On species delimitation, hybridization and population structure of cassava whitefly in Africa

Elfekih S[1,2a], Tay WT[3a†], Polaszek A[4], Gordon KHJ[3], Kunz D[3], Macfadyen S[3], Walsh TK[3], Vyskočilová S[5], Colvin J[5], De Barro PJ[6†]

(1) CSIRO, Australian Animal Health Laboratory, Geelong, VIC, Australia

(2) University of Melbourne, Bio21 Institute, Parkville, VIC, Australia

(3) CSIRO, Black Mountain Laboratories, ACT, Australia

(4) Department of Life Sciences, Natural History Museum, London, UK

(5) Natural Resources Institute, University of Greenwich, Kent, UK

(6) CSIRO, Ecosciences Precinct, Brisbane, Australia

(a) These authors contributed equally

(†) Corresponding authors

Key words: gene flow, speciation, hybridization, pest, genomics

Abstract

The *Bemisia* cassava whitefly complex includes species that cause severe crop damage through vectoring cassava viruses in eastern Africa. Currently, the cassava whitefly complex is divided into species and subgroups based on very limited molecular markers that did not allow clear definition of species and population structure. Based on 14,358 genome-wide SNPs from 63 cassava whitefly individuals belonging to sub-Saharan African (SSA1, SSA2 and SSA4) species, and using a well-curated mtCOI gene database, we show clear incongruities in previous taxonomic approaches underpinned by effects from pseudogenes. We show that the SSA4 species is part of the SSA2 species, and that populations of the SSA1 species comprise of south-eastern (Madagascar, Tanzania) and north-western (Nigeria, Democratic Republic of Congo and Burundi) sub-species that show signatures of allopatric incipient speciation, with a hybrid zone separating these adjacent sub-species. These findings provide the first genomic insights into the evolution and molecular ecology of a highly cryptic hemipteran insect complex in African, and allow the systematic use of genomic data in management and control strategies for this important cassava pest.

39 **Introduction**

40 Confidence in species identification underpins strategies for biological conservation
41 and management of natural resources, providing insight into the evolutionary forces that
42 drive species diversification. While morphological differences have enabled many species to
43 be identified, cryptic species complexes present a particular challenge due to their general
44 lack of readily distinguishable morphological characters. Collecting ecological and life
45 history data, including host plant use and mating behaviour, can be used effectively to
46 mitigate and address these challenges (Liu, et al. 2007; Vyskočilová, et al. 2019;
47 Vyskočilová, et al. 2018), but are often time-consuming and impractical, especially with
48 large numbers of samples. Incorporating DNA markers, i.e., diagnostic mitochondrial DNA
49 (mtDNA) markers (e.g., Behere, et al. 2008; Elfekih, et al. 2018b; Walsh, et al. 2019),
50 genome-wide sequencing (Anderson, et al. 2016; Elfekih, et al. 2018a), and whole-genome
51 resequencing (Anderson, et al. 2018), can increase confidence in differentiating between
52 cryptic species in a complex and identifying inter-species hybrids.

53 The challenges can become critical when individual cryptic species in a complex have
54 a serious impact on health or agricultural productivity. An example of the latter is presented
55 by insect vectors of virus diseases affecting cassava, the most important food crop grown in
56 Africa (55.2% global production share; average from 1994-2011; FAOSTAT 2017). Since
57 the 1990's, cassava production has been affected by a pandemic of cassava mosaic disease
58 (CMD) caused by a number of cassava mosaic geminiviruses (CMGs) (Legg, et al. 2014;
59 Patil and Fauquet 2009). Although cassava is also widely grown in Latin America (its
60 original home) and South East Asia, CMD has been reported in Africa since the early 1970's
61 (Macfadyen, et al. 2018) and more recently in South-East Asia (Minato, et al. 2019; Wang, et
62 al. 2016). While the vector responsible for the spread of the CMD in South-East Asia is not
63 fully understood, the epidemic in east and central Africa was associated with the distribution

64 of invasive Sub-Saharan African (SSA) species of the *Bemisia* whitefly complex (Legg, et al.
65 2014). Invasive whitefly species have become major pests of global agriculture, spreading
66 plant virus diseases and becoming resistant to chemical control agents (De Barro, et al. 2011).
67 In particular, the invasive cassava whitefly lineage SSA2 (previously 'Ug2'; Boykin, et al.
68 2017) has been found in regions affected by the CMD pandemic, and since 2007 has also
69 successfully expanded into southern Europe (Hadjistylli, et al. 2015), while another lineage,
70 the SSA1 (previously 'Ug1'; Boykin, et al. 2017), was found beyond the pandemic-affected
71 geographical regions (i.e., widespread throughout the African continent).

72 Defining the actual CMD vectors in Africa remains a challenge due to the cryptic
73 nature of the many species in the Aleyrodidae whitefly *Bemisia* genus (De Barro, et al. 2011;
74 Lee, et al. 2013; Mugerwa, et al. 2018; Vyskočilová, et al. 2018). At least 38 cryptic species
75 have now been recognised within what was previously considered a single *Bemisia tabaci*
76 (Gennadius) species (Martin and Mound 2007) through the characterisation of a partial
77 mitochondrial DNA cytochrome oxidase subunit I (mtCOI) 3' gene region of 657bp (De
78 Barro, et al. 2011; Dinsdale, et al. 2010; Kunz, et al. 2019b). However, re-analyses of partial
79 mtCOI sequences in GenBank relating to the Mediterranean (MED) species by Vyskočilová
80 et al. (2018), and of the *B. tabaci* standard partial mtCOI dataset (Boykin, et al. 2017) by
81 Kunz et al. (2019b), demonstrated that a substantial portion (*ca*. 60-65%) were artefacts of
82 nuclear mitochondrial DNA sequences (NUMTs) or other PCR artefacts (Elfekih, et al.
83 2018a; Kunz, et al. 2019b; Tay, et al. 2017a). This analysis, together with molecular
84 characterisation of the original voucher specimens, is beginning to disentangle the species
85 complex, identifying the Mediterranean (MED) species as the true *B. tabaci* (Tay, et al.
86 2012), the *B. tabaci* AsiaII_7 species as *B. emiliae* Corbett (Tay, et al. 2017b), assisting with
87 identifying historical species ranges (Kunz, et al. 2019a; Tay, et al. 2017b; Tay, et al. 2012),
88 and provided phylogenetic evidence that the cassava whitefly (i.e., SSA1, SSA2, SSA3,

89 SSA6, SSA9) were ancestral to the '*B. tabaci* cryptic species complex' (Kunz, et al. 2019b).
90 The advent of high-throughput sequencing (HTS) technologies has driven the availability of
91 genome-wide data that is further transforming the *Bemisia* taxonomy. HTS has confirmed
92 that the *B. tabaci* Middle East Asia Minor 2 (MEAM2), previously recognised as a distinct
93 species, is in fact the result of an artefact of sequencing nuclear mitochondrial DNA
94 sequences (NUMTs) (Tay, et al. 2017a ; Elfekih, et al. 2018a).

