multilocus dataset enabled us to resolve fine-scale genetic divergence among members of the *O. sphegodes* complex. Individuals of the same species (at least based on morphologic traits and scent emission) form well-supported clades, suggesting that insufficient informative characters in previous studies

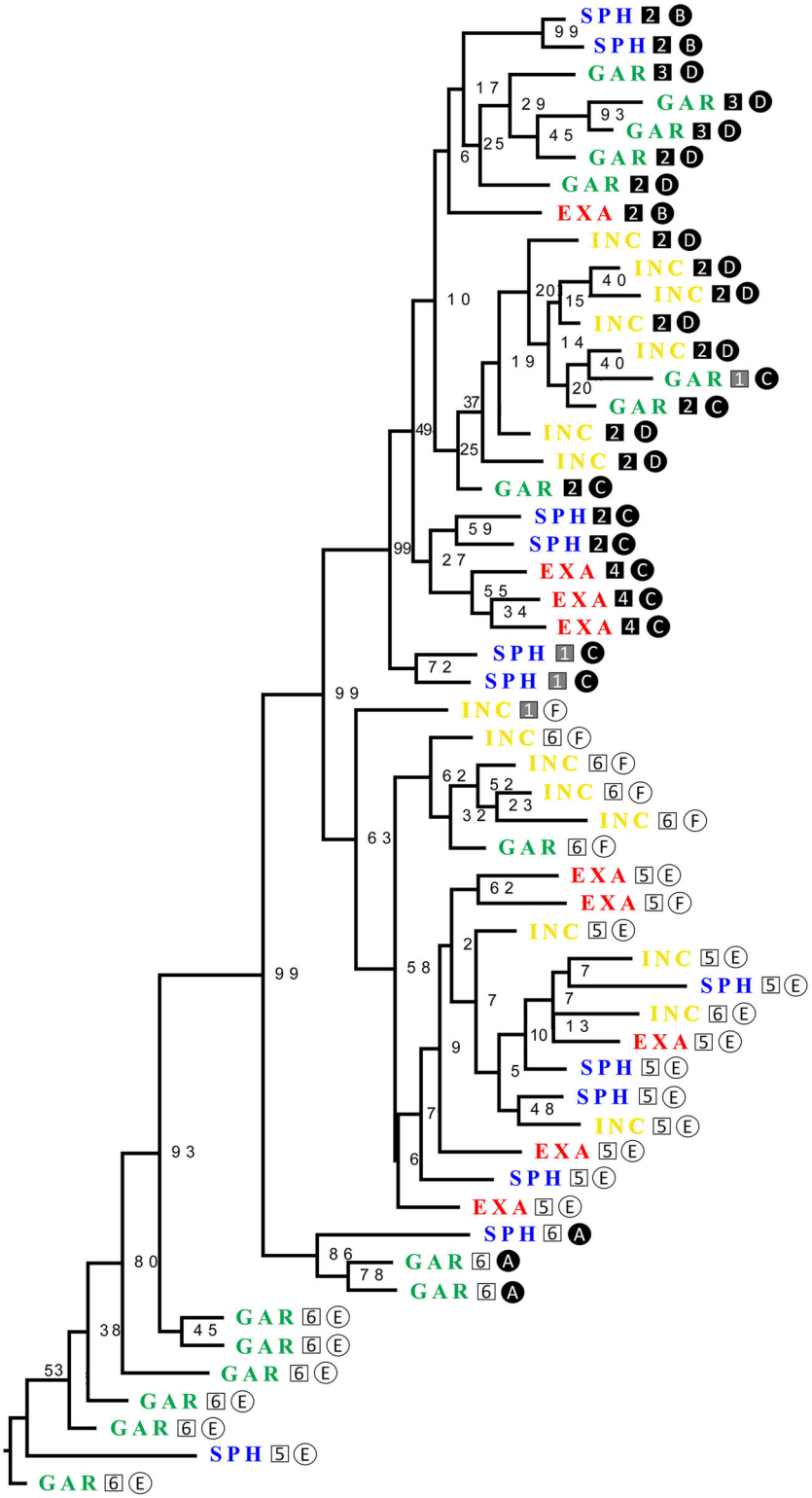

Fig. 4. RaxML tree of four species within the *Ophrys sphegodes* complex obtained by using the reduced supermatrix with loci shared by at least 70% of accessions and only including homozygous/organellar loci. Support values are derived from 1,000 bootstrap replicates. Letters in circles represent plastid haplotypes; numbers in squares represent mitochondrial haplotypes. EXA, *Ophrys exaltata*; GAR *O. garganica*; INC, *O. incubacea*; SPH, *O. sphegodes*.

were the major cause of poor resolution and confirmed the power and efficiency of multilocus approaches in identifying species circumscription and patterns of relationships among closely related species in *Ophrys*.

