A comprehensive framework for the delimitation of species within the *Bemisia tabaci* cryptic complex, a global pest-species group

Hua-Ling Wang[1,2,3#] , Teng Lei[4], Xiao-Wei Wang[2], Stephen Cameron[5], Jesús Navas-Castillo[6], Yin-Quan Liu[2], M. N. Maruthi[3], Christopher A. Omongo[7], Hélène Delatte[8], Kyeong-Yeoll Lee[9], Renate Krause-Sakate[10], James Ng[11], Susan Seal[3], Elvira Fiallo-Olivé[6], Kathryn Bushley[12], John Colvin[3#] and Shu-Sheng Liu[2#]

[1]College of Forestry, Hebei Agricultural University, No. 2596, Lekainan Street, Baoding, 071000, Hebei, China
[2]The Ministry of Agriculture Key Laboratory of Molecular Biology of Crop Pathogens and Insects, Institute of Insect Sciences, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, China
[3]Natural Resources Institute, University of Greenwich, Kent ME4 4TB, United Kingdom
[4]College of Life Sciences, Taizhou University, Taizhou, 318000, China
[5]Department of Entomology, Purdue University, 915 West State Street, West Lafayette, IN 479074
[6]Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora" (IHSM-UMA-CSIC), Consejo Superior de Investigaciones Científicas, 29750 Algarrobo-Costa, Málaga, Spain
[7]National Crops Resources Research Institute, Namulonge, P.O. Box 7084, Kampala, Uganda
[8] CIRAD, UMR PVBMT CIRAD, Pôle de Protection des Plantes, 7 chemin de l'IRAT, F-97410 Saint-Pierre
[9]School of Applied Biosciences, Kyungpook National University, Daegu 702-701, Republic of Korea
[10]UNESP, Faculdade de Ciências Agronômicas, 18610-307, Botucatu, Brazil
[11]Department of Microbiology and Plant Pathology, University of California, Riverside, California 92521, USA
[12]USDA Agricultural Research Service, 17123, Emerging Pests and Pathogens Research Unit, 538 Tower Road, Ithaca, New York, United States, 14853.
Correspondence should be addressed to Hua-Ling Wang, John Colvin and Shu-Sheng Liu.
wang_hual@126,com; j.colvin@greenwich.ac.uk; shshliu@zju.edu.cn.

Key words: *Bemisia tabaci*, cryptic species, phylogenetics, RAD-Seq, reproductive compatibility, whiteflies

Running Head: Decoding *Bemisia tabaci* cryptic species complex

Abstract

Identifying cryptic species poses a substantial challenge to both biologists and naturalists due to the morphological similarities. *Bemisia tabaci sensu lato* is a cryptic species complex containing more than 44 putative species; several of which are currently among the world's most destructive crop pests. Interpreting and delimiting the evolution of this species complex has proved problematic. To develop a comprehensive framework for the species delimitation and identification, we evaluated the performance of distinct data sources both individually and in combination among numerous samples of the *B. tabaci* species complex acquired worldwide. Distinct datasets include full mitogenomes, single-copy nuclear genes, restriction site–associated DNA sequencing, geographic range, host speciation, and reproductive-compatibility datasets. Phylogentically, our well-supported topologies generated from three dense molecular markers highlighted the evolutionary divergence of species of the *B. tabaci* complex and suggested that the nuclear markers sever as a more accurate representation of *B. tabaci* species diversity. Reproductive compatibility datasets facilitated the identification of at least 17 different cryptic species within our samples, confirming that the *B. tabaci* complex comprises multiple cryptic species. Native geographic range information provides a complementary assessment of species recognition, while the host ranges datasets provide low rate of delimiting resolution. We further summarized different data performance in species classification when compared with the reproductive compatibility indicating that combination of *mtCOI* divergence, nuclear markers, native geographic range (at least inclusion of collecting location) provide a complementary assessment of species recognition. Finally, we represent a model for understanding and untangling the cryptic species complexes

based on the evidence from this study and previously published articles.

Introduction

Cryptic species refers to taxa that are morphologically similar but biologically and genetically divergent (Pfenninger and Schwenk, 2007; Bickford *et al*., 2007). They are present in a wide range of taxonomic groups and represent a large amount of undiscovered genetic and functional biodiversity (Jörger and Schrödl 2013; Pante *et al*., 2015; Loxdale *et al*., 2016). Developing approaches for systematically studying cryptic species, would contribute significantly to our understanding of biodiversity, and connecting the study of taxonomy and phylogenetic patterns with evolutionary processes and ecosystems functioning, such as speciation (Struck *et al*. 2017). Cryptic species comprise a significant proportion of biodiversity in some regions (Smith *et al*. 2008). Additionally, many putative cryptic species have been identified in high-ranking threatened and agro-economically important species (Nater *et al*., 2017; De Barro *et al*., 2011), emphasizing the need for accurate identification of such cryptic species (Bickford *et al*., 2007). However, species delimitation of these morphologically similar cryptic species has posed a number of challenges to taxonomists and biologists even before the Linnaean classification system was adopted (Bickford *et al*., 2007). Cryptic species face two major challenges of species delimitation. First, many species form on a continuum, meaning that populations gradually become more divergent across space and time (Darwin 1859; Mayr 1942; Mallet 1995; De Queiroz 2007). Thus, there is disagreement on how much sequence divergence of DNA barcoding regions (Moritz & Cicero, 2004) or morphological traits is sufficient to name lineages as species (Gómez *et al*., 2002; McDaniel & Shaw, 2003). Second, speciation proceeds heterogeneously across many dimensions (Singhal *et al*., 2018). Therefore, an integrated approach is preferable in

addressing issues pertaining to species level taxonomic assignment (Padial *et al*., 2010; Dayrat 2005; Wiens & Penkrot, 2002; Schlick-Steiner *et al*., 2010).

We focus here on the whitefly, *Bemisia tabaci sensu lato* (Gennadius) (Hemiptera: Aleyrodidae), a species complex, that is widely distributed throughout tropical and subtropical regions. *B. tabaci* are agro-economically important phloem-feeding pests, causing extensive crop damage through direct feeding, deposition of honeydew or virus transmission (Dinsdale *et al*., 2010; Boykin *et al*., 2012; Mugerwa *et al*., 2018). The complex has been proposed to comprise more than 48 morphologically indistinguishable species largely based solely on DNA-sequencing methods using a single barcode region (*mtCOI*) (Dinsdale *et al*., 2010). Different cryptic species of the complex differ in many aspects of their biology and ecology, including host-plant range (Sun *et al*., 2013), resistance to insecticides (Wang *et al*., 2010; Horowitz *et al*., 2004; Sun *et al*., 2013), behavior (Liu *et al*., 2007; Crowder *et al*., 2010), and the capacity and specificity of virus transmission (Polston *et al*., 2014; Wei *et al*., 2014). The inability to distinguish between genetically and functionally different species poses a huge challenge for identifying and managing these pests. In particular, the ability of its members to vector plant viruses, predominantly the genus *Begomovirus*, poses a significant threat to farming communities across in many countries (Amari *et al*., 2008; NavasCastillo *et al*., 2011; Ghosh *et al*., 2019). For example, one putative species currently known as Sub-Saharan Africa 1, vectoring cassava mosaic disease (CMD) and cassava brown streak disease (CBSD) in cassava, causes annual losses of more than US$1.25 billion in the production of the staple-food cassava in nine East and Central African countries (Legg *et al*., 2006; Maruthi

et al., 2005; Mware *et al*., 2009). The Mediterranean (MED) and Middle East-Asia Minor 1 (MEAM1) putative species are globally invasive pests that have attracted attention owing to their highly invasive ability, polyphagia, the ability to transmit several emergent crop viruses, and the capacity for developing insecticide resistance (Naranjo *et al*., 2009; Navas-Castillo *et al*., 2011). Overall, regions affected by different cryptic species of whiteflies are continuing to expand, leading to outbreaks of plant diseases, causing hunger and food insecurity (Alicai *et al*., 2007; Hahn *et al*., 1985; Thresh *et al*., 1997; Legg *et al*., 2014). Given the significant economic losses caused by *B. tabaci* worldwide, the accurate identification of distinct species within this complex and understanding the relationships among them are clearly essential to ensure the implementation of appropriate quarantine regulations for restricting the spread of any particular *B. tabaci* species.

Initially, in the 1950s, *B. tabaci* populations were recognized as biotypes or host races after they were found to be specialized on specific host plants and/or had unique plant-virus transmission capability (Bird, 1957; Mound, 1963). Since then, the classification and systematics of *B. tabaci* have remained controversial (Brown *et al*., 1995; Gill & Brown, 2009; De Barro *et al*., 2011) because there is considerable overlap in their morphological characters present in ‘host-races’ or ‘biotypes’ during the fourth instar and pupal stages (Russell, 1957; Mound & Halsey, 1978). This controversy has generated confusion in the nomenclature (Boykin 2014), and much of the scientific literature still refers to *B. tabaci* as a single species. Consequently, *B. tabaci* has been listed as one of the world’s most destructive invasive species (Russell 1957; Lowe *et al*., 2004) without

reference to accurate species names.

Advances in molecular approaches and tools that are inexpensive and easily accessible have initiated a completely new era in species delimitation and taxonomy (Bickford *et al*., 2006; Herbert *et al*., 2003). The emergence of the *mtCOI* barcoding method in the late 1990s was a breakthrough, allowing the characterization of species based on genetic data. Phylogenetic analyses based on their 3′ partial *mtCOI* sequences (a 3.5% *mtCOI* divergence is used to designate different species) suggested that *B. tabaci* is a cryptic species complex (Frohlich *et al*., 1999; Boykin *et al*., 2007; Dinsdale *et al*., 2010; Boykin *et al*., 2012). Subsequently, crossing experiments that demonstrated reproductive incompatibility among some putative species that provided further support to the proposition that *B. tabaci* is a cryptic species complex (Maruthi *et al*., 2004; Wang *et al*., 2011; Liu *et al*., 2012). Since reproductive incompatibility is a widely accepted approach for confirming the existence of distinct *B. tabaci* species, we have adopted the biological species concept (BSC), which defines species as interbreeding populations that are reproductively isolated from each other (Mayr 1942, 1963; Noor 2002), as the definitive test of species delimitation. It has been recommended that the BSC be used to avoid “taxonomic over-inflation” (Isaac *et al*., 2004; Singhal *et al*., 2018). Reciprocal cross-breeding experiments, however, are technically demanding and often not possible for invasive pest-species where moving live material between countries to conduct such experiments is illegal. Although *B. tabaci* is now generally recognized as a species complex, questions remain about how many species exist and their relationships. A genetic framework that reflects BSC boundaries in *B. tabaci* and that could be applied

for distinct species identification is needed.

Although *B. tabaci* was recognized as a cryptic species complex based on a 3.5% divergence (657bp) of the *mtCOI* sequence (Frohlich *et al*., 1999; Boykin *et al*., 2007; Dinsdale *et al*., 2010; Boykin *et al*., 2012; Lee *et al*., 2011), Lee *et al*. (2011) suggested that this threshold percentage was not accurate and instead estimated the accurate threshold to be 4.0%. Based on these findings, 44 putative *B. tabaci* species were recognized by employing the 4.0% threshold (Kanakala & Ghanim, 2019). Because the presence of nuclear mitochondrial (NUMT) DNA/pseudogenes from the nuclear genome can result in an over-estimation of species diversity- for example, a previously known species MEAM2 (Delatte *et al*., 2007) was later shown to be an artifact (Tay *et al*., 2017). Therefore, the *mtCOI* sequence alone is not sufficient to unambiguously delimit and resolve the number of *B. tabaci* species (De Queiroz, 2005).