95 The taxonomy, population structure and evolutionary genetics of the cassava whitefly
96 *Bemisia* species complex (i.e., SSA1, SSA2, SSA3, SSA4, SSA6, and SSA9; see Kunz, et al.
97 2019b) remain poorly understood, despite the importance of defining and managing the
98 vectors of CMD. In a recent study of taxonomic structure and gene flow among genetically
99 diverse SSA1, SSA2, SSA3 and SSA4 species, Wosula et al. (2017) described serious
100 discrepancies between their results obtained either from mtCOI or by nextRAD sequencing,
101 with possible implications for understanding the dynamics of the CMD-transmitting species.
102 The findings of Wosula et al. (2017) have wider implications to the future usage of this
103 partial mtCOI gene as the current standard DNA marker for species identification of *B. tabaci*
104 and related *Bemisia* cryptic species. This is especially so in the sub-Saharan African *Bemisia*
105 species context, where species identified to-date would need to be reanalysed using a
106 genome-wide sequencing approach. The diversity within genome-wide SNP data, and
107 complex gene flow patterns existed between these newly defined SSA genetic groups
108 (Wosula, et al. 2017) would appeared to also have a different operational taxonomic unit to
109 challenge the traditional 'species' unit.

110 In this study, we seek to understand why the African cassava whitefly SSA1 (which is
111 also further refined into sub-groups 1, 2, and 3 in some studies; e.g., Legg, et al. (2014)),
112 SSA2, and SSA4 species, should have such conflicting species status and admixture patterns
113 depending on whether the mtCOI or nuclear DNA markers were used. To reconcile these

114 contrasting findings, we re-analysed the mtCOI data of Wosula et al. (2017) and compared it
115 to a strictly filtered set of mtCOI sequences for species status delimitation. Identification and
116 removal of NUMT-derived mtCOI resolved conflict of results from mtCOI discrepancies as
117 reported by Wosula et al. (2017). Reanalyses of genome-wide SNP data enabled insights into
118 SSA1-species-wide population substructure to identify the geographic contact zone and
119 signature of intra- and inter-species hybridization. Our findings provide powerful insights
120 into the evolutionary forces shaping incipient speciation and vector competence of this
121 agriculturally invasive species complex with the capacity to impact on food security in sub-
122 Saharan Africa.

123

124 **Material and Methods**

125 **mtCOI data handling**. The mtCOI sequences from the study by Wosula et al. (2017) were
126 downloaded from GenBank (accession numbers MF417578-MF417602), including all their
127 published sequences and published reference sequences to identify genetic groups within the
128 SSA species. These partial mtCOI sequences were imported into Geneious R11.1 and aligned
129 with reference sequences (Suppl. Fig. 1) using the MAFFT alignment program (Katoh and
130 Standley 2013) with default settings. We examined all sequences for insertion/deletions
131 (indels) as signatures of pseudogenes (NUMTs). Sequences that did not contain indels were
132 assessed also for amino acid substitution patterns at evolutionary conserved regions in the
133 partial COI gene region to assist with further identification of potential NUMTs.
134 Identification of amino acid residue conservation sites across the partial mtCOI gene regions
135 was as described in Kunz et al. (2019b).

136 **nextRAD data processing.** The raw data from nextRAD sequencing of Wosula et al. (2017)
137 was accessed from GenBank (SRP103541). The fastq sequences were screened using FastQC

138 for quality control. The retained reads were trimmed by quality to a total length of 151 bp
139 using Trimmomatic (Bolger, et al. 2014).

140 **Mapping quality**. We mapped the reads from each individual sample to the *B. tabaci*
141 MEAM1 genome (Chen, et al. 2016) using Burrows-Wheeler Aligner (BWA) v.0.7.12 (Li
142 and Durbin 2010). The alignments used the BWA-MEM algorithm with default settings. The
143 SAM files were converted to BAM output using Samtools (Li, et al. 2009). The BAM files
144 were subsequently sorted and indexed, and checked for quality and mapping percentages per
145 scaffold.

146 **SNP genotyping.** SNPs were first called using a *de novo* approach in PyRAD (Eaton 2014).
147 PyRAD is a pipeline designed for RADseq datasets that aims to capture variation at the
148 species/clade level. It allows clustering of highly divergent sequences and takes into
149 consideration indel variation. We also performed SNP calling in PyRAD, using reads mapped
150 to the MEAM1 reference genome (Chen, et al. 2016). Given the low-depth nature of the data,
151 we analysed the data using an additional pipeline. We used the BAM files generated in BWA
152 as input for the program ANGSD (Korneliussen, et al. 2014), which relies on a statistical
153 approach that takes into consideration genotype uncertainty (Fumagalli, et al. 2013).

154 **Genetic clusters and tests for introgression.** Genetic clusters within the dataset that consists
155 of 63 SSA samples were investigated by Principal Component Analysis (PCA) using the
156 SNPRelate R package (Zheng, et al. 2012). We used ADMIXTURE v.1.3.0 (Alexander, et al.
157 2009) on the 63-sample dataset to identify genetic clusters and estimate the genetic ancestry
158 of the SSA samples. We ran the program with K ancestral clusters varying from 1 to 10 with
159 each 'K' value repeated 100 times. A cross-validation test was performed to determine the
160 optimal K value. The program ANGSD v.0.911 (Korneliussen, et al. 2014) was used to run
161 an ABBA-BABA test (D-statistics) to investigate introgression patterns between the 63
162 samples (SSA1 and SSA2 species) and with the MEAM1 genome as an outgroup. The

163 ABBA-BABA test is based on the comparison of the number of tree topologies of the ABBA
164 and BABA patterns. In this study, the SSA species group patterns we investigated were as
165 defined by Wosula et al. (2017) and consisted of
166 [ESA+CA/SSA2+SSA4/ECA+WA/MEAM1], where 'ESA', 'CA', 'ECA', and 'WA'
167 represented East South Africa, Central Africa, East Central Africa, and Western Africa,
168 respectively. We used the MEAM1 species as outgroup for comparison and to infer
169 introgression patterns. We tested the significance of these ABBA/BABA patterns using the
170 corresponding Z-scores, obtained in ANGSD using a jack-knife procedure. An absolute Z-
171 score value of ≥ 3 is used as a cut-off value (Reich, et al. 2009).

172 **Phylogenetic analyses.** The topology of the phylogenetic relationships between individuals
173 /populations/subspecies and species within the dataset (63 SSA samples and three *B. afer*
174 individuals (TZ_CHA1, TZ_CHA2, TZ_KIL1) as the outgroup) was examined using a
175 maximum likelihood (ML) approach. The SSA species were first identified based on the
176 mtCOI marker, then, a phylogeny was reconstructed using the nextRAD genome-wide SNPs
177 excluding samples with low genotype quality to minimize biases that could potentially be
178 introduced by missing data. The phylogenetic reconstruction was carried out in RAxML
179 v.7.2.8 (Stamatakis 2006) using the GTR substitution model and GTRGAMMA as the
180 GAMMA model of rate heterogeneity, with 1,000 bootstrap replications. The phylogenetic
181 clustering obtained in RAxML was used to infer a population ML tree in TreeMix (Pickrell
182 and Pritchard 2012), in order to identify genetic mixing, the history of population splits and
183 admixture patterns. We ran the TreeMix simulations using 0 to 5 migrations events with 100
184 bootstrap replications.