Higher resolution of the multilocus dataset was already detected between the nuclear single-copy *LFY* gene and amplified fragment length polymorphism (AFLP) markers for resolving the phylogeny of the *Ophrys fusca* group

(Schlüter et al., 2011a). However, compared to the AFLP approach, GBS data overcome the difficulties associated with AFLP data in assessing fragment homology in the absence of knowledge of the underlying sequence (Althoff et al., 2007). Additionally, GBS data allow for a more robust assessment of relationships, because of the larger size of the available input data matrix and the fact that they provide a codominant source of data.

However, species resolution and phylogenetic relationships (with high bootstrap support) with GBS data have been obtained mainly when selecting the larger supermatrices with a higher number of missing data (loci shared by at least 10%, 30%, or 50% accessions; Figs. 2, S4). Interestingly, a progressive decay in species resolution occurs when selecting loci with higher representation, that is, fewer loci, but less missing data (i.e., by increasing the number of accessions sharing the loci from 70% to 90%) (Figs. S5, S6). This is associated with a corresponding decrease in both average bootstrap support of resulting trees and between-lineage differentiation of the loci used, as measured by global F_{ST} (Fig. 3).

There is a debate on how both the size of the matrix and the data matrix properties (i.e., the number of missing loci and whether they are randomly distributed) might contribute to successfully disclosing patterns of relationships (Lee et al., 2018). Recent empirical studies have confirmed that larger data matrices, despite their large amount of missing data (SNPs called in a lower number of accessions), result in better resolution in delimiting very closely related species (as in Lake Victoria cichlid fishes, see Wagner et al., 2013) and simulations have shown that a higher proportion of missing data in larger data matrices does not adversely affect phylogenetic accuracy as long as there is no systematic bias (Rubin et al., 2012). The most likely explanation provided was that a less stringent filtering (i.e., inclusion of loci shared by fewer samples) preferentially retains lineage-specific loci, which might allow coalescent methods to better delimit lineages (Huang & Knowles, 2014). Accordingly, Huang & Knowles (2014), by using simulated data, showed that low tolerance to missing data leads to a disproportionately high exclusion rate of loci with high mutation rate/substitution rate. These latter loci, with a higher amount of missing data, are therefore those that have differentiated among very recently diverged species (increased F_{ST}) and could be especially informative for phylogenetic analyses. Instead, when loci with missing data are excluded in favor of more highly represented (i.e., more conserved) loci across the dataset, there is a shift in the *spectrum* of mutation rates that negatively affects the power of phylogenetic resolution. Indeed, loci conserved between distant relatives are expected to be slowly evolving. This translates into a disproportionately low number of SNPs and consequently a weak phylogenetic signal (Leaché et al., 2015). Furthermore, those loci that increase species differentiation are more likely to be under divergent/positive selection and evolving quickly, whereas the slowly evolving loci might be more likely to be neutral and thus particularly prone to be retained as ancestral polymorphisms, a phenomenon particularly relevant in very recent divergent species, such as those belonging to the *O. sphegodes* complex (Breitkopf et al., 2015). More ancestral loci (shared among many accessions) are also those with lower F_{ST} values (i.e., the more stringently filtered datasets have lower global F_{ST} value and, correspondingly, less bootstrap support). This is consistent with the idea that pollinator-driven ecological speciation in *Ophrys* might first result in divergent selection and accelerated evolution on few large-effect genes in the genome that are linked to pollinator-mediated reproductive isolation (Schlüter et al., 2011b; Sedeek et al., 2014).