To date, the phylogenetic relationships of the major lineages of the *B. tabaci* cryptic species complex remain unresolved. Phylogenetic trees inferred from molecular datasets, especially large-scale datasets, are recognized as a necessary framework for comparative study of a wide spectrum of biological organisms (Hall *et al*., 2002). In particular, phylogenies of multiple genes have proven successful, along with traditional methods for accurate identification of different organisms (Doyle *et al*., 2003), both within and outside the traditional boundaries of evolutionary biology (Nadler 1995). Previous phylogenetic markers used in molecular phylogeny analyses of diverse samples of *B. tabaci* have included *mtCOI*, ribosomal ITS1, and a few single-copy nuclear genes, but

182 firm conclusions have yet to be reached owing to the short lengths of these sequences or
183 limited sample sizes (De Barro *et al*., 2000; Hsieh *et al*., 2014). Mitochondrial genomes
184 and single nucleotide polymorphisms (SNPs) derived from Restriction site–associated
185 DNA sequencing (RAD-Seq) are also utilized to identify species from sub-Saharan Africa
186 clades that previously were thought to comprise a single *mtCOI*-defined species
187 (Mugerwa *et al*., 2020; Vyskočilová *et al*., 2018; Wosula *et al*., 2017; Elfekih *et al*., 2021).
188 However, the performance of these markers with respect to the resolution globally of this
189 important cryptic species complex has not been completely explored and compared.
190
191 Apart from the molecular phylogenetic approach, tremendous efforts have been directed
192 towards delimiting members of the *B. tabaci* cryptic species using other methods,
193 including reproductive compatibility (Liu *et al*., 2012) and machine learning based on
194 differences in puparium morphology (Macleod *et al*., 2022). For example, crossing
195 experiments have proven partial, or complete reproductive isolation between many
196 populations identified as distinct species based on *mtCOI* sequences (Liu *et al*., 2012; Qin
197 *et al*., 2015; Vysko˘cilová *et al*., 2019; Mugerwa *et al*., 2021). Recently, Macleod *et al*.
198 (2022) demonstrated that the 15 species that they analyzed could be discriminated
199 successfully according to the puparium morphology differences alone using a machine
200 learning approach. These methods offer support for the existence of multiple cryptic
201 species in *B. tabaci* and provide alternative means of species identification. However,
202 each individual method displays deficiencies and limitations, and fails to recognize
203 species in some cases (Sukumaran & Knowles, 2017; Frézal & Leblois, 2008). For
204 instance, the *mtCOI* distance-based method did not recover the same species groups as

the mating approach in the Asia II (Qin *et al*., 2016), MED (Vyskočilová *et al*., 2018), and Sub-Saharan Africa 1 (Mugerwa *et al*., 2021) groups of species. Additionally, there is a considerable lack of congruence between relationships generated from mitochondrial genes versus nuclear genes in delimiting *B. tabaci* cryptic species (Mugerwa *et al*., 2021). Given the considerable failure rate when single data sources and approaches were used for species delimitation, an integrated, multisource approach was proposed to increase the rigor of species delimitation (Dayrat, 2005; Tan *et al*., 2009; Schlick-Steiner *et al*., 2010; Palandaéiæ *et al*., 2017; Dzhembekova *et al*., 2020; Macleod *et al*., 2022).

It has been suggested that the combination of different datasets, including but not limited to comprehensive genome datasets, geographic distribution, reproductive incompatibilities and host ranges (Heraty *et al*., 2007; Liu *et al*., 2012; Struck *et al*., 2017; Wosula *et al*., 2017; Macleod *et al*., 2022), can either provide additional support for a species hypothesis or uncover contradictions between datasets. However, no systematic comparison has yet been carried out on the performance of different data sources and approaches in identifying the distinct species of *B. tabaci.* The objectives of this research are therefore to utilize the previously published data sources, as well as those generated from this current research to develop a comprehensive framework for species delimitation and identification by evaluating the performance of distinct data sources both individually and in combination.

Our study involved analyzing and integrating the following information: molecular, mating, and geographical divergence. Molecular data is commonly used for interpreting

species limits and relationships as it contains a large number of characters that have high level of resolution in distinguishing population groups (phylogeographic patterns). This level of resolution is not often achieved by other types (e.g., morphological) of datasets (Bickford *et al*., 2006). Since incongruences has been observed between mitochondrial /*mtCOI* and nuclear genes in delimiting the SSA1 species (Mugerwa *et al*., 2021), we aimed to further test whether this incongruence occurs across the entire *B. tabaci* cryptic species complex by obtaining frequently used genetic markers, such as mitogenomes, single-copy nuclear genes, and genome level SNPs from selected species across all lineages in *B. tabaci*.

Levels of reproductive incompatibility through mating experiments are generally considered a more direct and conclusive approach (Paterson *et al.*, 2016) to confirm true cryptic species. Many *B. tabaci* species have been verified using mating experiments, but the consistency of this approach has not been compared with the other methods of species delimitation across the phylogenetic lineages of *B. tabaci* on a large scale. We thus tested the hypothesis that integrating both molecular genetics and mating data could greatly aid in species delimitation of *B. tabaci*.

Geographical divergence was recognized as one of the key forces driving the diversification of the *B. tabaci* species complex, as this diversification was thought to be associated with the separation of continental landmasses (Boykin *et al*., 2013; De Barro *et al*., 2005). However, it has been argued that the role of the insect's host plants in the divergence of the species complex is unlikely because most species have the ability to

colonize multiple plant hosts (De Barro 2005; De Barro *et al*., 2005). To clarify these opposing perspectives, we determined whether evolved differences in geographical range and host utilization could have played a role in the diversification of the *B. tabaci* species complex. Outcomes from our analyses would potentially provide additional support for the operational designation of distinct species based on both mating as well as molecular phylogenic approaches or help define species when these other approaches fail or disagree.

To test this integrated approach across the global distribution of *B. tabaci*, we first generated mitogenomes, four single-copy nuclear genes, and RAD-Seq data for 25 specimens, representing the global-scale distribution of *B. tabaci*. These data were used to construct phylogenetic relationships and to determine the congruencies among different molecular phylogenies. Based on this new molecular phylogeny, we conducted a literature survey of mating experiments that have been conducted among the specimens that we used and compared them against the phylogenetic tree. Next, geographic information and host range variation, which could also aid in species delimitation, were investigated and compared to the phylogenetic framework. Finally, we synthesized our results to propose an integrated framework that can be applied for the identification of distinct species. We also addressed how our results may direct future analyses of the systematics of the *B. tabaci* species complex and other cryptic species complexes.

Materials and methods

Taxonomic sampling. Twenty five members of the whitefly cryptic species complex

were selected as test samples, using the 3.5% *mtCOI* sequence divergence cutoff (*SI Appendix*, Table S1) (hereafter called *mtCOI* defined species) within the *B. tabaci* complex (Dinsdale *et al*., 2010). These members cover 10 major clades according to the phylogeny inferred by the *mtCOI* (*SI Appendix*, Fig. S1, Table S2) across the world (Boykin *et al*., 2007). Specimen names, GenBank accession numbers, localities of collection and distributions are listed in Table S1. The specimen purity was determined by *mtCOI* sequencing. Three samples including Mediterranean (MED), Mediterranean_Sudan (MED_Sudan) and Mediterranean_Uganda_Sweet Potato (MED_Uganda_SP) were considered as a single MED species based on a 3.5% *mtCOI* sequence divergence cutoff. Genomic DNA was extracted from these samples with the Qiagen Blood and Tissue kit (Valencia, CA, USA).

Sequencing strategy. Sanger sequencing was used to obtain mitogenomes and single-copy nuclear genes from single females of the 25 *B. tabaci* samples and from two specimens of a closely related species *Bemisia afer*. Illumina HiSeq 2000 was used to obtain RAD-Seq datasets from 20 individuals of each sample. Due to small sample size, six *B. tabaci* and one *B. afer* were eliminated, therefore, only 19 *B. tabaci* and 1 *B. afer* specimen were used for RAD-Seq sequencing. The RAD library construction, sample indexing and pooling followed the descriptions of Baird *et al*. (2008). Briefly, EcoRI was selected to digest the genomic DNA after testing various restriction enzymes. A combinatorial pooled sequencing strategy (Cao *et al*., 2016) was used as it provides a cost-effective alternative to sequencing individuals separately. Four multiplexed sequencing libraries were constructed, in which each DNA sample was assigned a unique

297 nucleotide multiplex identifier for barcoding (*SI Appendix*, Table S3). Illumina HiSeq
298 2000 was used to perform single-end sequencing.
299
300 **Sanger dataset preparation and sequencing.** For the complete mitogenomes, primer
301 sequences and amplification protocols followed Wang *et al*. (2013). The *Adh*, *Ef-1α* and
302 *RNA II* single-copy nuclear genes were amplified using primer pairs as follows: forward
303 (F), 5′GGATGCTTGAGCAATTCTTTGT3′ and reverse (R),
304 5′GCTTTAGAAATTGGTTACCGTCA3′; F,
305 5′ACCATACCTGGTTTGATMACTCCRGT3′ and R,
306 5′CMTGGTTCAAGGGATGGCARAT 3′; and F, 5′TCGGAGACACAATTGCT 3′ and R,
307 5′TNCTGTACATTCCAAATCATAC 3′, respectively. The *Prp8* locus was amplified
308 using F, 5′GCCTTGGGAGGTGTTGAAG 3′ and R, 5′GGCTTGCATCCAGGGTACC 3′
309 (Hsieh *et al*., 2014). The PCR reaction consisted of denaturation for 3 min at 94°C, then
310 35 cycles of denaturation for 30 s at 94°C, annealing for 30 s at 54–60°C, extension at
311 72°C for 2 min and a final extension for 10 min at 72°C. PCR fragments were purified
312 and ligated into the pGEM-T Easy Vector (Promega Corp, Madison, WI, USA) for
313 sequencing in both directions using the ABI BigDye 3.1 at GenScript (Nanjing, China).
314 All sequences were deposited in GenBank (*SI Appendix*, Table S2). The full mitogenomes
315 were obtained for all species except Uganda and Asia II 1, for which the second control
316 regions were not completely obtained because of many hidden random repeats in this
317 region. In total, we generated 23 newly sequenced mitogenomes from a representative
318 member of each distinct population of *B. tabaci* species groups.
319

Annotation of the mitogenomes. The protein-coding genes (PCGs) of 13 mitochondrial genomes were annotated using DOGMA (Wyman *et al*., 2004). The tRNAs were identified using tRNAscan-SE (Lowe and Eddy, 1997) or recognized manually (Wolstenholme, 1992). The 5′-end of the *rrnL* gene was assumed to be delimited by the ends of *trnV*. The 3′-end of the *rrnS* gene was determined through comparison between the mitochondrial orthologous genes described in this study and those available in GenBank (Thao *et al*., 2004; Tay *et al*., 2014; Wang *et al*., 2013).

Phylogenetically informative SNPs from sequenced RAD tags. Raw sequencing data were generated by Illumina base-calling software CASAVAv1.8.2 (http://support.illumina.com/sequencing/sequencing_software/casava.ilmn) according to the manufacturer's manual. Contaminated reads, such as those containing adaptors or primers, were identified by SeqPrep (https://github.com/jstjohn/SeqPrep) with parameters: '-q 20 -L 25 -B AGATCGGAAGAGCGTCGTGT -A AGATCGGAAGAGCACACGTC'. Sickle (https://github.com/najoshi/sickle) was applied to trim Illumina paired-end reads with default parameters. Sequence reads were decoded to identify the reads that belong to different samples according to the pooling signature for each sample. Clean data passing the above quality control processes were used in further analyses.

So far, among four published genomes (MEAM1, MED, SSA1-SG1 and Asia II 1) of *B. tabaci* species, MEAM1 has the best genome assembly compared to the other three genomes (Chen *et al*., 2016, 2019; Xie *et al*., 2017; Hussain *et al*., 2019). Thus, the high-quality sequencing reads were aligned to the MEAM1 genome sequences (the public data

of the MEAM1reference genome were obtained from NCBI, http://ncbi.nlm.nih.gov/Traces/wgs/?val=MAMS01#contigs) using BWA software (http://bio-bwa.sourceforge.net/). After removing PCR-duplication reads by SAMtools (http://samtools.sourceforge.net/) software with the command of 'rmdup', the sequencing depth and coverage were calculated based on the alignments by custom Perl scripts. The valid BAM file was used to detect SNPs by GATK 'Unified Genotyper' function (http://www.broadinstitute.org/gatk/). The SNPs with a quality value of more than 20 (GQ>20) and MQRankSum < 10 were considered to be validated. The SNPs that varied between individuals of the same member species were marked with degeneracy. Then all the SNP loci were sequentially concatenated by mapping against to the MEAM1 genome. Gap sequences were added to represent missing reads.

Sequence alignment. We carried out alignments of 13 PCGs, 22 tRNAs, and 2 rRNAs for the 27 specimen mitogenomes and four single-copy nuclear genes using MEGA 5.2.2 (Tamura *et al*., 2011). The mitogenome sequences of New World 1 (AY521259), MED (KU579279), *B. afer*_Africa (KF734668) and *B. afer*_China (GQ139515) were downloaded from GenBank. Alignments of single-copy nuclear genes were reconstructed with MAFFT 6.864 (L-INS-i option) (Katoh & Standley, 2013). Then the complete mitochondrial genome and four single-copy nuclear genes were concatenated respectively. MUSCLE (Edgar, 2004) was employed to perform the protein alignment. RNA sequences (rRNAs and tRNAs) were aligned using their secondary structure. The PCGs, tRNA and ribosomal RNA genes were aligned separately and then concatenated using Mesquite (Maddison & Maddison, 2001).

Missing data and pairwise analyses. For mitochondrial genomes, single-copy nuclear genes and RAD-Seq datasets (for RAD-Seq, minimum samples were specified as 10, percentages of missing data for the three datasets were calculated using a Perl script (available upon request). The heat maps of the nucleotide identities were drawn by R script (available upon request).