185

186 **Results**

187 **Integrity of African cassava *Bemisia* SSA mtCOI sequences.** Of the 11 SSA4 sequences
188 (originally labelled as 'Sub-Saharan III: Western Africa-Cameroon/Cassava'; see Berry et al.
189 (2004), Table 1), only two (Cam Ayos 1 WO2 and Cam Ayos 2 WO3) did not exhibit any
190 premature stop codons at the amino acid level or have indels at the nucleotide level (Suppl.
191 Fig. 1). Both of these sequences were also much shorter than the remaining nine sequences
192 within the Sub-Saharan III: Western Africa-Cameroon/Cassava clade. Comparison of amino
193 acid composition against 21 high-throughput sequencing-generated mtCOI genes
194 representing 12 evolutionary divergent *Bemisia* cryptic species (Kunz, et al. 2019b), enabled
195 Cam Ayos 2 WO3 to be identified as a potential NUMT sequence. Indels, stop codons, non-
196 conservation of amino acid residues in highly conserved mtCOI region therefore ruled out 10
197 of the 11 SSA4 sequences as potential NUMTs. The remaining single, short but 'functional'
198 Cam Ayos 1 WO2 sequence (AF344246) that clustered with the 10 NUMT-related SSA4
199 sequences suggested that as a whole, these SSA4 sequences were likely of nuclear genomic
200 origin. The clustering of SSA4 sequences to SSA2 indicates it likely to be of SSA2 nuclear
201 origin, similar to that reported for *B. tabaci* MEAM2 NUMT that originated from the nuclear
202 genome of MEAM1 (Elfekih, et al. 2018a; Tay, et al. 2017a).

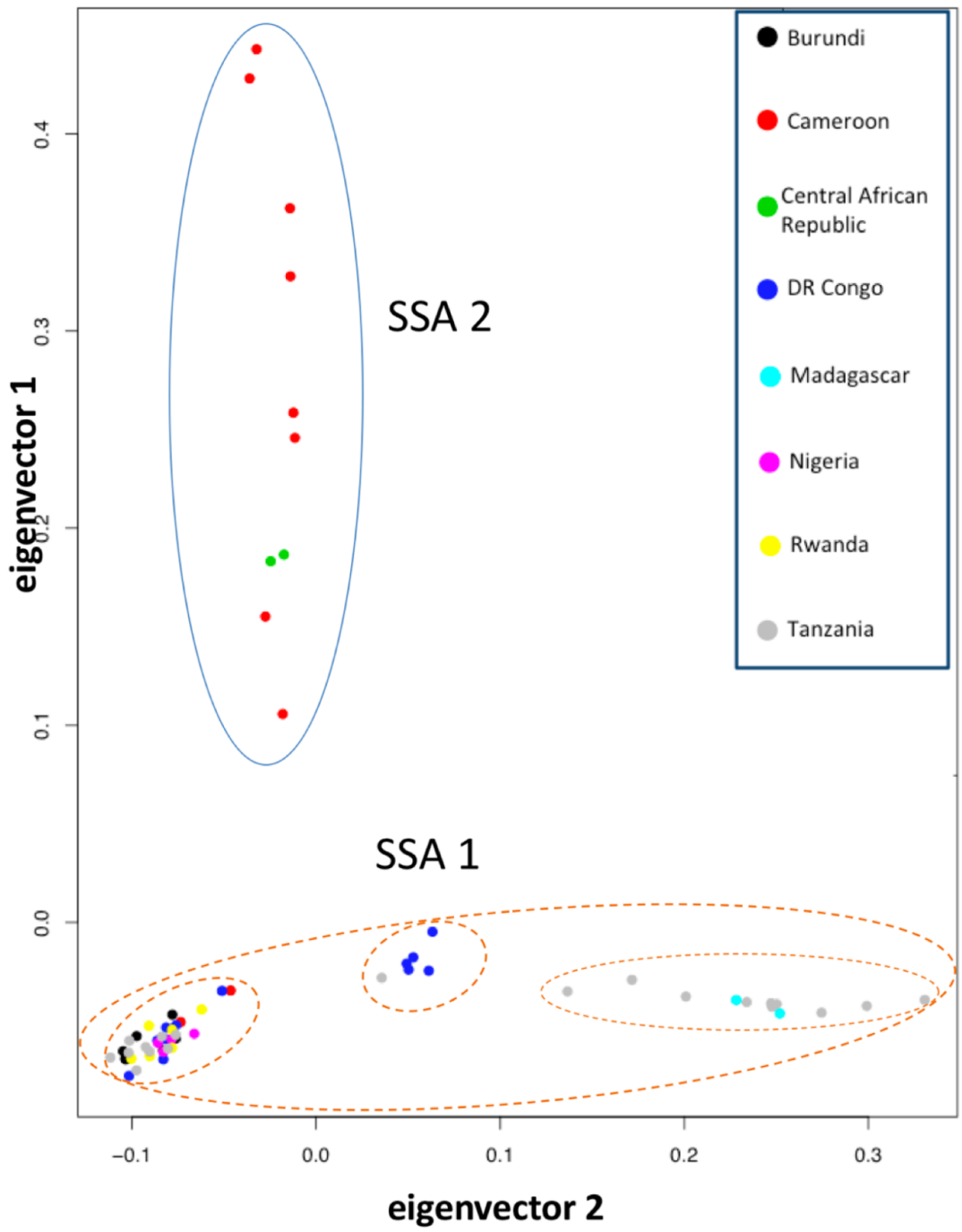
203 The examination of nucleotide substitution patterns identified indels-affected
204 sequences and inconsistencies with respect to number of base substitutions for distinguishing
205 'subgroups' (SG) within the SSA1 species. For example, the SSA1-SG1/SG2 sequences
206 (accession numbers: MF417578-MF417602) were affected by indels that should not be
207 present in a mitochondrial DNA COI protein coding gene, suggesting that these sequences
208 were associated with sequencing errors and/or derived from non-coding NUMT DNA
209 sequences. The re-assessment of the two SSA1-SG5 sequences (MF417585, MF417586)
210 showed that they differed from the SSA1-SG3 sequences by only three C/T substitutions that
211 resulted in 0.56% nucleotide difference. The basis for suggesting MF417585 and MF417586

212 belong to different 'sub-groups' was therefore unclear, given that similar numbers of
213 nucleotide substitutions between sequences (e.g., MF417602 and MF417587: SSA1-SG2;
214 two transitions (i.e., 2xC/T), 0.37% nucleotide difference), and sequences with greater
215 numbers of nucleotide substitutions (e.g., MF417590, MF417582: SSA1-SG1; three
216 transitions (2xG/A, 1xT/C) and one transversion (A/T); 0.75% nucleotide difference) were
217 placed within the same subgroups.