The supermatrices with a high number of loci (shared by at least 10% and 30% of accessions) allow differentiating the four orchid species with *O. garganica* sister to a clade with the remaining three species (Fig. 2). Here, *O. incubacea* is sister to the inner lineage of *O. sphegodes* and *O. exaltata*. This pattern of relationships suggests a transition of pollinators between *Andrena* species (in *O. garganica*, *O. incubacea*, and *O. sphegodes*) and *Colletes* (in *O. exaltata*), a scenario congruent with a pollinator-mediated progenitor-derivative speciation (Schlüter et al., 2011a) driven by genetic change affecting flower odor emission (as hypothesized by Xu & Schlüter, 2015; see also Sedeek et al., 2016). However, as no outgroup was included, we cannot infer the direction of this evolutionary transition.

Basal relationships among species have higher support in the supermatrices with loci shared by at least 30% of accessions than in the largest supermatrix with loci shared by at least 10% of accessions that, in contrast, has higher support in the terminal clades. A potential explanation for this discrepancy is that this latter supermatrix includes a high number of loci that are shared by few individuals only (i.e., roughly 10% of the accessions) so preferentially increasing only the strength of terminal relationships.

Notably, the phylogenetic tree based on the supermatrix with a high number of loci (i.e., loci shared by at least 30% of accessions) has a clear lack of correspondence with the organellar networks (Fig. 2). The incongruence between the phylogenetic signals from organellar and fast-evolving nuclear genes, as detected in the larger supermatrix, can progressively affect bootstrap support and tree topology of the smaller supermatrices (as those built on an increased number of accessions sharing the loci, i.e., 70% and 90%; Figs. S5, S6). Indeed, by progressively selecting for more conserved, shared loci we favored the retention of organellar loci with low mutation rates and are present in most of the samples. Thus, as we aimed to clearly identify the organellar phylogenetic signal in the small supermatrix (with loci shared by at least 70% of accessions), we used an additional selective filtering for homozygous/organellar loci. The resulting smaller supermatrix generates a phylogenetic tree identifying two main supported clades (bootstrap support $\geq$90%) (Fig. 4). In these clades, individuals cluster independently from taxonomy, according to their plastid and mitochondrial haplotypes. For instance, individuals characterized by cpDNA haplotype E belong to all four distinct species. Conversely, five distinct cpDNA haplotypes (A, C, D, E, and F) are attributed to *O. garganica* individuals (Fig. 4).

The two most common cpDNA haplotypes (E and D) are five mutations different from each other and would therefore be considered as two independent evolutionary units based on the haplotype network analysis. All four *Ophrys* species contain at least one cpDNA haplotype from each of the two haplotype clusters. Network analysis of

mitochondrial DNA (mtDNA) identifies two main haplotype lineages. Almost all accessions carrying cpDNA haplotypes of lineages A–D have the mtDNA haplotype of lineages 1–4.

By analyzing the small reduced supermatrix (i.e., homozygous/organellar loci shared by at least 70% of accessions) for the presence of plastid and mitochondrial reads, we confirmed that this supermatrix included organellar loci (Table 1). However, as the *Ophrys* mitochondrial genome is not available, it could be that we misidentified several putative mitochondrial loci. Further, with these filtering strategies we also potentially selected for nuclear coding and non-coding loci that underwent purifying selection before species differentiation and hence have very little phylogenetic information (Williamson et al., 2014). Therefore, even a few informative mutations in each of the organellar genomes could determine the predominant tree signal.

The exclusion of these homozygous/organellar loci from the larger dataset (loci shared between at least 30% of accessions) used in the STRUCTURE analysis allows us to identify five groups fully corresponding to the four species assignment, with *O. incubacea* including two groups (Fig. S7). Our result contrasts with the finding of K = 6 found in the analysis of Sedeek et al. (2014). We argue that their stringent settings (≤55% missing data per accession and ≤10% missing data per locus, 1,233 loci) and the retention of homozygous/organellar loci (coupled with presence of different haplotypes in *O. incubacea*), explain the different results between the analysis presented in Sedeek et al. (2014) and in the present study.

The different gene genealogies of rapidly evolving nuclear loci compared to the slower organellar loci could explain the incongruence between the tree topologies we observe from our large (loci shared by at least 30% of accessions) and stringent (homozygous/organellar loci shared by at least 70% of accessions) supermatrices. In a rapid radiation, there has not been enough time for lineage coalescence of conserved organellar loci in each new species (Neigel & Avise, 1986).