Phylogenetic analyses. PartitionFinder (Lanfear *et al*., 2012) was used to determine the best substitution model for the complete mitogenomes and the four single-copy nuclear genes. For the concatenated alignment of 13 PCGs, 22 tRNAs and two rRNAs, PartitionFinder selected a scheme with 12 partitions. Partitioning with exclusion of the third codon position resulted in 9 partitions, partitioning by mitochondrial genes resulted in 15 partitions and partitioning by four single-copy nuclear genes resulted in 3 partitions. For RAD-Seq datasets, 5 different matrices were proposed to assess whether SNP matrix size (i.e. length of the alignment of base pairs) influenced the outcome of phylogenetic inference (details are listed in *SI Appendix* 'Three molecular data resources'). Two major phylogenetic approaches, Bayesian and Maximum likelihood (ML) analyses were conducted. Bayesian analysis was conducted using MrBayes 3.2 (Ronquist *et al*., 2012) in combination with an exact model of molecular evolution generated by PartitionFinder. The algorithm MCMC was applied in parallel (four processors) on the IBM High Performance Computing Cluster (Bio-macromolecules Analysis Lab, Analysis Center of Agrobiology and Environmental Sciences, Zhejiang University). The analysis was run for 30 million to 60 million generations, with trees sampled and saved every 1,000

389 generations. All runs reached a plateau for likelihood score, which was indicated by the
390 standard deviation of split frequencies (0.0015). The potential scale reduction factor was
391 close to 1, indicating that convergence was achieved. Convergence of the runs was also
392 checked using Tracer v.1.6.0 (Rambaut *et al*., 2010) and the effective sample size (ESS)
393 values were well above 200 for each run. Maximum likelihood analyses on all three
394 datasets were performed in the MPI-parallelized RAxML 7.2.8-ALPHA (Stamatakis 2006)
395 using a fast bootstrapping algorithm (Stamatakis *et al*., 2008). For mitochondrial and
396 single-copy nuclear genes, bootstrap values were obtained using 1,000 replicates under
397 the GTRGAMMA + CAT approximation of the GTR + Γ model (Stamatakis 2006). For
398 RAD datasets (matrix: minimum sample= 6,8,10, 12 and 14), bootstrap values were
399 obtained using 1,000 replicates under the GTRCAT approximation (Stamatakis *et al*.,
400 2008) with the ascertainment bias correction added in operation (Leaché *et al*., 2015).
401 Figtree v1.4. 0 (Rambaut, 2012) was used to view and trim the topology.
402
403 BEAUti v1.8.2 (Drummond & Rambaut 2007) was used to generate the xml file for
404 BEAST runs. Four independent runs of BEAST were conducted, each consisting of two
405 chains resulting in 8 independent runs. The tree prior was set to 'Speciation: Birth-Death
406 Process' and the clock model was set to 'lognormal relaxed clock'. For each run, one
407 taxon set for the *Bemisia* species was defined and forced to be a monophyletic model and
408 specified as HKY. Base frequencies were estimated, and the site heterogeneity model was
409 set as Gamma plus invariant sites, with four gamma categories, partitioned into codon
410 positions (1 + 2), 3. The MCMC were run for 50 million generations and sampled every
411 1000 generations. Tracer v1.6.0 (Rambaut & Drummond, 2007) was used to check the

412 convergence of the multiple runs and the ESS values were well above 200 for each run.
413 Two independent BEAST runs were completed and the two tree files were combined
414 using Logcombiner (Rambaut & Drummond, 2015). TreeAnnotator (Rambaut &
415 Drummond, 2013) was used to generate a final tree, which was viewed in FigTree
416 v1.4.20 (Rambaut, 2012).
417
418 **Mating Crosses.** To give a global view of the crossing experiments conducted among the
419 different cryptic species, crossing data were (i) obtained from previously published
420 reports and (ii) generated for the current research project. In the current study, reciprocal
421 crosses were set up between two populations (MED_Sudan, MED). Control crosses
422 consisted of a virgin female and males from the same population with cotton plants (cv.
423 Zhe-Mian 1793) as the host plant. Four types of mating combinations between 10 females
424 and 10 males (MED♀ × MED♂, MED_Sudan♀ × MED_Sudan♂, MED♀ × MED_Sudan♂,
425 and MED_Sudan♀ × MED♂) were performed in home-made 'Lock and Lock' rearing units
426 containing cotton plants, with three replicates per combination. Rearing unit, plant
427 cultural solution and methods of collecting newly emerged virgin adults were as
428 described in Wang *et al*. (2011). For each combination, the newly emerged adults were
429 introduced into the rearing units and, 5 days later, all adults were collected and stored at
430 −20°C for subsequent RAPD-PCR confirmation of identity by diagnostic PCR detection
431 (De Barro & Driver, 1997). After another 30 days, the top cage of the rearing unit with
432 the first generation (F1) progeny whiteflies was placed in a freezer at −20°C for
433 subsequent counting and sexing.
434
435 Statistical analyses were performed using R statistical software (www.R-project.org). To

436 determine significant differences between the numbers of progeny as well as the female
437 ratios of progeny, one-way analysis of variance (ANOVA) was used. When a significant
438 difference (at $P < 0.05$) was detected among groups, the Least Significant Difference test
439 was used for multiple comparisons (Qin *et al*., 2016).
440
441 **Biological species assignment.** For the crossing experiments, if the fertility of the F1
442 female progeny produced in both intrapopulation crosses and inter-population crosses
443 was confirmed through carrying out reciprocal crossings with their male siblings
444 produced in interpopulation crosses, we took this as support that the two tested
445 populations represent the same species. However, if no female progeny produced by the
446 two tested populations, or even very few F1 female progeny were produced, but the F1
447 female did not have viability, the two tested populations would be assigned as separate
448 species.
449
450 **Geographic range and host datasets.** To further understand the role of geographic range
451 and host utilization in driving the diversification of the *B. tabaci* species complex,
452 information of the geographic distribution and the host plants of the cryptic species that
453 we studied were obtained in two ways: (i) relevant information compiled from published
454 literature associated with each specific species, as well as the relative collected location
455 and the host information deposited in GenBank that include the metadata mentioned
456 above; (ii) generated from the current study. The geographic distribution of each cryptic
457 species was summarized and marked onto the map. The host information was used as
458 input for the HEATMAP.2 function of the R “gplots” package (R Core Team, 2017).

Results

Dataset information. We carefully selected a set of 25 representative members of the *B. tabaci* cryptic species complex as test samples. Complete mitogenomes, four single-copy nuclear genes, and RAD-Seq were obtained for all the collected samples. The details of these datasets are described in *SI Appendix* ('Three molecular data resources', 'Strategy on handling RAD-Seq datasets' and 'General characterization of the three datasets', Figs. S2 and S3, Tables S3 and S4).

Molecular Phylogenetics. After getting datasets from whole mitogenomes, single-copy nuclear genes, and RAD-Seq, we generated phylogenetic trees (Figure 1) to screen the molecular relationships of the 25 selected members. The results were as follows: (i) The tree generated from full mitogenome sequences (37 genes, 14,599 bp total alignment) included 11 clades (Asia II, Australia, Asia I, China, Italy, Africa/Middle-East/Asia-Minor, New World, Unknown, sub-Saharan Africa, Uganda, Japan 2) and provided strong nodal support for most branches with different partition schemes and inferred methods (Fig. 1A and *SI Appendix*, Fig. S4A, Table S5); (ii) The single-copy nuclear-gene trees (5,325 bp) recovered many of the same clades as the mitogenomic analysis, but identified differences in backbone relationships, depending on the partition scheme (*SI Appendix*, Table S5) and inference methods used (Fig. 1B and *SI Appendix*, Fig. S4B). For instance, the Asia II species are consistently resolved with high nodal support into two separate groups. Some differences were attributable to non-significant nodal support in the single-copy nuclear-gene tree, but differences such as the position of

MED_Uganda SP had significant support which represented a genuine conflict between nuclear and mitogenomic datasets; (iii) In contrast, the RAD-Seq datasets produced tree topologies with high support regardless of inference method and the SNP matrix used (*SI Appendix*, Tables S3–S6, Fig. S5). The RAD-Seq trees recovered the same major clades as the other two analyses, but the topology was more like the trees for nuclear genes than those for mitogenomes (Fig. 1C, *SI Appendix*, Fig. S5; SNP distributions around the MEAM1 genome are shown in Fig. 1D). Both nuclear topologies recovered two major clades with one consisting of New World 1 and New World 2 and the other one consisting of Asia II 3 and Asia II 9. In summary, discrepancies can be seen among the phylogenetic trees built by the three datasets.

Reproductive compatibility among putative species. For further validating phylogenetic results, we collected data of crossing experiments from the literature and performed reciprocal crossing experiments among putative species, and compared these results with the molecular phylogenetic species. As a test case, the data of reciprocal-crossing experiments associated with the 16 *mtCOI*-defined species defined in previous studies (Liu *et al*., 2012; Qin *et al*., 2016), together with one additional MED population called MED_Sudan were analyzed. The crossing experiments showed a range of outcomes, from completely compatible, to an intermediate level of compatibility, to completely incompatible (Fig. 2). The intermediate levels showed a range of partially incompatible phenotypes, ranging from low numbers of F_1 hybrid females to successful mating in only one direction, and/or F_1 progeny with reduced viability and fertility.17 species in 25 samples used here were identified (using data of 37 reciprocal crossings among 17 putative species) that are largely agreed with the number of species proposed

by 3.5% *mtCOI* barcode criteria. And the reproductive incompatibilities co-varied with the presence of major phylogenetic clades, although limited incongruence between the two approaches was also evident (Fig. 2). The Asia II 3 and Asia II 9 groups from China, for instance, had a partial *mtCOI* divergence of 5.26% (*SI Appendix*, Table S2), indicating that they were two *mtCOI*-defined species, but showed near complete reproductive compatibility in one direction (Asia II 3 male × Asia II 9 female) and partial reproductive compatibility in the other direction (Qin *et al*., 2016) (Fig. 2). The other notable exception was between MED and MED_Sudan, which showed *mtCOI* divergence of only 0.31% and thus would be considered the same *mtCOI* defined species, but showed complete incompatibility without the production of F_1 hybrids in either direction (*SI Appendix*, Table S6). Interestingly, with reference to the nuclear topologies, the putative species 'Asia II 3' and 'Asia II 9' seem genetically indistinguishable in the RAD-Seq topology (despite being divergent in the mitochondrial data), while the putative species 'MED' and 'MED_Sudan' seem genetically indistinguishable in single-copy nuclear topology. This suggests a combination of molecular phylogenies and reciprocal crossing data can achieve a better classification than the molecular experiments analyses alone.

Geographic distribution. The geographic information of each *B. tabici* species covered within this manuscript were analyzed based on the metadata deposited in GenBank (Fig. 3). The result showed these species mainly occurred around their local geographic regions; for example, the species belonging to the China clade was distributed only in China, while Japan2 was restricted to Japan and Korea, and two species in the New World clade were mostly found in the Americas with New World 1 mostly occurring in

North America, while New World 2 being restricted to South America. However, there were four exceptions to the geographic separation of species: (i) the highly invasive MEAM1 and MED species have a wide and overlapping geographic range, except that MED has not yet been recorded in Australia; (ii) the Australian species classified in the Australia clade were also found in Indonesia; (iii) New World 1 in the New World clade was found in Sudan; and finally, (iv) the Sub-Saharan Africa 2 was found in Europe (Fig. 3). By mapping this information to the reproductive compatibility data, we found that the reproductive barriers between the *mtCOI* defined species, MED and MED_Sudan might contribute to the geographic isolation. Because the MED sample was collected from a population in east China, which is believed to have been introduced from the Mediterranean region approximately 15 years ago, and the latter was collected from Sudan, in East Africa. The results further suggest that a combination of geographical and reciprocal crossing data provide a complementary assessment of species recognition.

Host plant associated speciation. Host-associated differentiation plays an important role in speciation of various phytophagous insects (Berlocher & Feder, 2002; Stireman, *et al.*, 2005). By conducting a literature survey, we found that plants belonging to 30 families have been documented as potential hosts of the *B. tabaci* cryptic species complex. Of the 30 families, nine families comprise crop plants that could be utilized by many of the species, while the remaining hosts were mainly wild plants that could be utilized by only a few species. A heatmap (Fig. 4) showed that Euphorbiaceae, Malvaceae, and Solanaceae are the top three host families utilized. Apart from these three families, Convolvulaceae, Asteraceae, Fabaceae and Cucurbitaceae were identified as commonly

shared by most *B. tabaci* species. Clustering analysis clustered the *B. tabaci* species into three groups based on their host range. One group comprised by Sub-Saharan Africa 1, MED, MEAM1, Asia I and Asia II 1 species, and in particular for MED and MEAM1, infested the highest number of hosts including most crop plants, illustrating their economic importance (Fig. 4). Most species fed on more than one family of host plants, except for Australia_E, Asia II 9 and Italy 3. Overall, species in the complex differ in their host range, with many of them showing a potential to use at least seven plant families. However, these datasets could not clearly distinguish among species, indicating that host utilization datasets available so far are not a good indicator of species.

Lessons from comparison of different datasets with reproductive compatibility.