218 **Species delimitation and introgression.** The genome-wide SNP-based Principal Component
219 Analysis (PCA) discriminated between SSA1 and SSA2, suggesting that they are genetically
220 distinct and likely to be different species, whereas samples from SSA4 clustered with SSA2
221 (Fig. 1). Three groupings were identified within SSA1, suggesting a level of intra-species
222 substructure. These observations were further explored using ML phylogenies generated with
223 RAxML. An initial such analysis based on the mtCOI sequences (Fig. 2A) broadly divided
224 these 63 SSA individuals into three clusters containing the populations of (Wosula, et al.
225 2017): (i) SSA2 cluster containing individuals previously identified as either SSA4 or SSA2
226 by mtCOI gene sequences with 100% bootstrap confidence, (ii) SSA1 cluster of the
227 populations defined as 'SSA-CA/SSA-ESA' (node confidence 89%), and (iii) SSA1 cluster
228 that defined as 'SSA-ECA/SSA-WA' populations (node confidence 96%).

229

230 **Fig. 1. Principal component analysis of SSA *Bemisia* samples.** Principal component
231 analysis (PCA) was based on 14,358 genome-wide SNPs across the 63 sub-Saharan African
232 cassava *Bemisia* samples. Clustering of SSA2 (blue circle) from SSA1 (red circles) supported
233 their genetic distinctness as separate species. Three groupings are identified within SSA1,
234 suggesting a level of intra-species substructure including the potential gene flow contact
235 zone. Country of origins for these SSA1 and SSA2 *Bemisia* species are shown by unique
236 colour codes.



237

238

239 A comprehensive RAxML phylogenomic analysis based on the 14,358 genome-wide
240 SNPs (Figs. 2B, 2C) confirmed SSA1 and SSA2 species status and is consistent with that
241 earlier identified using partial mtCOI gene sequence identity. This analysis clearly showed
242 SSA1 to comprise two clusters, denoted SSA1-SE (South-Eastern) and SSA1-NW (North-
243 Western) (Fig. 2C), respectively, reflecting the geographic distribution of the populations

identified within them (SSA1-SE comprising samples from Democratic Republic of Congo (DRC), Tanzania (TZ), and Madagascar (MAD); and SSA1-NW comprising samples from Cameroon (CMR), Nigeria (NGA), Democratic Republic of Congo (DRC), Rwanda (RWA), Burundi (BUR) and Tanzania (TZ)). Within SSA1-NW, the 'SSA-ECA' and 'SSA-WA' populations as defined by Wosula et al. (2017) are clustered separately with 99% and 96% confidence values. Within SSA1-SE, the 'SSA-ESA' and 'SSA-CA' populations as defined by Wosula et al. (2017) are partially resolved, sharing only 34% confidence and suggesting insufficient genomic evidence for this phylogenetic analysis to confirm them as separate genetic groups.

Fig. 2. Phylogeny of SSA populations. A: Phylogeny based on mtCOI of 63 SSA *Bemisia* individuals. The phylogeny was constructed using RAxML with 1,000 bootstrap replications. Sample names are as in Fig. S1, and GenBank accession numbers for the sequences used, and corresponding mtCOI SSA1-subgroups (SG) of either SG1, SG2 and SG3, are listed on the right. The *Bemisia* SSA2-SSA4 clade is indicated by yellow branches, and the lineages corresponding to the SSA1 populations analysed are indicated as follows: CA (pink), ESA (black), WA (green) and ECA (red). **B:** Phylogeny based on 14,358 genome-wide SNPs of the individuals analysed in A. The phylogeny was constructed using RAxML with 1,000 bootstrap replications. Sample names and lineage colours are as for **A** above. **C:** Detail of the RAxML phylogeny in **B** showing only the SSA1 and SSA2-SSA4 lineages, with population names added. Bootstrap confidence percentages are visible for the main branches.