Although *Ophrys* is relatively old (7.1–2.9 Ma), some of the species complexes (including the *O. sphegodes* complex) are estimated to be extremely young (Breitkopf et al., 2015).

This consistent pattern of variation found at organellar DNA loci implies that phylogenetic reconstructions based on organellar loci could be regarded as gene genealogies representing the older evolutionary history of the *Ophrys* lineage rather than a phylogeny that reflects the most recent organismal history (i.e., the *O. spegodes* complex). Even though cpDNA (and mtDNA) phylogenetic distributions can lack concordance with species boundaries, they still could bear the signature of the phylogeographic history of the lineages. The observed haplotype patterns, in particular the fact that cpDNA haplotypes belonging to two phyletic clusters are shared among the four species, suggests an admixture of two *Ophrys* lineages in a common ancestor of the investigated species group (i.e., the retention of haplotype diversity that was present prior to speciation in the descendant species), a scenario that could be tested by analyzing the plastid network in the genus context. These two distinct lineages might, for instance, have segregated (and diverged) in different refugia and later hybridized in secondary contact zones (Widmer & Lexer, 2001) prior to radiation within the *O. sphegodes* complex. This is consistent with the low amounts of differentiation among actual species, only detectable when using the most variable nuclear loci or pollinator-relevant phenotypic traits. Both ancestral polymorphism and signature of old hybridization are more evident in organellar than in fast-evolving nuclear regions. In the latter, polymorphism is very recent and likely emerged after the ancestral putative hybridization event. These rapidly evolving regions (with higher substitution rate) are those preferentially recovered in the large supermatrices (i.e., loci shared by at least 10% or 30% of accessions) and are likely to be the most important tool for detecting the very recent phylogenetic signal among extremely young species (Wagner et al., 2013).

Past hybridization has been advocated at the bases of recent species radiations, as in the Hawaiian silverswords (Barrier et al., 1999) and in African cichlid fishes (Meier et al., 2017). Hybridization occurring when allopatric lineages come into secondary contact could fuel the onset of an adaptive radiation by providing a new genetic background for novel trait combinations or for increasing genotypic diversity (Abbott et al., 2013) that, in *Ophrys*, can allow the exploitation of new available pollinator niches and, consequently, the evolution of premating isolation (Breitkopf et al., 2013, 2015; Vereecken et al., 2010). Although incomplete lineage sorting is difficult to distinguish from reticulation, our results including fast-evolving loci suggest that current hybridization is at least unlikely to occur frequently among the four species in the sympatric study region. This has been further corroborated by local experimental studies confirming premating isolation among the four *Ophrys* species due to pollinator isolation (Xu et al., 2011; Sedeek et al., 2014).

In conclusion, we present a well-resolved phylogenetic tree from a group that has represented a challenge due to its recent origin and weak genomic differentiation (Breitkopf et al., 2013). Although the different levels of information contained in GBS loci with different substitution rates and genealogy should be properly accounted for, our finding that these sympatric *Ophrys* species form well-supported lineages highlights the power that NGS-based data holds for resolving species boundaries, particularly in groups with complex evolutionary histories.

Acknowledgements

This research was carried out in the frame of Programme STAR, financially supported by UniNA and Compagnia di San Paolo, as well as the Swiss National Science Foundation (SNF grant 31003A_155943 to PMS). We thank Riccardo Aiese Cigliano and Walter Sanseverino (Sequentia Biotech sl) for bioinformatics support and Laura Piñeiro Fernandez and Karl Duffy for comments on the MS. We also thank Mary Longrigg for language revision.

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, Butlin RK, Dieckmann U, Eroukhmanoff F, Grill A, Cahan SH, Hermansen JS, Hewitt G, Hudson AG, Jiggins C, Jones J, Keller B, Marczewski T, Mallet J, Martinez-Rodriguez P, Möst M, Mullen S, Nichols R,

Nolte AW, Parisod C, Pfennig K, Rice AM, Ritchie MG, Seifert B, Smadja CM, Stelkens R, Szymura JM, Väinölä R, Wolf JBW, Zinner D. 2013. Hybridization and speciation. *Journal of Evolutionary Biology* 26: 229–246.