Among all the available datasets, reproductive incompatibility is a more direct and conclusive approach (Paterson *et al.*, 2016) to confirm true cryptic species. To further learn the role of the various datasets in aiding *B. tabaci* species classification, we thus compared the other datasets to reproductive compatibility. We found that host range datasets could be ruled out, as many of *B. tabaci* species share overlapped hosts. In addition, phylogenetic topologies constructed by mitochondrial genome could improve our knowledge on the relationships of different cryptic species, but show similar topology as *mtCOI* sequences topology (Boykin *et al.*, 2007). Overall, by mapping the other data sources to the reproductive incompatibility, we found that single-copy nuclear gene or RAD-Seq phylogenetic topology, *mtCOI* barcode criteria, and geographical distribution could be used for explaining the reproductive compatibility or mating barriers observed most cases. Specifically, what we learned could be summarized as follows (Fig. 5): (i)

similar with previous reports, the *mtCOI* barcode criteria are useful, but in some cases could not recover the same species groups as the mating approach such as in the case of Asia II and MED groups of species. (ii) geographically, different distribution pattern yields different level of reproductive compatibility. For example, when the species are sympatric, complicated situations may arise (Fig. 5A): when the *mtCOI* divergence is higher than 5.26%, complete reproductive incompatibility nearly always occurs, however, when the *mtCOI* divergence is lower than 3.5%, either complete reproductive incompatibility or intermediate level of compatibility may occur. When species are allopatric, even when the *mtCOI* divergence is lower than 3.5%, complete reproductive incompatibility may occur. Allopatric, for instance, could be partly used to explain the mating barrier between MED and MED_Sudan but is in conflict with the previously proposed *mtCOI* barcode criteria in species recognition (Fig. 5B); (iii) nuclear genes/SNPs based phylogeny provides additional determinants for recognizing relationships between different species and explaining the results of crossing experiments, such as cases reported in MED (Vyskočilová *et al*., 2018), and Sub-Saharan Africa 1 (Mugerwa *et al*., 2021) groups of species. Overall, by mapping the other data sources to the reproductive incompatibility, we found that *mtCOI* barcode criteria and single-copy nuclear gene or RAD-Seq phylogenetic provide the first useful separation of species, with the data of geographical distribution to aid in further analysis. We, thus, promote that the integration of these data sources has the potential for accurate determination of species status within the *B. tabaci* whitefly complex. Finally, based on the findings from this study as well as previously accumulated knowledge we propose a comprehensive framework for guiding classification of cryptic species of this whitefly complex (Fig 5C).

Discussion

Species, as the basic units in evolution and biodiversity, provides a direct link to the knowledge about an organism (Schlick-Steiner, 2010). Currently, a range of data sources has been used for delimiting different *B. tabaci* cryptic species, including the 3.5% sequence divergence threshold of the *mtCOI* (Dinsdale *et al*., 2010), reproductive compatibility (Liu *et al*., 2012), and machine learning based on differences in puparium morphology (Macleod *et al*., 2022). However, a lack of systematic studies offers little guidance on how congruence or incongruence among different datasets. To this end, we evaluated the power of datasets that have been used in distinguishing different species and the results showed the inconsistency in the operational designation of "cryptic species" by each of these single datasets. Then, we applied comparative analysis of different datasets against with the reproductive compatibility datasets to see how many and which data sources should be integrated to delimit species most accurately in the *B. tabaci* species complex. Based on our results, apart from reproductive compatibility, we propose *mtCOI* divergence, nuclear markers, native geographic range (at least inclusion of collecting location) as assistant potential diagnostic markers for recognition of new *B. tabaci* species.

In comparison to the phylogenetic relationships derived from *mtCOI* data (Boykin *et al*., 2007; Dinsdale *et al*., 2010; Kanakala & Ghanim, 2019), our results provide a more accurate view of the diversity and molecular phylogenetic relationships among members of the *B. tabaci* species complex using dense molecular markers combining mitogenome,

single copy nucleotide genes and RAD-seq with appropriate worldwide sampling. RAD-seq has been used to the study of a wide range of organisms, including *B. tabaci* from the Subsaharan African group (Cariou *et al*., 2013; Rubin *et al*., 2012; Wosula *et al*., 2017, Wosula *et al*., 2017, Elfekih *et al*., 2021). The method is highly recommended when there is a reliable reference genome, which decreases errors in the assessment of orthologous RAD loci (Wagner *et al*., 2013). In this study, 729,953–2,550,155 bp of SNP information with high corresponding average depth of polymorphic loci was used for generation of a phylogenetic tree. This resulted in a phylogeny with improved resolution and high bootstrap support that has not been reported previously. The associated RAD sequencing datasets provide a reference model for further application of this technology to other cryptic pest species (*SI Appendix*, Implications from RAD-Seq). The resulting phylogeny is consistent with the topology inferred by single-copy nuclear genes, although it is incongruent with the topology inferred from mitochondrial genomes. Overall, our results agree with those of Mugerwa *et al*. (2021), who found that genome-wide SNPs provide a more accurate representation of *B. tabaci* species diversity within the Sub-Saharan African populations compared with inferences based on mitochondrial genomes (Dinsdale *et al*., 2010).

Although our results provide a more accurate molecular phylogenetic relationship of members within the *B. tabaci* species complex, these very rich datasets can also be misleading (De Queiroz, 2007). To truly confirm a cryptic species, the biological species concept, which posits that new species are formed when they are reproductively isolated (Paterson *et al*., 2016), can be applied. By investigating the reciprocal-crossing

644 experiments associated with 16 putative species defined by *mtCO1*, complete
645 reproductive incompatibility appeared in most studied species pairs that were largely co-
646 varied with the presence of major lineages in phylogenetic trees. These results were not
647 unexpected as most pairs used for crossing experiments had more distant relationships
648 along lineages. In the process of species divergence, the farther along a lineage a species
649 resides, a larger number of differences expected to have evolved compared to the species
650 pairs that have closer relationships (De Queiroz, 2007). Two exceptions arose in the MED
651 and Asia II group -MED and MED_Sudan; Asia II 3 and Asia II9 -with the former one
652 designated as the same *mtCOI*-defined species showing near complete incompatibility
653 without producing hybrid offspring in either direction of crosses and the latter one
654 designated as two *mtCOI*-defined species showing partial reproductive compatibility in
655 one direction. Consistent with the results of Vyskočilová *et al*. (2018), who demonstrated
656 the existence of at least two distinct species within MED species, we also found that
657 MED and MED_Sudan previously designated as the same species are indeed two
658 separate biological species.
659
660 The current *B. tabaci* species distribution was consistent with a mode of speciation based
661 on geographic separation, as different species were associated with particular continental
662 regions (Boykin *et al*., 2013; De Barro *et al*., 2005). In line with that, we found that most
663 species were confined to local geographic regions, indicating the allopatric divergence
664 mechanisms likely contribute to speciation in the *B. tabaci* complex. Substantial species
665 diversity was found in China and Uganda (Mugerwa *et al*., 2018), highlighting the
666 evolutionary importance of these regions. However, these results might also be

attributable to uneven sampling efforts by researchers across the globe. Four species were found to have a wider distribution range than originally reported, suggesting that current species distribution ranges may have changed over time (Berlocher *et al*., 2000; Hofreiter *et al*., 2009). MEAM1 and MED had a wide geographic range, as expected, because they are highly invasive and are the world's most destructive crop pests (Tay *et al*., 2017; De Barro *et al*., 2011). Interestingly, an Australian species was also found in Indonesia, a New World 1 species in Sudan, and a SSA2 species was also reported in Europe, supporting a new hypotheses of *B. tabaci* speciation that is linked to the original large Gondwana landmass (Cawood & Buchan, 2007). This result agrees with the discovery of Mugerwa *et al*. (2018), who reported two new species named sub-Saharan Africa 10 and 11 that clustered together with the New World species also found in Africa. These and our results provide new evidence for a close evolutionary linkage between the Old and New World species. Consequently, it seems likely other unidentified geographic ranges of certain members exist within the *B. tabaci* complex.

Patterns of host utilization could serve as a major factor in the diversification of herbivorous insects (Hamm and Fordyce 2015). De Barro (2005) proposed that the role of the host in divergence in the complex is unlikely. However, Malka *et al.* (2018) reported species in the complex that differ in their host range and found that the 10 *B. tabaci* populations could be divided into four groups based on their host association. In accordance with the findings of Malka *et al.* (2018), the 25 *B. tabaci* populations here could be roughly clustered into three groups on the basis of the host range: (i) one group consisting of Sub-Saharan Africa 1, MED, MEAM1, Asia I and Asia II 1 species with the

690 widest host range, (ii) the second group having an extended host range, and (iii) the third
691 group having a more or less restricted host range. Since many specimens clustered
692 together in this analysis, our results indicate that host plant ranges might not be good for
693 distinguishing the individual *B. tabaci* species. Anyhow, we do not feel confident of these
694 results, as over 600 plants have been reported as hosts of *B. tabaci* (Mound and Halsey
695 1978), and only very limited information could be extracted from metadata. Notably,
696 some host plants were reported only once for some species, and thus might only serve as
697 temporary hosts for *B. tabaci*, and not as a 'documented' host used as a food resource.
698 Even for a single cryptic species, MED for example, different mitochondrial groups had
699 differing but overlapping host plant ranges (Vyskočilová *et al*., 2019). Much more effort
700 is required in the realm of field sampling and experimental host adaption to further
701 identify the role of host plants in the speciation of *B. tabaci* (Funk 2012; Malka *et al*.,
702 2018).
703
704 Since reproductive incompatibility is generally considered a more direct and conclusive
705 approach (Paterson *et al*., 2016) to confirm true cryptic species, we further compared
706 different datasets against reproductive compatibility to investigate the congruence or
707 incongruence in species delimiting. The results showed that the combination of *mtCOI*
708 divergence, nuclear markers, native geographic range (at least inclusion of collecting
709 location) provide a complementary assessment of species recognition. For example,
710 complete reproductive incompatibility appeared in most studied species pairs that were
711 largely co-varied with the *mtCOI*-defined species, as well as the presence of major
712 lineages in phylogenetic trees. The RAD-Seq or single-copy nuclear phylogenetic trees

showed the genetic similarity or difference of the species in the two expectations raised in reproductive incompatibility/compatibility. These results provide further support for the hypothesis that a combination of phylogenies derived from nuclear datasets and reciprocal crossing data can achieve a better classification than molecular data alone and further highlighting that nuclear markers are a good complementary data source for *B. tabaci* cryptic species identification or any re-examination of species diversity within the *B. tabaci* group (Mugerwa *et al*., 2021; Ally *et al*., 2019).

A lack of mating barriers between 'Asia II 3' and 'Asia II 9' might attribute to the lineage still in the early stages of divergence that has not yet evolved the properties to clearly distinguish them through reproductive isolation (De Queiroz, 2007). However, this interpretation does not fit for why mating barriers occurred between 'MED' and 'MED_Sudan', even though they share closer relationship than that between 'Asia II 3' and 'Asia II 9', and alternative factor should be searched for illustrating the mating barriers between them. In addition, these findings, while preliminary, suggest that a series of crosses among closely related populations should be conducted, and only complete reproductive incompatibility could be severed as a point of reference to improve the inference about species delimitation. In addition, reproductive barriers observed between the MED and MED_Sudan could have been caused by limited gene flow (Irwin, 2002), due to the large geographic distance, supporting that the proposition of allopatric speciation in the evolution of *B. tabaci* cryptic species. Two *mtCOI*-defined species, Asia II 3 and Asia II 9, demonstrated reproductive incompatibility only in one direction of reciprocal crosses, which might attributable to similar geographic distributions. If this is

the case, it would mean that sympatric speciation may be occurring in the evolution of these *B. tabaci* cryptic species. Specifically, invasive species in the complex, may be able to mate with diverse incompatibility gradients. Invasive MEAM1 (Perring 2000; Cheek & Macdonald, 1994) showed the potential to mate with resident species China 1 and Asia II 3, suggesting that competition may be occurring between the invasive and the resident species. Consistently, Taquet *et al*. (2022) found around 2% hybrids between MEAM1 and Indian Ocean species using 11 nuclear microsatellite loci, indicating the presence of incomplete reproductive isolation between the two species. Our study corroborates previous findings (Manel *et al*., 2003; Manni *et al*., 2004) that geographic information is key to distinguish differentiation that appears within species because of isolation by distance.

Collectively, we found that different data sources we used performed differently in identifying distinct *B. tabaci* species. Our results suggest the following insights regarding the study of species relationships within the *B. tabaci* species complex: (i) Because the *mtCOI* and full mitogenome sequencing data performed very similarly with regard to the number of delimited species, for cost-effectiveness, universal and widely used *mtCOI* barcodes can be used for preliminary species delineation; (ii) nuclear phylogenies performed better for discovery of hidden genetic divergence (Elfekih *et al*., 2021; Mugerwa *et al*., 2021), which is definitely needed before a new species can be accepted; (iii) reproductive compatibility is a reliable criterion and quality check for species recognition as new species are formed when they are reproductively isolated (Paterson *et al*., 2016); (iv) geographical position provides an additional dimension for species

identification (Bolnick and Fitzpatrick, 2007; Coyne & Orr, 2004) when the species has a narrow and separate distribution range. In fact, the combination of mating information and geographic information can lead to reliable species recognition results; (v) host range is a poor choice as a criterion for delimiting *B. tabaci* species because the association between host plants and diversification of *B. tabaci* is unclear (De Barro, 2005; De Barro *et al*., 2005). In conclusion, in practical terms, recognition of a new *B. tabaci* species requires information on *mtCOI*, nuclear markers, geographical location, and reproductive compatibility. Due to the fact that separate species might respond to control measures differently, the current lack of precision in making routine species-level identifications in many insect pests has limited the effectiveness of control programs (Packer *et al*., 2018). Our results suggest that an integrated approach based on multiple data types can more confidently delimit *B. tabaci* species and provide a better overall taxonomic resolution of the entire species complex. Given the considerable damage that *B. tabaci* infestations inflict on agricultural crops, our study paves the way for the improved identification and diagnosis, as well as prevention of the spread of disease caused by whitefly-transmitted begomoviruses and implementation of biological control (Bikford *et al.*, 2006).