Fig. 2A

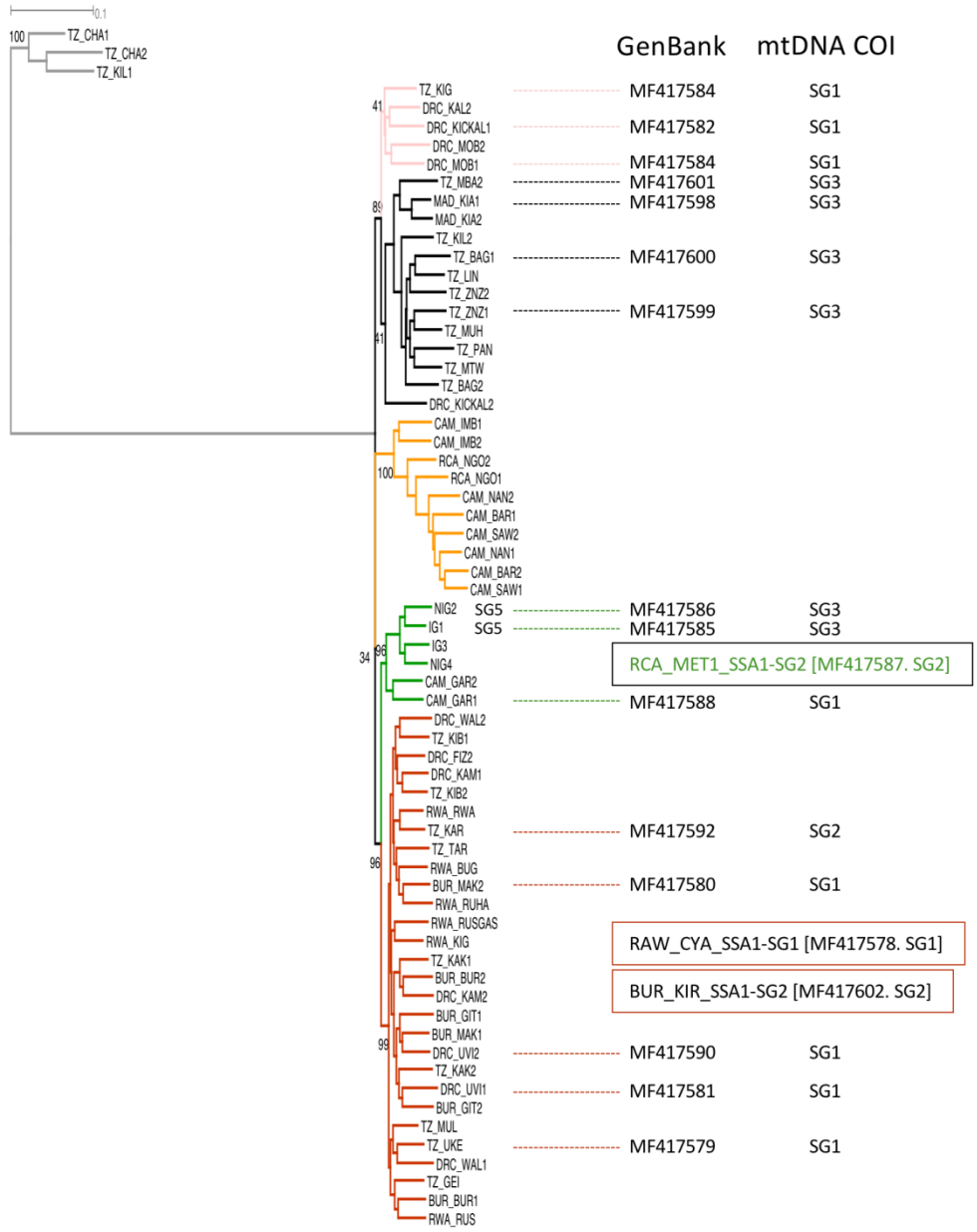
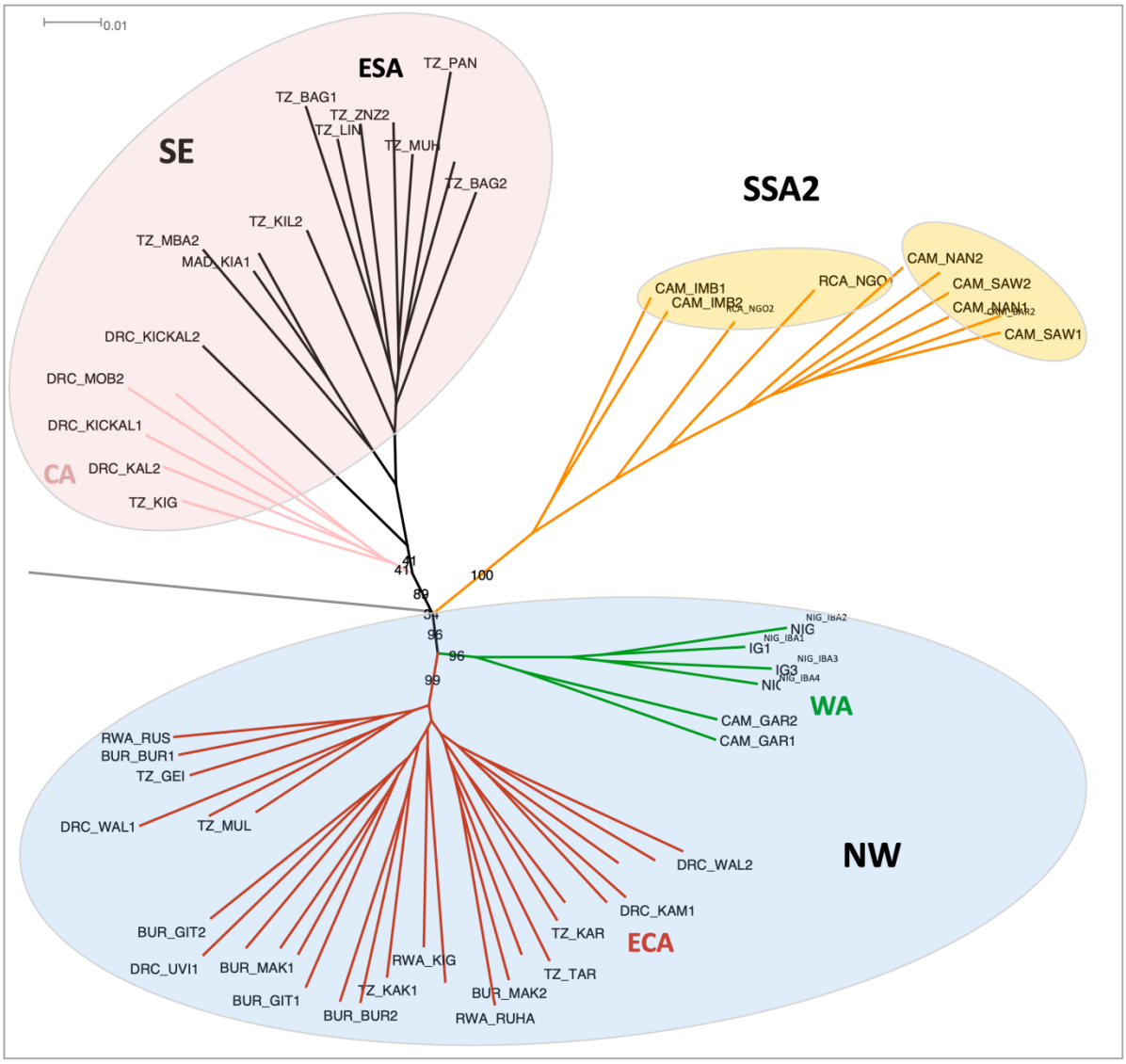

0.1
GenBank
mtDNA COI
TZ_CHA1
TZ_CHA2
TZ_KIL1
TZ_KIG
DRC_KAL2
DRC_KICKAL1
DRC_MOB2
DRC_MOB1
TZ_MBA2
MAD_KIA1
MAD_KIA2
TZ_KIL2
TZ_BAG1
TZ_LIN
TZ_ZNZ2
TZ_ZNZ1
TZ_MUH
TZ_PAN
TZ_MTW
TZ_BAG2
DRC_KICKAL2
CAM_IMB1
CAM_IMB2
RCA_NGO2
RCA_NGO1
CAM_NAN2
CAM_BAR1
CAM_SAW2
CAM_NAN1
CAM_BAR2
CAM_SAW1
NIG2 SG5
IG1 SG5
IG3
NIG4
CAM_GAR2
CAM_GAR1
DRC_WAL2
TZ_KIB1
DRC_FIZ2
DRC_KAM1
TZ_KIB2
RWA_RWA
TZ_KAR
TZ_TAR
RWA_BUG
BUR_MAK2
RWA_RUHA
RWA_RUSGAS
RWA_KIG
TZ_KAK1
BUR_BUR2
DRC_KAM2
BUR_GIT1
BUR_MAK1
DRC_UVI2
TZ_KAK2
DRC_UVI1
BUR_GIT2
TZ_MUL
TZ_UKE
DRC_WAL1
TZ_GEI
BUR_BUR1
RWA_RUS
MF417584 SG1
MF417582 SG1
MF417584 SG1
MF417601 SG3
MF417598 SG3
MF417600 SG3
MF417599 SG3
MF417586 SG3
MF417585 SG3
RCA_MET1_SSA1-SG2 [MF417587. SG2]
MF417588 SG1
MF417592 SG2
MF417580 SG1
RAW_CYA_SSA1-SG1 [MF417578. SG1]
BUR_KIR_SSA1-SG2 [MF417602. SG2]
MF417590 SG1
MF417581 SG1
MF417579 SG1

Fig. 2B

0.01
TZ_CHA2
TZ_CHA1
TZ_PAN
TZ_BAG1
TZ_LIN
TZ_BAG2
TZ_MBA2
CAM_IMB1
RCA_NGO1
CAM_BAR2
DRC_KICKAL2
DRC_MOB2
DRC_KICKAL1
DRC_KAL2
TZ_KIG
NIG2
IG3
CAM_GAR2
RWA_RUS
TZ_GEI
TZ_MUL
DRC_WAL2
DRC_KAM1
BUR_GIT2
BUR_MAK1
BUR_GIT1
RWA_KIG
TZ_TAR
RWA_RUHA

274

275 **Fig. 2C**



276

277

278 Admixture analysis identified three genetic clusters (Fig. 3) that separated the SSA2
279 species (highlighted in green) from the SSA1 species, which was again divided into two
280 distinct genetic backgrounds corresponding to SSA1-NW (containing SSA-WA and SSA-
281 ECA (highlighted in blue)) and SSA1-SE (containing SSA-CA and SSA-ESA (highlighted in
282 red)). Some SSA2 and all the SSA1-CA samples showed evidence for low levels of
283 admixture. Interestingly, F_{st} analysis (Table 1) supported the status of the two sub-groups