Althoff DM, Gitzendanner MA, Segraves KA. 2007. The utility of amplified fragment length polymorphisms in phylogenetics: A comparison of homology within and between genomes. *Systematic Biology* 56: 477–484.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* 3: e3376.

Barrier M, Baldwin BG, Robichaux RH, Purugganan MD. 1999. Interspecific hybrid ancestry of a plant adaptive radiation: Alloploidy of the Hawaian silversword alliance (Asteraceae) inferred from floral homeotic gene duplication. *Molecular Biology and Evolution* 16: 1105–1113.

Bateman RM, Bradshaw E, Devey DS, Glover BJ, Malmgren S, Sramko G, Murphy Thomas M, Rudall PJ. 2011. Species arguments: Clarifying competing concepts of species delimitation in the pseudo-copulatory orchid genus *Ophrys*. *Botanical Journal of the Linnean Society* 165: 336–347.

Bateman RM, Sramkó G, Paun O. 2018. Integrating restriction site-associated DNA sequencing (RAD-seq) with morphological cladistic analysis clarifies evolutionary relationships among major species groups of bee orchids. *Annals of Botany* 121: 85–105.

Breitkopf H, Onstein RE, Cafasso D, Schlüter PM, Cozzolino S. 2015. Multiple shifts to different pollinators fuelled rapid diversification in sexually deceptive *Ophrys* orchids. *New Phytologist* 207: 377–389.

Breitkopf H, Schlüter PM, Xu S, Schiestl FP, Cozzolino S, Scopece G. 2013. Pollinator shifts between *Ophrys sphegodes* populations: Might adaptation to different pollinators drive population divergence? *Journal of Evolutionary Biology* 26: 2197–2208.

Bruneaux M, Johnston SE, Herczeg G, Merilä J, Primmer CR, Vasemägi A. 2013. Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Molecular Ecology* 22: 565–582.

Cafasso D, Widmer A, Cozzolino S. 2004. Chloroplast DNA inheritance in the orchid *Anacamptis palustris* using single-seed polymerase chain reaction. *Journal of Heredity* 96: 66–70.

Cariou M, Duret L, Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. *Ecology and Evolution* 3: 846–852.

Carstens BC, Pelletier TA, Reid NM, Satler JD. 2013. How to fail at species delimitation. *Molecular Ecology* 22: 4369–4383.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: An analysis tool set for population genomics. *Molecular Ecology* 22: 3124–3140.

Darwin C. 1862. On the various contrivances by which British and foreign orchids are fertilised by insects: And on the good effect of intercrossing. Cambridge: Cambridge Library Collection.

Davey JW, Blaxter ML. 2010. RADSeq: Next-generation population genetics. *Briefings in Functional Genomics* 9: 416–423.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology &. *Evolution* 24: 332–340.

Delforge P. 2016. Guide des orchidées d'Europe, d'Afrique du Nord et du Proche-Orient, Paris: Delachaux et Niestle.

Devey DS, Bateman RM, Fay MF, Hawkins JA. 2008. Friends or relatives? Phylogenetics and species delimitation in the controversial European orchid genus *Ophrys*. *Annals of Botany* 101: 385–402.

Doyle JJ. 1997. Trees within trees: Genes and species, molecules and morphology. *Systematic Biology* 46: 537–553.

Eaton DAR. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30: 1844–1849.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63: 1–19.

Edwards SV, Beerli P. 2000. Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54: 1839–1854.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution* 29: 51–63.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6: e19379.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences USA* 107: 16196–16200.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology* 14: 2611–2620.

Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. *Trends in Genetics* 28: 342–350.

Foll M, Gaggiotti OE. 2008. A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* 180: 977–993.

Gielly L, Taberlet P. 1994. The use of chloroplast DNA to resolve plant phylogenies: Noncoding versus *rbcL* sequences. *Molecular Biology and Evolution* 11: 769–777.

Hernández-León S, Gernandt DS, de laRosa JAP, Jardón-Barbolla L. 2013. Phylogenetic relationships and species delimitation in *Pinus* section *Trifoliae* inferred from plastid DNA. *PloS One* 8: e70501.

Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, Manos PS. 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. *PloS One* 9: e93975.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PloS Genetics* 6: e1000862.