Since cryptic species complexes are being documented at an ever-increasing rate on most major branches of the animal tree of life (Bickford *et al*., 2007; Wosula *et al*., 2017; Jörger and Schrödl 2013; Pante *et al*., 2015; Loxdale *et al*., 2016; Nygren 2014), lessons derived from our present study will help guide the assessment of species complexes of other similar pests, especially those with the same evolutionary patterns as the *B. tabaci* species complex. The suggestions for insects' cryptic species identification include the

following: (i) The top priority during the initial stage is that the sampling of species or populations of interest should cover all lineages; (ii) effort must be given for specific *mtCOI* barcode development for species of interest that damage economically important crops, so as to increase the accuracy of species delimitation; (iii) reproductive compatibility should be at least investigated in representative populations and tested against *mtCOI* barcode species boundaries to determine reasonable species delimitation criteria; (iv) comprehensive nuclear genes are good targets for accurately predicting species diversity and evolutionary history; and (v) finally, further characterization of host range, geographic range, behavior, and other possible factors that drive speciation could be assessed to assist in delimiting species of interest.

Author Contributions

Hua-Ling Wang, Teng Lei, Xiao-Wei Wang and Stephen Cameron, John Colvin and Shu-Sheng Liu conceived and designed the experiments; Hua-Ling Wang analyzed the data, Hua-Ling Wang and Kathryn Bushley wrote the manuscript, and all authors commented on the manuscript; John Colvin andmShu-Sheng Liu, Jesús Navas-Castillo, Yin-Quan Liu, M. N. Maruthi, Christopher A. Omongo, Hélène Delatte, Kyeong-Yeoll Lee, Renate Krause-Sakate, James Ng, Susan Seal and Elvira Fiallo-Olivé collected experimental data. All authors read and approved the manuscript.

Acknowledgments

We thank Tao Ye, Shanghai Biozeron Biotechnology Co. Ltd., and Laura Boykin, University of Western Australia, for providing sequencing platform and bioinformatics

804 support. Kevin Gorman, Rothamsted Research UK, for providing whitefly specimens of
805 Mediterranean_Sudan. We acknowledge the support of Jodie Wetherall, Douglas Barry
806 for giving access to the High Performance Computing facility at the University of
807 Greenwich. We acknowledge the support of San-Ling Wu for giving access to the IBM
808 High Performance Computing Cluster of Bio-macromolecules Analysis Lab, Analysis
809 Center of Agrobiology and Environmental Sciences, Zhejiang University. We also thank
810 Peter Sseruwagi and Joseph Ndunguru, Mikocheni Agriculture Research Institute,
811 Tanzania, for collecting samples from Uganda. J.N.C. is supported by grants (AGL2013-
812 48913-C2-1-R and AGL2016-75819-C2-2-R) from the Ministerio de Economía y
813 Competitividad (MINECO, Spain), co-financed by ERDF. E.F.O. is a recipient of a 'Juan
814 de la Cierva-Incorporación' postdoctoral contract from MINECO.

815 **Funding**

816 This work was supported by the National Natural Science Foundation of China (Grant
817 numbers 31501878 and 31272104), China Agriculture Research System (CARS-23-D07)
818 and the Bill & Melinda Gates Foundation (African cassava whitefly project,
819 OPP1058938).

820 **Data availability**

821 The mitogenomes and single-copy nuclear genes sequences of this study were deposited
822 in GenBank. The raw sequence data of RAD-Seq were deposited in the National Center
823 for Biotechnology Information (NCBI) Sequence-Read Archive (SRA) database. The
824 Accession numbers are listed in *SI Appendix*, Table S1.

825 **Declarations**

826 **Conflict of Interest** The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication Not applicable.

References

Alicai T, Omongo CA, Maruthi MN, Hillocks RJ, Baguma Y, Kawuki R, *et al*. (2007) Re-emergence of cassava brown streak disease in Uganda. *Plant Disease*, 91: 24-9.

Ally HM, Hamss HE, Simiand C, Maruthi MN, Colvin J, Omongo CA, *et al*. (2019) What has changed in the outbreaking populations of the severe crop pest whitefly species in cassava in two decades? *Scientific Reports*, 9: 14796.

Amari K, Gonzalez-Ibeas D, Gómez P, Sempere RN, Sanchez-Pina MA, Aranda MA, *et al*. (2008) Tomato torrado virus is transmitted by *Bemisia tabaci* and infects pepper and eggplant in addition to tomato. *Plant Disease*, 92: 1139.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, *et al*. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Public Library of Science*, 3: e3376.

Berlocher SH and Feder JL. (2002) Sympatric speciation in phytophagous insects: Moving beyond controversy? *Annual Review of Entomology*, 47: 773-815.

Bolnick DI and Fitzpatrick BM (2007) Sympatric speciation: models and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, 38: 459-487.

Boykin LM, Armstrong KF, Kubatko L and De Barro P. (2012) Species delimitation and global biosecurity. *Evolutionary Bioinformatics*, 8: 1-37.

Boykin LM, Bell CD, Evans G, Small I, and De BarroPJ. (2013) Is agriculture driving the diversification of the *Bemisia tabaci* species complex (Hemiptera: Sternorrhyncha: Aleyrodidae)?: Dating, diversification and biogeographic evidence revealed. *BMC Evolutionary Biology*, 13: 228.

Boykin LM. (2014) *Bemisia tabaci* nomenclature: lessons learned. *Pest Management Science*, 70:1454-1459.

Boykin LM, Shatters Jr RG, Rosell RC, Mckenzie CL, Bagnall RA, De Barro P, *et al*. (2007) Global relationships of *Bemisia tabaci* (Hemiptera: Aleyrodidae) revealed using Bayesian analysis of mitochondrial COI DNA sequences. *Molecular Phylogenetics and Evolution*, 44: 1306-1319.

Brown JK, Frohlich D and Rosell R (1995) The sweetpotato or silverleaf whiteflies: biotypes of *Bemisia tabaci* or a species complex? *Annual Review of Entomology*, 40: 511-534.

Cao CC and Sun X. (2016) Combinatorial pooled sequencing: experiment design and decoding. *Quantitative Biology*, 4: 36-46.

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

862 Cawood PA and Buchan C. (2007) Linking accretionary orogenesis with supercontinent assembly. *Earth-*
863 *Science Reviews*, 82: 217-256.
864 Cheek S and Macdonald O. (1994) Statutory controls to prevent the establishment of *Bemisia tabaci* in the
865 United Kingdom. *Pesticide Science*, 42: 135-137.
866 Chen W, Hasegawa DK, Kaur N, Kliot A, Pinheiro PV, Luan JB, *et al.*(2016) The draft genome of whitefly
867 *Bemisia tabaci* MEAM1, a global crop pest, provides novel insightsinto virus transmission, host
868 adaptation, and insecticide resistance. *BMC Biology*, 14(1): 110.
869 Chen W, Wosula EN, Hasegawa DK, Casinga C, Shirima RR, Fiaboe KKM, *et al*. (2019) Genome of the
870 African cassava whitefly *Bemisia tabaci* and distribution and genetic diversity of cassava-colonizing
871 whiteflies in Africa. *Insect Biochemistry and Molecular Biology*, 110: 112–120.
872 Coyne JA and Orr HA. (2004) Speciation. Sinauer Associates, Sunderland, MA.
873 Crowder DW, Horowitz AR, De Barro PJ, Liu SS, Showalter AM, Kontsedalov S, *et al*. (2010) Mating
874 behaviour, life history and adaptation to insecticides determine species exclusion between whiteflies.
875 *Journal of Animal Ecology*, 79: 563-570.
876 Darwin C. (1859) The origin of species by means of natural selection: or, the preservation of favored races
877 in the struggle for life. London, UK: John Murray.
878 Dayrat B. (2005) Towards integrative taxonomy. *Biological Journal of the Linnean Society*, 85: 407-415.
879 De Barro PJ and Driver F. (1997) Use of RAPD PCR to distinguish the B biotype from other biotypes of
880 *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae). *Australian Journal of Entomology*, 36: 149-
881 152.
882 De Barro PJ, Driver F, Trueman JWH and Curran J. (2000) Phylogenetic relationships of world populations
883 of *Bemisia tabaci* (Gennadius) using ribosomal ITS1. *Molecular Phylogenetics and Evolution*, 16: 29-
884 36.
885 De Barro PJ, Liu SS, Boykin LM and Dinsdale AB. (2011) *Bemisia tabaci*: a statement of species status.
886 *Annual Review of Entomology*, 56: 1-19.
887 De Queiroz K. (2005) A unified concept of species and its consequences for the future of taxonomy.
888 *Proceedings of the California Academy of Sciences*, 56: 196-215.
889 De Queiroz K. (2007) Species concepts and species delimitation. *Systematic Biology*, 56:879-86.
890 Delatte H, Reynaud B, Granier M, Thornary L, Lett JM, Goldbach R and Peterschmitt M. (2005) A new
891 silverleaf-inducing biotype Ms of *Bemisia tabaci* (Hemiptera: Aleyrodidae) indigenous to the islands
892 of the south-west Indian Ocean. *Bulletin of Entomological Research*, 95: 29-35.
893 Dinsdale A, Cook L, Riginos C, Buckley Y and De Barro PD. (2010) Refined global analysis of *Bemisia*
894 *tabaci* (Hemiptera: Sternorrhyncha: Aleyrodoidea: Aleyrodidae) mitochondrial cytochrome oxidase 1
895 to identify species level genetic boundaries. *Annals of the Entomological Society of America*, 103:
896 196-208.
897 Doyle JJ and Luckow MS. (2003) The rest of the iceberg: Legume diversity and evolution in a
898 phylogenetic context. *Plant Physiology*, 131: 900-910.

899 Drummond AJ and Rambaut A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC*
900 *Evolutionary Biology*, 7: 214-214.
901 Edgar RC. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput.
902 *Nucleic Acids Research*, 32: 1792-1797.
903 Egea E, David B, Choné T, Laurin B, FéralJP and Chenuil A. (2016) Morphological and genetic analyses
904 reveal a cryptic species complex in the echinoid *Echinocardium cordatum* and rule out a stabilizing
905 selection explanation. *Molecular Phylogenetics and Evolution*, 94: 207-220.
906 Elfekih S, TayW, Polaszek A, Gordon K, Kunz D, Macfadyen S, *et al*. (2021) On species delimitation,
907 hybridization and population structure of cassava whitefly in Africa. *Scientific Reports*, 11: 1-11.
908 Firdaus S, Vosman B, Hidayati N, Supena J, Darmo E, Visser RG, *et al*. (2013) The *Bemisia tabaci* species
909 complex: additions from different parts of the world. *Insect Science*, 20: 723-733.
910 Frézal L and Leblois R. (2008) Four years of DNA barcoding: current advances and prospects. *Infection,*
911 *Genetics and Evolution*, 8: 727-736.
912 Funk DJ. (2012) Of "host forms" and host races: terminological issues in ecological speciation. *Acta*
913 *Oecologica*, 2012: 506957.
914 Gómez A,Serra M,Carvalho GR and Lunt DH. (2002) Speciation in ancient cryptic species complexes:
915 evidence from the molecular phylogeny of *Brachionus plicatilis* (Rotifera). *Evolution*, 56:1431-1444.
916 Hammer Øyvind, Harper D and Ryan P (2001) PAST–Palaeontological statistics, ver. 1.89. *Palaeontologia*
917 *Electronica*, 4.
918 Frohlich DR, Torres-Jerez I, Bedford ID, MarkhamPG and Brown JK. (1999) A phylogeographical analysis
919 of the *Bemisia tabaci* species complex based on Mitochondr DNA markers. *Molecular Ecology*, 8:
920 1683-1691.
921 Gill RJ and Brown JK. (2009) Systematics of *Bemisia* and *Bemisia* relatives: can molecular techniques
922 solve the *Bemisia tabaci* complex conundrum-a taxonomist's viewpoint. In "*Bemisia: Bionomics and*
923 *Management of a global pest*", eds Stabaly PA & Naranjo SE, (Springer), pp, 5-29.
924 Ghosh S, Kanakala S, Lebedev G, Kontsedalov S, Silverman D, Alon T, *et al*. (2019) Transmission of a
925 new polerovirus infecting pepper by the whitefly *Bemisia tabaci*. *Journal of Virology*, 93: e00488-19.
926 Habibu M, Wang HL, Peter S, Susan S and Colvin J. (2021) Whole-genome single nucleotide
927 polymorphism and mating compatibility studies reveal the presence of distinct species in sub-Saharan
928 Africa *Bemisia tabaci* whiteflies. *Insect Science*, 28:1553-1566.
929 Hahn SK and Janet K. (1985) Cassava: a basic food of Africa. *Outlook on Agriculture*, 14: 95-9.
930 Hall AE, Fiebig A and Preuss D. (2002) Beyond the Arabidopsis genome: opportunities for comparative
931 genomics. *Plant Physiology*, 129:1439-1447.
932 Hamm CA and Fordyce JA. (2015) Patterns of host plant utilization and diversification in the brush-footed
933 butterflies. *Evolution*, 69:589-601.