284 SSA1-NW and SSA-SE as intermediate at the species-level differentiation between any of
285 the SSA1 populations and the pooled SSA2 populations (the latter ranging from 0.28582 –
286 0.33406). This was particularly evident for SSA1-NW versus the SSA-ESA population of
287 SSA1-SE, with F_{st} of 0.22605 - 0.24304. F_{st} differentiation was less clear for the more
288 weakly defined SSA-CA population, whose F_{st} scores versus SSA1-NW were not
289 significantly different to those between populations within the two major SSA1 groups.
290 While the overall differentiation between SSA1-NW and SSA1-SE is consistent with them
291 being sub-species of SSA1, the observations concerning the SSA-CA population suggest
292 gene flow between CA and neighbouring populations from both sub-species.
293
294 **Fig. 3. Populations and species of *Bemisia* across sub-Saharan Africa.** Population
295 structure analysis using Admixture of 63 individuals from the sub-Saharan *Bemisia* cryptic
296 species clade based on 14,358 SNPs (K=3). Individuals were clustered into the SSA2 species
297 group, as well as two genetic clusters (red, blue) that belonged to SSA1 based on partial
298 mtCOI gene signature. The samples are shown on the map: green squares represent SSA-
299 WA, red squares represent SSA-ECA, pink squares represent SSA-CA and black squares
300 represent SSA-ECA individuals. Yellow and blue squares represent SSA2 and 'SSA4'
301 individuals, respectively (the latter being NUMT artefacts of SSA2 – see text). One SSA1
302 genetic cluster in the Admixture panel included the SSA-WA and SSA-ECA groups (35 blue
303 individuals and one red-blue individual), and the other SSA1 genetic cluster included five
304 SSA-CA (blue-red/blue-red-green) and 12 SSA-ESA (red) individuals. Evidence of gene
305 flow between SSA1 genetic clusters were predominantly detected in SSA-CA individuals
306 (four red-blue individuals in the Admixture analysis). Three inter-specific hybrids (two from
307 SSA2; one from SSA-CA with SSA1 (red-blue) and SSA2 (green) genetic backgrounds)
308 were also detected in the SSA *Bemisia* individuals analysed. The gene flow contact zone

between populations on the NW-SE transect is indicated by the thick blue dashed line. SSA1-NW and SSA1-SE sub-species boundary lines are indicative only. Landscape elevations across the sampling sites are shown in insert.

Fig. 3A

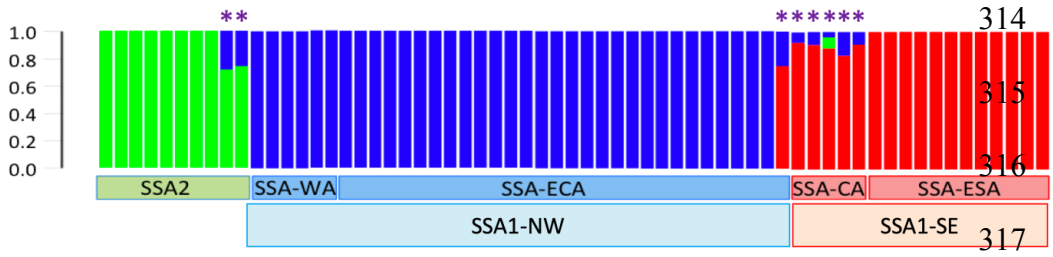


Fig. 3B

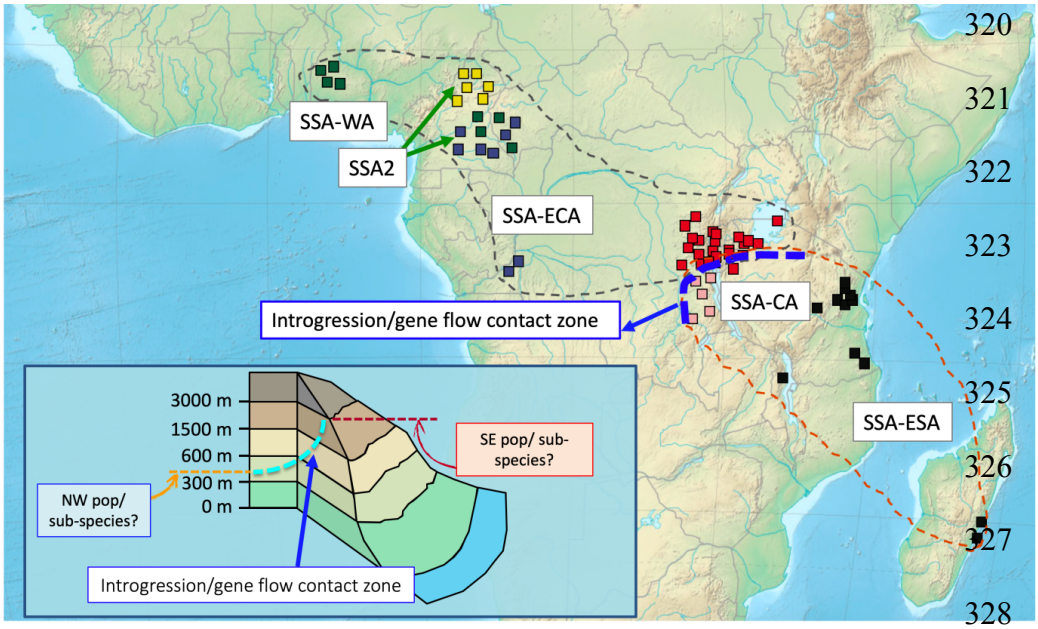


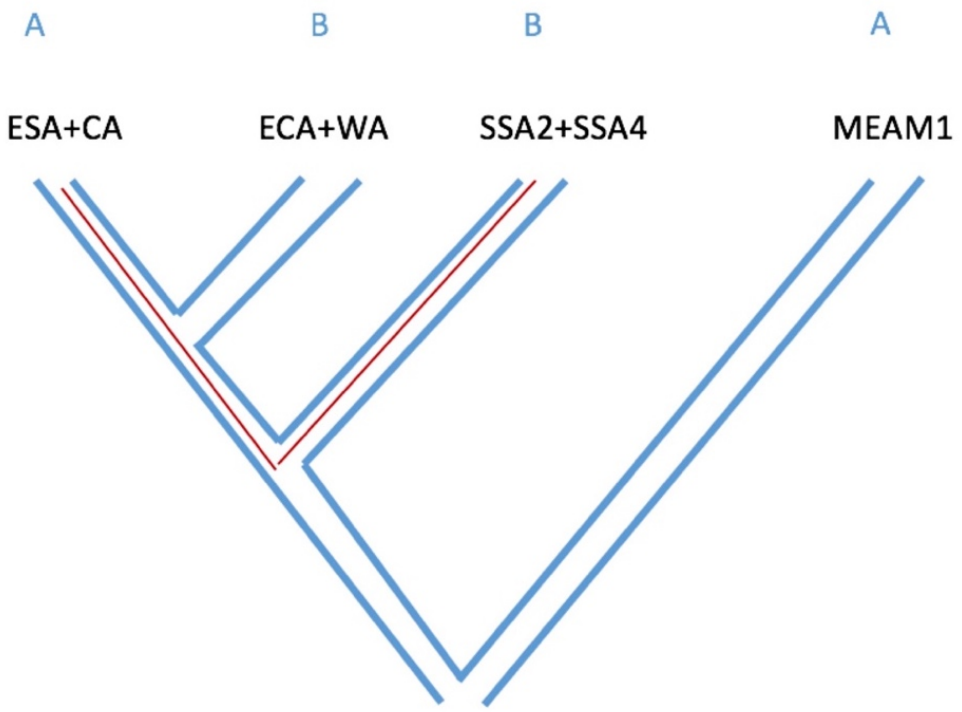
Table 1. Population substructure based on the *F*-statistics (F_{st}) of (Weir and Cockerham 1984). Treatment of SSA1 as single species against the SSA2/SSA4 species showed significant population substructure (F_{st} range 0.28 – 0.33). This limited gene flow between SSA2/SSA4 and the SSA1 (i.e., SSA-WA, SSA-ECA, SSA-CA. SSA-ESA) supported their