Huang H, Knowles LL. 2014. Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Systematic Biology* 65: 357–365.

Knowles LL. 2009. Statistical phylogeography. *Annual Review of Ecology Evolution and Systematics* 40: 593–612.

Knowles LL, Maddison WP. 2002. Statistical phylogeography. *Molecular Ecology* 11: 2623–2635.

Kullenberg B. 1961. Studies in *Ophrys* pollination. *Zoologiska Bidrag Fran Uppsala* 34: 1–340.

Leaché AD, Banbury BL, Felsenstein J, DeOca ANM, Stamatakis A. 2015. Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology* 64: 1032–1047.

Lee KM, Kivelä SM, Ivanov V, Hausmann A, Kaila L, Wahlberg N, Mutanen M. 2018. Information dropout patterns in RAD phylogenomics and a comparison with multilocus Sanger data in a species-rich moth genus. *Systematic Biology* 67: 925–939.

Leigh JW, Bryant D. 2015. PopART: Full-feature software for haplotype network construction. *Methods in Ecology and Evolution* 6: 1110–1116.

Lewis ZA, Shiver AL, Stiffler N, Miller MR, Johnson EA, Selker EU. 2007. High-density detection of restriction-site-associated DNA markers for rapid mapping of mutated loci in Neurospora. *Genetics* 177: 1163–1171.

Li H 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: 1303.3997.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Lischer HEL, Excoffier L, Heckel G. 2014. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: Implications for the evolutionary history of *Microtus voles*. *Molecular Biology and Evolution* 31: 817–831.

Maddison Wayne P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297–1303.

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications* 8: 14363.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17: 240–248.

Neigel JE, Avise JC. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Karlin S, Nevo E eds. *Evolutionary Processes and Theory*. New York: Academic Press. 515–534.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327: 92–94.

Paulus HF, Gack C. 1990. Pollinators as prepollinating isolation factors: evolution and speciation in Ophrys (Orchidaceae). *Israel Journal of Botany* 39: 43–79.

Petit RJ, Duminil J, Fineschi S, Hampe A, Salvini D, Vendramin GG. 2005. Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology* 14: 689–701.

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution* 54: 396–402.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

Roma L, Cozzolino S, Schlüter PM, Scopece G, Cafasso D. 2018. The complete plastid genomes of *Ophrys iricolor* and *O. sphegodes* (Orchidaceae) and comparative analyses with other orchids. *PloS One* 13: e0204174.

Roure B, Baurain D, Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution* 30: 197–214.

Rowe HC, Renaut S, Guggisberg A. 2011. RAD in the realm of next-generation sequencing technologies. *Molecular Ecology* 20: 3499–3502.

Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PloS One* 7: e33394.

Sang T, Crawford DJ, Stuessy TF. 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *American Journal of Botany* 84: 1120–1136.

Schlüter PM, Harris SA. 2006. Analysis of multilocus fingerprinting data sets containing missing data. *Molecular Ecology Notes* 6: 569–572.

Schlüter PM, Kohl G, Stuessy TF, Paulus HF. 2007. A screen of low-copy nuclear genes reveals the *LFY* gene as phylogenetically informative in closely related species of orchids (*Ophrys*). *Taxon* 56: 493–504.

Schlüter PM, Ruas PM, Kohl G, Ruas CF, Stuessy TF, Paulus HF. 2011a. Evidence for progenitor-derivative speciation in sexually deceptive orchids. *Annals of Botany* 108: 895–906.

Schlüter PM, Xu S, Gagliardini V, Whittle E, Shanklin J, Grossniklaus U, Schiestl FP. 2011b. Stearoyl-acyl carrier protein desaturases are associated with floral isolation in sexually deceptive orchids. *Proceedings of the National Academy of Sciences USA* 108: 5696–5701.

Scopece G, Musacchio A, Widmer A, Cozzolino S. 2007. Patterns of reproductive isolation in Mediterranean deceptive orchids. *Evolution* 61: 2623–2642.

Sedeek KE, Scopece G, Staedler YM, Schönenberger J, Cozzolino S, Schiestl FP, Schlüter PM. 2014. Genic rather than genome—wide differences between sexually deceptive *Ophrys* orchids with different pollinators. *Molecular Ecology* 23: 6192–6205.