Heraty JM, Woolley JB, Hopper KR, Hawks DL, Kim JW and Buffington M. (2007) Molecular phylogenetics and reproductive incompatibility in a complex of cryptic species of aphid parasitoids. *Molecular Phylogenetics and Evolution*, 45: 480-493.

Hillis DM, Bull JJ. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42:182-92.

Holsinger KE,Weir BS. (2009) Genetics in geographically structure populations: defining, estimating and interpreting F (ST). *Nature Reviews Genetics*, 10: 639-65.

Horowitz AR, Kontsedalov S and Ishaaya I. (2004) Dynamics of resistance to the neonicotinoids acetamiprid and thiamethoxam in *Bemisia tabaci* (Homoptera: Aleyrodidae). *Journal of Economic Entomology*, 97: 2051-2056.

Hsieh CH, Ko CC, Chung CH and Wang HY. (2014) Multilocus approach to clarify species status and the divergence history of the *Bemisia tabaci* (Hemiptera: Aleyrodidae) species complex. *Molecular Phylogenetics and Evolution*, 76: 172-180.

Huelsenbeck JP and Rannala B. (2004) Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*, 53:904-13.

Hussain S, Farooq M, Malik HJ, Amin I, Scheffler BE, Scheffler JA, *et al.* (2019) Whole genome sequencing of Asia II 1 species of whitefly reveals that genes involved invirus transmission and insecticide resistance have genetic variances between Asia II 1 and MEAM1 species. *BMC Genomic*, 20: 507.

Irwin DE. (2002) Phylogeographic breaks without geographic barriers to gene flow. *Evolution*, 56: 2383-2394.

Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, *et al*. (2016) HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, 4: 1600016.

Jörger KM and Schrödl M. (2013) How to describe a cryptic species? Practical challenges of molecular taxonomy. *Frontiers in Zoology*, 10: 59.

Kanakala S and Ghanim M. (2019) Global genetic diversity and geographical distribution of B*emisia tabaci* and its bacterial endosymbionts. *PLoS One*, 14(3):e0213946.

Katoh K and Standley DM. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30: 772-780.

Lanfear R, Calcott B, Ho SYW and Guindon S. (2012) Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, 29: 1695-1701.

Larsen K. (2001) Morphological and molecular investigation of polymorphism and cryptic species in tanaid crustaceans: implications for tanaid systematics and biodiversity estimates. *Zoological Journal of the Linnean Society*, 131: 353-379.

970 Leaché AD, Banbury BL, Felsenstein J, De Oca ANM and Stamatakis A. (2015) Short tree, long tree, right
971 tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology*,
972 64: 1032-1047.
973 Legg JP, Owor B, Sseruwagi P and Ndunguru J. (2006) Cassava mosaic virus disease in East and Central
974 Africa: epidemiology and management of a regional pandemic. *Virus Research*, 67: 355-418.
975 Legg JP, Shirima R, Tajebe LS, Guastella D, Boniface S, Jeremiah S, *et al*. (2014) Biology and
976 management of *Bemisia* whitefly vectors of cassava virus pandemics in Africa. *Pest Management*
977 *Science*, 70:1446-53.
978 Liu SS, Colvin J and De Barro PJ. (2012) Species concepts as applied to the whitefly *Bemisia tabaci*
979 systematics: how many species are there? *Journal of Integrative Agriculture*, 11: 176-186.
980 Liu SS, De Barro PJ, Xu J, Luan JB, Zang LS, Ruan YM, *et al*. (2007) Asymmetric mating interactions
981 drive widespread invasion and displacement in a whitefly. *Science*, 318: 1769-1772.
982 Lowe S, Browne M and Boudjelas S. (2004) 100 of the world's worst invasive alien species. *A section from*
983 *the global invasive species database.* Invasive Species Sepcialist Group, Auckland, New Zealand.
984 Lowe TM and Eddy SR. (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes
985 in genomic sequence. *Nucleic Acids Research*, 25: 955-964.
986 Loxdale HD, Davis BJ and Davis RA. (2016) Known knowns and unknowns in biology. *Biological Journal*
987 *of the Linnean Society*, 117: 386-398.
988 Macleod N, Canty RJ and Polaszek A. (2022) Morphology-based identification of *Bemisia tabaci* cryptic
989 species puparia via embedded group-contrast convolution neural network analysis. *Systematic Biology*,
990 71(5): 1095-1109.
991 Maddison WP and Maddison DR. (2001) Mesquite: a modular system for evolutionary. Available at
992 analysis.
993 Malka O, Santos-Garcia D, Feldmesser E, Sharon E, Krause-Sakate R, Delatte H, *et al*. (2018) Species-
994 complex diversification and host-plant associations in *Bemisia tabaci*: a plant-defense, detoxification
995 perspective revealed by RNAseq analyses. *Molecular Ecology*, 27: 4241-4256.
996 Mallet J. (1995) A species definition for the modern synthesis. *Trends in Ecology and Evolution*, 10: 294-
997 299.
998 Mallet J. (2008) Hybridization, ecological races and the nature of species: empirical evidence for the ease
999 of speciation. *Proceedings of the Royal Society B: Biological Sciences*, 363, 2971-2986.
1000 Manel S, Schwartz MK,Luikart G and Taberlet P. (2003) Land scape genetics: combining landscape
1001 ecology and population genetics. *Trends in Ecology and Evolution*, 18:189-197.
1002 Manni F, Guérard E and Heyer E. (2004) Geographic patterns of (genetic, morphologic, linguistic)
1003 variation: how barriers can be detected by using Monmonier's algorithm. *Human Biology*, 76:173-90
1004 Maruthi MN, Hillocks RJ, Mtunda K, Raya MD, Muhanna M, Kiozia H, *et al*. (2005) Transmission of
1005 Cassava brown streak virus by *Bemisia tabaci (*Gennadius). *Journal of Phytopathology*, 312: 307–312.
1006 Mayr, E. (1942) *Systematics and the origin of species.* Columbia University. Press, New York.

Mayr, E. (1963) *Animal species and evolution.* The Belknap Press of Harvard University. Press, Cambridge, U.K

McDaniel SF and Shaw AJ. (2003) Phylogeographic structure and cryptic speciation in the trans-Antarctic moss *Pyrrhobryum mnioides*. *Evolution*, 57: 205-215.

Moritz C and Cicero C. (2004) DNA barcoding: promise and pitfalls. *PLoS Biology*, 2: e354.

Mound LA. (1963) Host-correlated variation in *Bemisia tabaci* (Gennadius)(Homoptera: Aleyrodidae). *Physiological Entomology*, 38: 171-180.

Mound LA and Halsey SH. (1978) Whitefly of the world. A systematic catalogue of the *Aleyrodidae* (*Homoptera*) with host plant and natural enemy data. (John Wiley and Sons), pp, 340.

Mugerwa H, Wang HL, Sseruwagi P, Seal S and Colvin J. (2020) Whole-genome single nucleotide polymorphism and mating compatibility studies reveal the presence of distinct species in sub-Saharan Africa *Bemisia tabaci* whiteflies. *Insect Science*, 28: 1553-1566.

Mugerwa H, Rey MEC, Tairo F, Ndunguru J and Sseruwagi P. (2019) Two sub-Saharan Africa 1 populationsof *Bemisia tabaci* exhibit distinct biological differences in fecundityand survivorship on cassava. *Crop Protection*, 117: 7–14.

Mugerwa H, Seal S, Wang HL, Patel MV, Kabaalu R, Omongo CA, *et al.*(2018) African ancestry of New World, *Bemisia tabaci*-whitefly species. *Scientific Reports*, 8:1-11.

Mware B, Narla R, Amata R, Olubayo F, Songa J, Kyamanyua S, et al. (2009) Efficiency of cassava brown streak virus transmission by two whitefly species in coastal Kenya. *Journal of General and Molecular Virology*, 1: 40–45.

Nadler SA. (1995) Advantages and disadvantages of molecular phylogenetics: a case study of ascaridoid nematodes. *Journal of Nematology*, 27:423-32.

Naranjo SE, CastleSJ, De Barro PJ and Liu SS. (2009) Population dynamics, demography, dispersal and spread of *Bemisia tabaci*. In "*Bemisia*: Bionomics and Management of a Global Pest", eds Stansly PA & Naranjo SE, Springer, pp. 185-226.

Nater A,Mattle-Greminger MP, Nurcahyo A, Nowak MG,de Manuel M, Desai T, *et al.* (2017) Morphometric, behavioral,and genomic evidence for a new Orangutan species. *Current Biology*, 27:3487-3498.

Navas-Castillo J, Fiallo-Olivé E and Sánchez-Campos S. (2011) Emerging virus diseases transmitted by whiteflies. *Annual Review of Phytopathology*, 49: 219-248.

Noor MAF. (2002) Is the biological species concept showing its age? *Trends in Ecology and Evolution*, 17:153–154.

Nygren A. (2014) Cryptic polychaete diversity: a review. *Zoologica Scripta*, 43: 172-183.

Packer L, Monckton SK, Onuferko TM and Ferrari RR. (2018) Validating taxonomic identifications in entomological research. *Insect Conservation and Diversity*, 11: 1-12.

Padial JM, Miralles A, De La Riva I and Vences M. (2010) The integrative future of taxonomy. *Frontiers in Zoology*, 7: 16.

1044 Pante E, Puillandre N, ViricelA, Arnaud-Haond S, Aurelle D, Castelin M, *et al*. (2015) Species are
1045 hypotheses: avoid connectivity assessments based on pillars of sand. *Molecular Ecology*, 24: 525-544.
1046 Paterson ID, Mangan R, Downie DA, Coetzee JA, Hill MP, Burke AM, *et al*. (2016) Two in one: cryptic
1047 species discovered in biological control agent populations using molecular data and crossbreeding
1048 experiments. *Ecology and Evolution*, 6: 6139-50.
1049 PerringTM. (2001) The *Bemisia tabaci* species complex. *Crop Protection*, 20: 725-737.
1050 Pfenninger M and Schwenk K. (2007) Cryptic animal species are homogeneously distributed among taxa
1051 and biogeographical regions. *BMC Evolutionary Biology*, 7: 121.
1052 Polston JE, De Barro P and Boykin LM. (2014) Transmission specificities of plant viruses with the newly
1053 identified species of the *Bemisia tabaci* species complex. Pest Management Science, 70: 1547-1552.
1054 Qin L, Pan LL and Liu SS. (2016) Further insight into reproductive incompatibility between putative
1055 cryptic species of the *Bemisia tabaci* whitefly complex. *Insect Science*, 23: 215-224.
1056 R Core Team. (2017) R: A language and environment for statistical computing. Vienna, Austria: R
1057 Foundation for Statistical Computing.
1058 Rambaut A. (2012) FigTree version 1.4. 0. Available at http://tree.bio.ed.ac.uk/software/figtree.
1059 Rambaut A and Drummond AJ. (2007) Tracer v1. 4. Available at http://beast.bio.ed.ac.uk/Tracer.
1060 Rambaut A and DrummondAJ. (2013) TreeAnnotator v1. 7.0. Available at
1061 http://beast.bio.ed.ac.uk/TreeAnnotator.
1062 Rambaut A and Drummond AJ. (2015) LogCombiner v1. 8.2. Available at
1063 http://beast.bio.ed.ac.uk/LogCombiner.
1064 Rissler LJ and Apodaca JJ. (2007) Adding more ecology into species delimitation: ecological niche models
1065 and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*).
1066 *Systematic Biology*, 56: 924-942.
1067 Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, *et al*. (2012) MrBayes 3.2:
1068 efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic*
1069 *Biology*, 61: 539-542.
1070 Rubin BER, Ree RH and Moreau CS. (2012) Inferring phylogenies from RAD sequence data. *PLoS One*, 7:
1071 1-12.
1072 Russell LM. (1957) Synonyms of *Bemisia tabaci* (Gennadius)(Homoptera: Aleyrodidae). *Bulletin of the*
1073 *Brooklyn Entomological Society*, 52: 122-123.
1074 Schlick-Steiner BC, Seifert B, Stauffer C, Christian E, Crozier RH and Steiner FM. (2007) Without
1075 morphology, cryptic species stay in taxonomic crypsis following discovery. *Trends in Ecology and*
1076 *Evolution*, 22: 391-392.
1077 Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E and Crozier RH. (2010) Integrative
1078 taxonomy: a multisource approach to exploring biodiversity. *Annual Review of Entomology*, 55: 421-
1079 438.

Smith MA, Rodriguez JJ, Whitfield JB, Deans AR, Janzen DH, Hallwachs W and Hebert PD. (2008) Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology, and collections. *Proceedings of the National Academy of Sciences of the United States of America*, 105:12359-12364.

Stamatakis A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22: 2688-2690.

Stamatakis A, Hoover P and Rougemont J. (2008) A rapid bootstrap algorithm for the RAxML web servers. *BMC Systems Biology*, 57: 758-771.