separate species status. Treatment of the four SSA1 populations as two well-defined groups, the proposed sub-species, showed that that comprising SSA-WA and SSA-ECA overall shared intermediate levels of gene flow with the SSA-ESA population. Gene flow at the contact zone (SSA-CA) is highest with the SSA-WA, SSA-ECA, and SSA-ESA genetic groups as evident by the low F_{st} values.

Species	**SSA2**	**SSSA1**			
Sub-species		**SSA1-NW**		**SSA1-SE**	
Populations	**SSA2/SSA4**	**SSA-WA**	**SSA-ECA**	**SSA-CA**	**SSA-ESA**
SSA2/4					
SSA-WA	0.29068				
SSA-ECA	0.33406	0.17048			
SSA-CA	0.28582	0.1658	0.1575		
SSA-ESA	0.33173	0.24304	0.22605	0.16026	

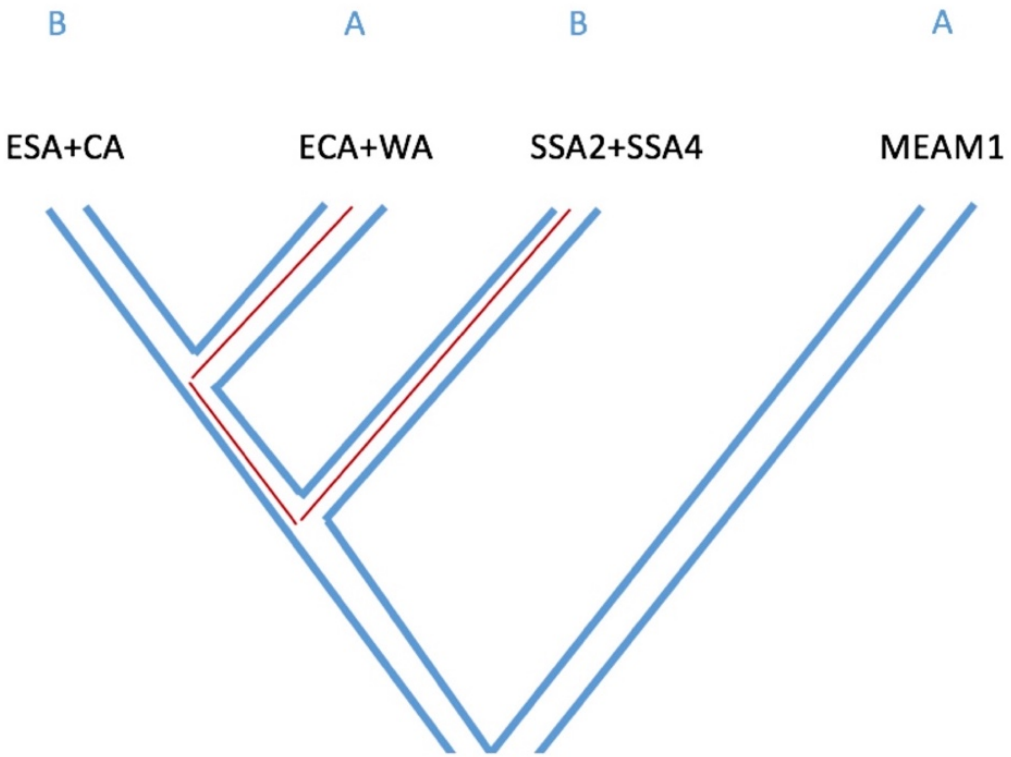
Within sub-species SSA1-NW, the SSA-WA population is predominantly found in the lowlands (*ca*. 300m - 1500m), whereas the SSA-ECA/SSA-CA populations were found at higher altitudes (from *ca*. 600m - ≥1500m). Individuals belonging to the SSA-ESA group were predominantly from coastal landscape of Tanzania and Madagascar (Fig. 3). A gene flow contact zone was detected at the SSA-CA region represented by individuals from DRC (n = 5), and one TZA individual from SSA-ECA group. Hybridization was detected in two SSA2 individuals (green/blue genetic backgrounds) and in one DRC individual from SSA-CA (red/green/blue in Fig. 3).

350 Results from our D-statistics test (ABBA-BABA test) between three groups
351 [ESA+CA], [SSA2+SSA4] and [ECA+WA] using MEAM1 as an outgroup identified strong
352 signals of introgression between the [ESA+CA] and [ECA+WA] groups which belonged to
353 the same SSA1 species based on mtCOI gene analysis, where the strong introgression signals
354 between [ESA+CA] and [ECA+WA] could potentially represent well-defined sub-
355 populations/sub-species occupying different geographical landscapes, with gene flow
356 occurring between these sub-populations/sub-species at the contact zone (Fig. 4).
357
358 **Fig. 4. Introgression site patterns using a four-Taxon ABBA/BABA test. A.** The ABBA
359 and BABA sites used in the test are derived from ESA+CA and ECA+WA belonging to
360 SSA1 species, compared to the SSA2+ SSA4 (**Fig. 4B**) group which are previously reported
361 as two distinct species but were identified as belonging to one species (SSA2) in this study.
362
363 **Fig. 4A**