Sedeek KE, Whittle E, Guthörl D, Grossniklaus U, Shanklin J, Schlüter PM. 2016. Amino acid change in an orchid desaturase enables mimicry of the pollinator's sex pheromone. *Current Biology* 26: 1505–1511.

Small RL, Cronn RC, Wendel JF. 2004. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* 17: 145–170.

Soliva M, Kocyan A, Widmer A. 2001. Molecular phylogenetics of the sexually deceptive orchid genus Ophrys (Orchidaceae) based on nuclear and chloroplast DNA sequences. *Molecular Phylogenetics and Evolution* 20: 78–88.

Stölting KN, Nipper R, Lindtke D, Caseys C, Waeber S, Castiglione S, Lexer C. 2013. Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology* 22: 842–855.

Vereecken NJ, Cozzolino S, Schiestl FP. 2010. Hybrid floral scent novelty drives pollinator shift in sexually deceptive orchids. *BMC Evolutionary Biology* 10: 103.

Vereecken NJ, Streinzer M, Ayasse M, Spaethe J, Paulus HF, Stoekl J, Cortis P, Schiestl FP. 2011. Integrating past and present studies on *Ophrys* pollination — a comment on Bradshaw et al. *Botanical Journal of the Linnean Society* 165: 329–335.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* 22: 787–798.

Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Featherstone AW, Pellicer J, Buggs RJA. 2013. Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Molecular Ecology* 22: 3098–3111.

Wang X, Ye X, Zhao L, Li D, Guo Z, Zhuang H. 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the

phylogeny of temperate bamboos (Poaceae: Bambusoideae). *Scientific Reports* 7: 11546.

Widmer A, Lexer C. 2001. Glacial refugia: Sanctuaries for allelic richness, but not for gene diversity. Trends in Ecology & *Evolution* 16: 267–269.

Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLOS Genetics* 10: e1004622.

Wolfe KH, Sharp PM, Li WH. 1989. Rates of synonymous substitution in plant nuclear genes. *Journal of Molecular Evolution* 29: 208–211.

Xu S, Schlüter PM. 2015. Modeling the two-locus architecture of divergent pollinator adaptation: How variation in SAD paralogs affects fitness and evolutionary divergence in sexually deceptive orchids. *Ecology and Evolution* 5: 493–502.

Xu S, Schlüter PM, Scopece G, Breitkopf H, Gross K, Cozzolino S, Schiestl FP. 2011. Floral isolation is the main reproductive barrier among closely related sexually deceptive orchids. *Evolution* 65: 2606–2620.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12493/suppinfo:

Fig. S1. Statistical parsimony haplotype network based on four mitochondrial loci filtered from the GBS data and shared by all individuals. Square size is proportional to haplotype frequency. In parentheses the number of individuals.

Fig. S2. Distribution of number of SNPs per locus in the 54 individuals, from the original dataset from Sedeek et al. (2014).

Fig. S3. RAXML tree obtained by using the software RRHS on the supermatrix with loci shared by at least 30% of accessions. EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1000 bootstrap 776 replicates.

Fig. S4. RAXML tree obtained by using the supermatrix with loci shared by at least 50% of accessions. EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1000 bootstrap replicates.

Fig. S5. RAXML tree obtained by using the supermatrix with loci shared by at least 70% of accessions. EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1000 bootstrap replicates.

Fig. S6. RAXML tree obtained by using the supermatrix with loci shared by at least 90% of accessions. EXA = *Ophrys exaltata*, GAR = *O. garganica*, INC = *O. incubacea*, SPH = *O. sphegodes*. Support values are derived from 1000 bootstrap replicates.

Fig. S7. Neighbor Joining (NJ) tree, Bayesian assignment bar graph and Plot of delta K values from the Structure analyses based on: (A) the dataset with homozygous/organellar loci shared by at least 70% of accessions; (B) the dataset with loci shared by at least 30% of accessions; (C) the dataset with loci shared by at least 30% of accessions after exclusion of homozygous/organellar loci. Red = *Ophrys exaltata*; Green = *O. garganica*; Yellow/orange = *O. incubacea*; Blue = *O. sphegodes*. Dark grey and white circles represent the two plastid haplotype lineages identified in the network analysis.Supporting information.