Stireman JO, Nason JD and Heard SB. (2005) Host-associated genetic differentiation in phytophagous insects: General phenomenon or isolated exceptions? Evidence from a goldenrod-insect community. *Evolution*, 59: 2573-2587.

Struck TH, Feder JL, Bendiksby M, Birkeland S, Cerca J, Gusarov VI, *et al*. (2017) Finding evolutionary processes hidden in cryptic species. *Trends in Ecology and Evolution*, 33: 153-163.

Sukumaran J and Knowles LL. (2017) Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences of the United States of America*, 114: 1607-1612.

Sun DB, Li J, Qin L, Xu J and Li FF. (2013) Competitive displacement between two invasive whiteflies: insecticide application and host plant effects. *Bulletin of Entomological Research*, 5: 1-10.

Singhal S, Hoskin CJ, Couper P, Potter S and Moritz C. (2018) A Framework for resolving cryptic species: a case study from the lizards of the Australian wet tropics. *Systematic Biology*, 67: 1061-1075.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M and Kumar S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28: 2731-2739.

Taquet A, Jourdan-Pineau H, Simiand C, Grondin M, Barrès B and DelatteH. (2022) Distribution of invasive versus native whitefly species and their pyrethroid knock-down resistance allele in a context of interspecific hybridization. *Scientific Reports*, 12:8448.

Tay WT, Elfekih S, Court L, Gordon KH and De Barro PJ. (2014) Complete mitochondrial DNA genome of *Bemisia tabaci* cryptic pest species complex Asia I (Hemiptera: Aleyrodidae). *Mitochondrial DNA*, 27: 972-973.

Thao ML, Baumann L and Baumann P. (2004) Organization of the mitochondrial genomes of whiteflies, aphids, and psyllids (Hemiptera, Sternorrhyncha). *BMC Ecology and Evolution*, 4: 25.

Thresh JM, Otim-Nape GW, Legg JP and Fargette D. (1997) African cassava mosaic virus disease: the magnitude of the problem. *African Journal of Root and Tuber Crops*, 2:13-9.

Vyskočilová S, Tay WT, Van Brunschot S, Seal S and Colvin J. (2018) An integrative approach to discovering cryptic species within the *Bemisia tabaci* whitefly species complex. *Scientific Reports*, 8: 1-13.

Vyskočilová S, Seal S and Colvin J. (2019) Relative polyphagy of "Mediterranean" cryptic *Bemisia tabaci* whitefly species and global pest status implications. *Journal of Pest Science*, 92: 1071-1088.

1117 Wang HL, Yang J, Boykin LM, Zhao QY, Li Q, Wang XW, *et al*. (2013) The characteristics and
1118 expression profiles of the mitochondrial genome for the Mediterranean species of the *Bemisia tabaci*
1119 complex. *BMC Genomic*, 14: 401.
1120 Wang P, Sun DB, Qiu BL and Liu SS. (2011) The presence of six cryptic species of the whitefly *Bemisia*
1121 *tabaci* complex in China as revealed by crossing experiments. *Insect Science*, 18: 67-77.
1122 Wang Z, Yan H, Yang Y and Wu Y. (2010) Biotype and insecticide resistance status of the whitefly
1123 *Bemisia tabaci* from China. *Pest Management Science*, 66: 1360-1366.
1124 Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, *et al*. (2013) Genome-wide RAD
1125 sequence data provide unprecedented resolution of species boundaries and relationships in the Lake
1126 Victoria cichlid adaptive radiation. *Molecular Ecology*, 22: 787-798.
1127 Wei J, Zhao J, Zhang T, Li F, Ghanim M, Zhou X, *et al*.(2014) Specific cells in the primary salivary glands
1128 of the whitefly *Bemisia tabaci* control retention and transmission of begomoviruses. *Journal of*
1129 *Virology*, 88: 13460-13468.
1130 Wiens JJ and Penkrot TA. (2002) Delimiting species using DNA and morphological variation and
1131 discordant species limits in spiny lizards (*Sceloporus*). *Systematic Biology*, 51: 69-91.
1132 Wolstenholme DR (1992) Animal mitochondrial DNA: structure and evolution. International Review of
1133 *Cytology*, 141: 173-216
1134 Wosula EN, Chen W, Fei Z and Legg JP. (2017) Unravelling the genetic diversity among cassava *Bemisia*
1135 *tabaci* whiteflies using NextRAD sequencing. *Genome Biology and Evolution*, 9: 2958-2973.
1136 Wyman SK, Jansen RK and Boore JL. (2004) Automatic annotation of organellar genomes with DOGMA.
1137 *Bioinformatics*, 20: 3252-3255.
1138 Xie W, Chen C, Yang Z, Guo L, Yang X, Wang D, *et al.* (2017) Genome sequencing of the sweetpotato
1139 whitefly *Bemisiatabaci* MED/Q. *GigaScience*, 6: 1.
1140 Xu J, De Barro PJ and Liu SS. (2010) Reproductive incompatibility among genetic groups of *Bemisia*
1141 *tabaci* supports the proposition that the whitefly is a cryptic species complex. *Bulletin of*
1142 *Entomological Research*, 100: 359-66.
1143 Yang Z and Rannala B. (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings*
1144 *of the National Academy of Sciences of the United States of America*, 107: 9264-9269.

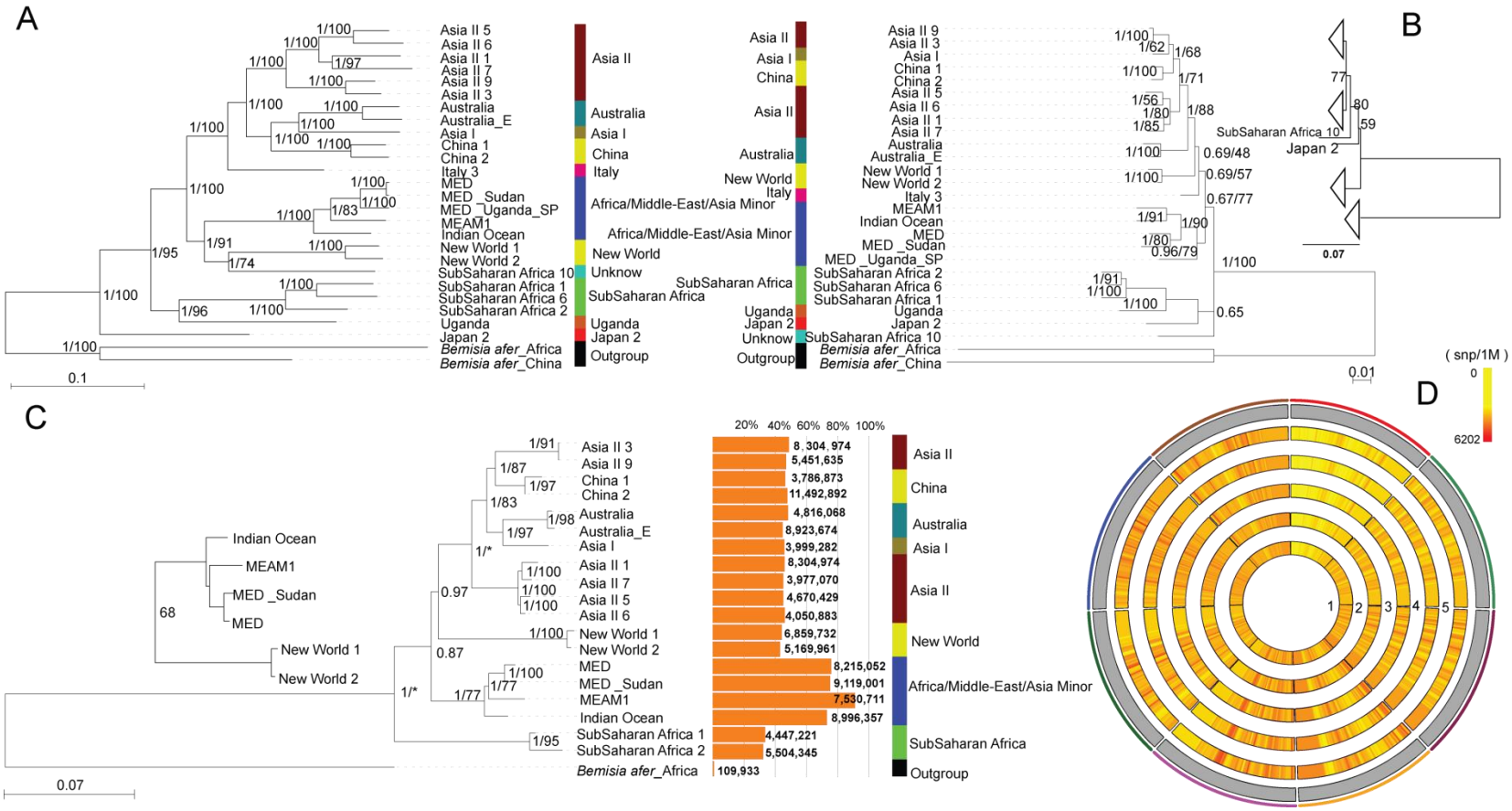

Fig. 1. Phylogenetic trees of *B. tabaci* (all reconstructed using the Bayesian and ML method). A Tree based on full mitochondrial genomes. **B** Tree based on four single-copy nuclear genes. **C** Tree based on the RAD-Seq dataset (minimum samples= 6) and the mapping rate (reads mapped to the MEAM1 genome) and the number of mapping reads. **D** SNP distribution by mapping all samples to the MEAM1 genome (the number in the figure represents different matrices: 5 represents "matrix=6", 4 represents "matrix=8", 3 represents "matrix=10", 2 represents "matrix=12" and 1 represents "matrix=14"). To view the SNP distribution around the MEAM1 genome, scaffolds of the MEAM1 genome were connected to eight bigger scaffolds according to length of the scaffolds and then using these eight scaffolds connected as a cycle. For all phylogenetic trees, numbers at nodes detail BI posterior probabilities greater than 0.95/ML bootstrap values over 50%.

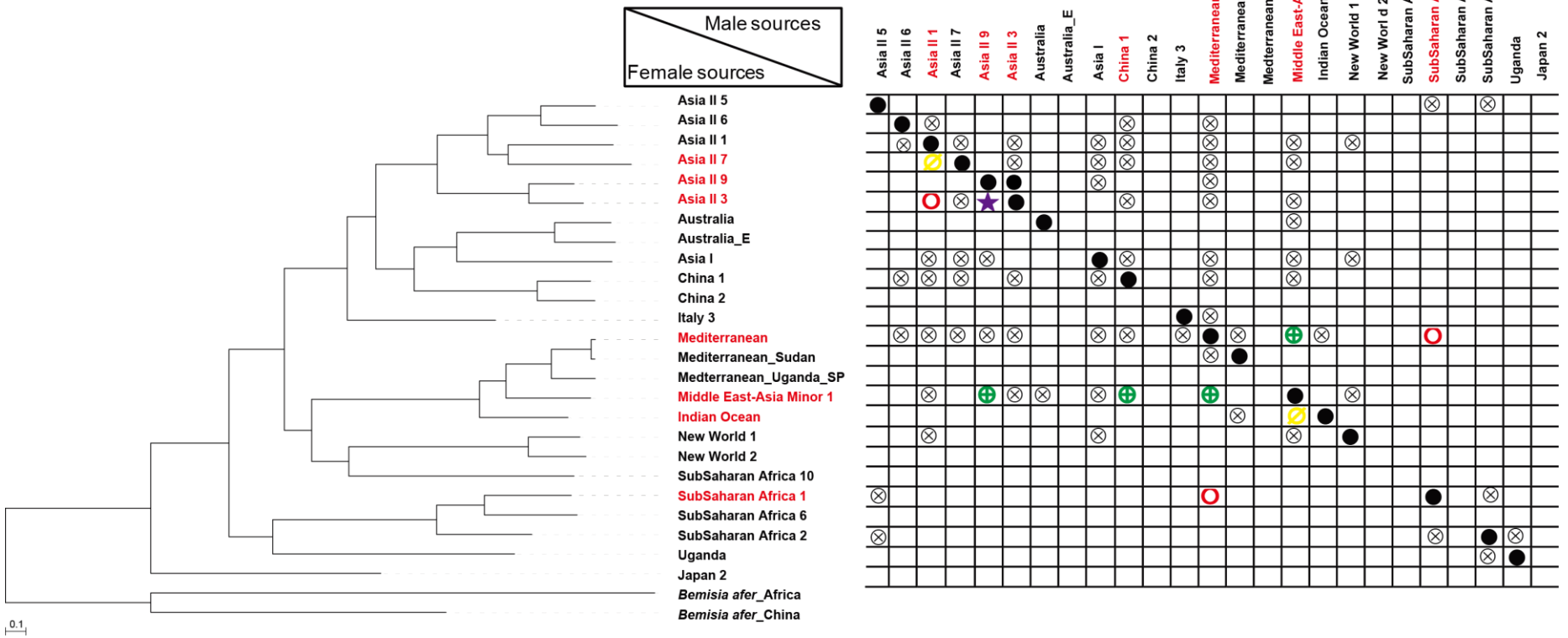
Male sources
Female sources
Asia II 5
Asia II 6
Asia II 1
Asia II 7
Asia II 9
Asia II 3
Australia
Australia_E
Asia I
China 1
China 2
Italy 3
Mediterranean
Mediterranean_Sudan
Medterranean_Uganda_SP
Middle East-Asia Minor 1
Indian Ocean
New World 1
New World 2
SubSaharan Africa 10
SubSaharan Africa 1
SubSaharan Africa 6
SubSaharan Africa 2
Uganda
Japan 2
Bemisia afer_Africa
Bemisia afer_China
0.1

Fig. 2. Summary of reproductive incompatibility among putative species of the *B. tabaci* cryptic species group from published crossing studies, as well as unpublished data from an experiment conducted for this paper.
The mating experiments carried out linked to the BI tree generated by mitogenomic data. Specimens marked with red color represent a conflict between the reproductive compatibility and phylogenetic clade boundaries.
The codes of male source correspond to those of female source listed in the top column.
Symbols marked with shapes and colors indicate levels of reproductive incompatibility.

- ● Complete reproductive compatibility.
- ⊗ No F_1 hybrid females produced.
- ⊕ Marked with green color represent low number of F_1 hybrid females produced in both directions of cross but the hybrids were sterile and/or characterized by reduced viability and fertility.
- ★ Marked with purple color represent lower number of F_1 hybrid females produced, and hybrid females were fertile.
- ∅ Marked with yellow color represent low number of F_1 hybrid females produced in one direction of cross, but hybrids had reduced viability and fertility.
- ○ Marked with red color represent low number of F_1 hybrid females produced but fertility of F_1 females was not tested.

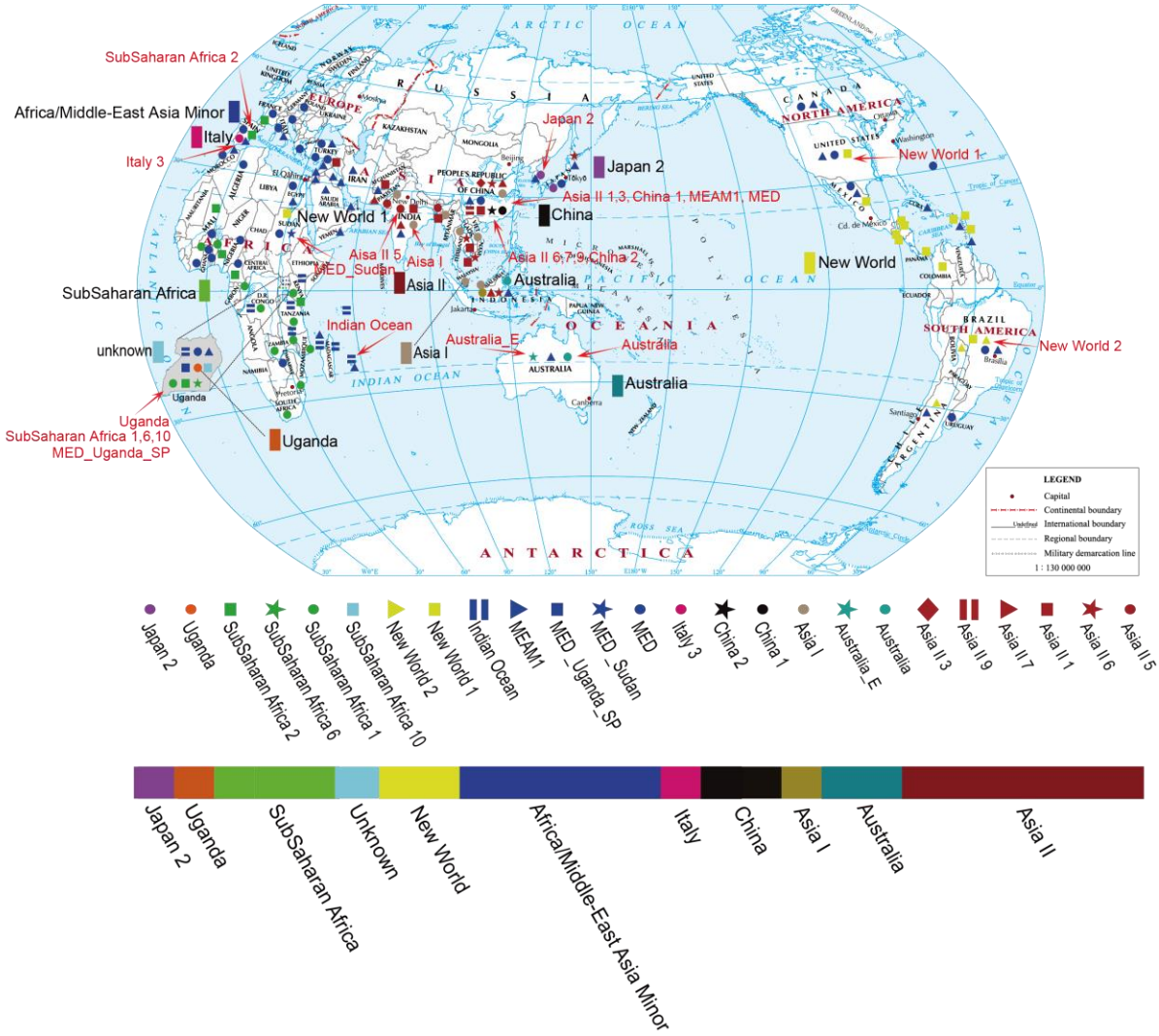

Fig. 3. Geographical distribution of different putative cryptic species. Geographical distribution of different putative cryptic species was summarized from literature reviews. Different clades are shown in different colors and the specimens under the same clades are noted by different shapes of symbols. Red arrows in the map represent the geographical origin of the 25 analyzed samples.

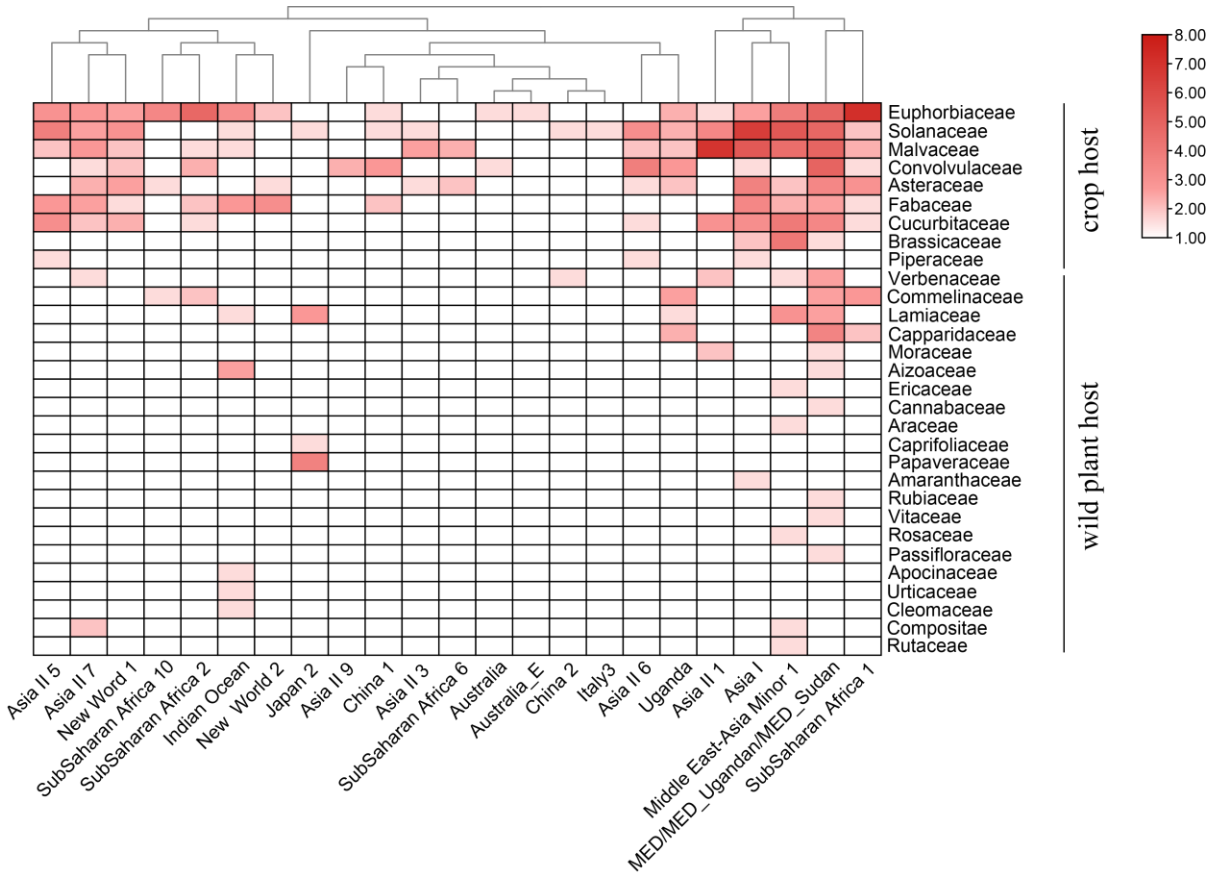

Fig. 4. The host range of species groups in the *Bemisia tabaci* complex. The reported number of host families utilized by different species were transformed into log2 and used for heatmap and clustering analysis.

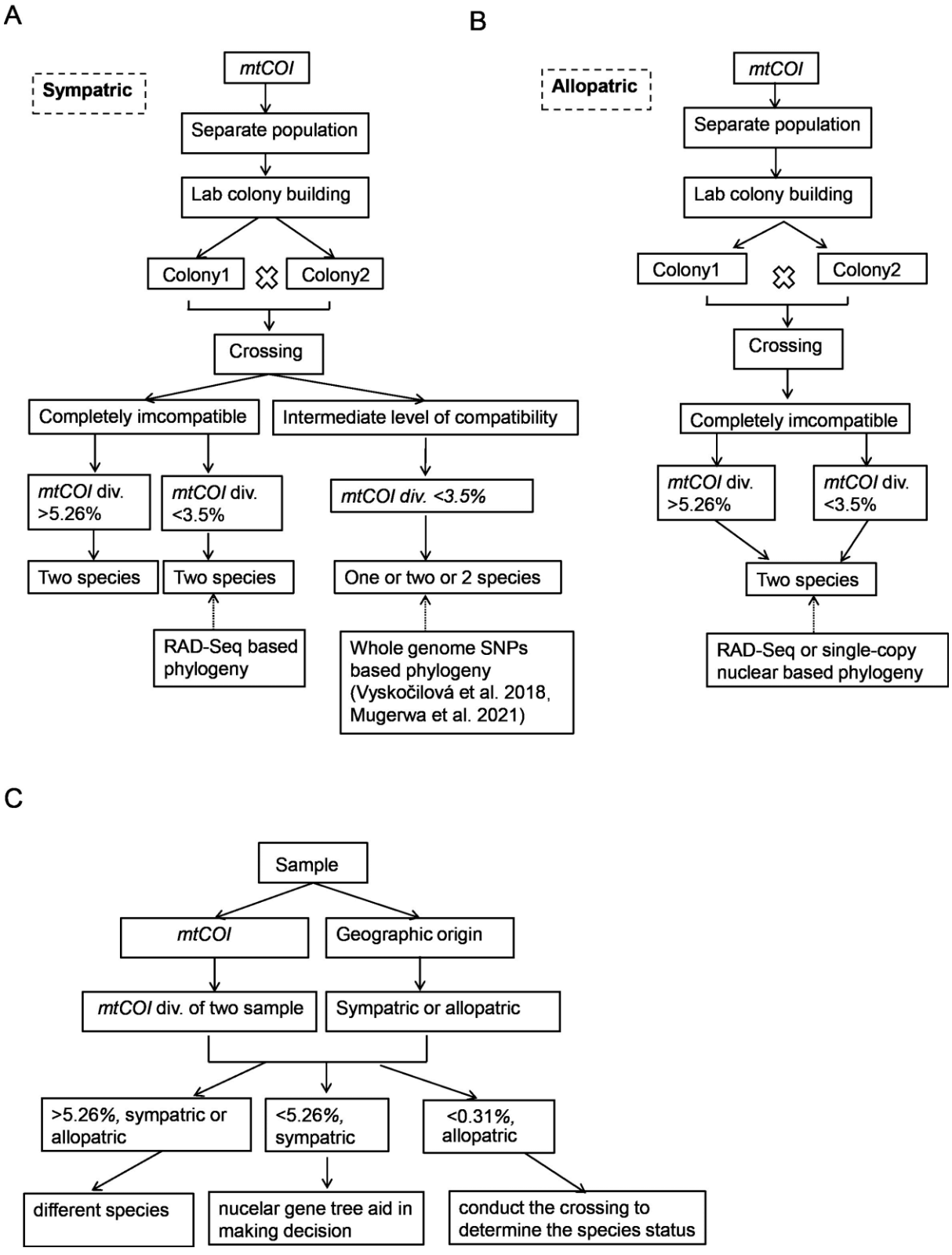

Fig. 5. Summary of the performance of different datasets in *B. tabaci* species classification as judged by reproductive compatibility datasets (A,B) and a new strategy proposed for species classification (C).