364
365

366 **Fig. 4B**



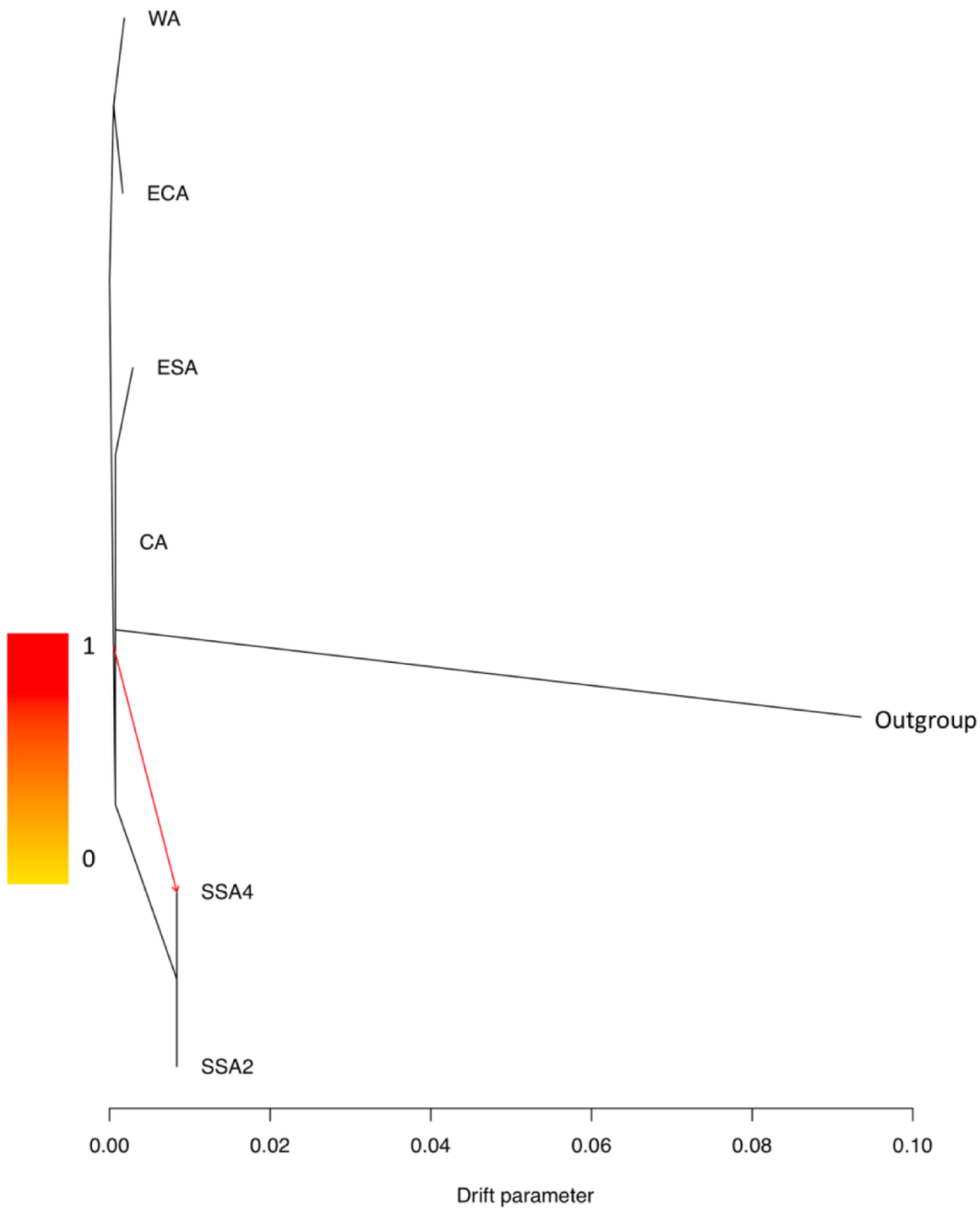
367

368

369 The test also detects signals of introgression within the same geographical region
370 between the SSA2 and SSA1 (i.e., [ECA+WA]) species. The same signatures of interspecies
371 admixture were also evident from the Treemix results (Fig. 5). The ML tree generated in
372 Treemix also supported the SSA4 species as part of SSA2. The populations/subspecies SSA-
373 ECA and SSA-WA cluster together which is consistent with their geographical origin (i.e.,
374 SSA1-NW), whereas the SSA-CA population clusters with the SSA-ECA population (i.e.,
375 representing the south-eastern geographic origin). Gene flow between these SSA1
376 populations is especially evident at the contact zone (i.e., at the SSA-CA geographical
377 location). The migration edges (m = 2) revealed gene flow between SSA2 and the SSA1
378 species from central Africa (i.e., SSA-CA) and supports detection of hybrid individuals
379 identified in the admixture analysis (Fig. 3).

380

381 **Fig. 5. Gene flow between SSA species.** TreeMix analysis showing gene flow from SSA1 to
382 SSA2/SSA4 (red arrow). Clustering between ECA and WA, and between ESA and CA
383 indicated close evolutionary relationship. SSA2 and SSA4 shared a genetic clade with no
384 branch depth indicating these two species were identical.
385



386
387

388 **Discussion**

389 We combined genome-wide SNPs and mtCOI markers to gain unprecedented
390 evolutionary ecological and population genomic insights into the sub-Saharan African
391 cassava whitefly species complex. By taking into consideration the impact of NUMTs, we
392 showed that there is no conflict between nuclear and mitochondrial DNA markers in terms of
393 defining species status, contrary to the study of Wosula et al. (2017) based on the same data
394 set. A thorough filtering of genome-wide SNP markers also discerned the evolutionary forces
395 shaping incipient speciation and admixture in the *Bemisia* SSA1 whitefly species. Various
396 evolutionary genetic studies based on the partial mtCOI gene have proposed the presence of
397 population subgroups (e.g., Ghosh, et al. 2018; Legg, et al. 2014) in the SSA1 species. Based
398 on genome-wide SNP markers, we found no evidence to support the subgroup classification,
399 although for the first time, the SSA1 species as a whole, could be broadly divided into two
400 well-defined north western (NW) and south eastern (SE) populations/sub-species, with a
401 contact zone separated by the eastern highland and central/western lowland regions. Evidence
402 for interspecies hybridisation was also detected between the SSA1 and SSA2 species, where
403 bidirectional genetic contributions were observed between SSA2 and SSA1, and will have
404 important implications to understanding the evolutionary genetics associated between plant
405 virus transmission capacity and the socioeconomically important SSA *Bemisia* whitefly pest
406 complex.

407 In the study of Elfekih et al. (2018a), mtCOI sequences from the MEAM1, MED,
408 Indian Ocean (IO) and Australia (AUS) *B. tabaci* cryptic species were interrogated prior to
409 generating nextRAD genome-wide SNP data to investigate the history of global introduction
410 pathways of the MED and MEAM1 species. Elfekih et al. (2018a), Elfekih et al. (2015),
411 Elfekih et al. (2016a), and Elfekih et al. (2016b) showed that species status of these cryptic *B.*
412 *tabaci* species as ascertained by the partial mtCOI gene was corroborated by genome-wide

413 SNP data. Similar to this study, genome-wide SNP markers also detected genomic signatures
414 of hybridization between the MED, MEAM1 and IO, while also revealing the role played by
415 NUMTs in confounding our effort to define species status such as between the MEAM1 and
416 MEAM2 species (Delatte, et al. 2005; Elfekih, et al. 2018a; Tay, et al. 2017a).

417 Wosula et al. (2017) adopted the nextRAD genome-wide sequencing method outlined
418 by Elfekih et al. (2015), Elfekih et al. (2016a), and Elfekih et al. (2016b) to investigate gene
419 flow patterns and diversity in the African cassava whitefly *Bemisia* SSA1, SSA2, SSA4
420 species complex, but overlooked the impact of NUMTs associated with the *B. tabaci* mtCOI
421 partial sequences. This oversight led to the major conclusion that the widely used mtCOI
422 gene was an unsuitable marker for differentiating between African *Bemisia* cryptic species.
423 Increasingly, reanalyses of the *B. tabaci* complex cryptic species partial mtCOI gene dataset
424 have detected the presence of significant NUMTs and pseudogenes (Kunz, et al. 2019b; Tay,
425 et al. 2017a; Vyskočilová, et al. 2018). These studies highlighted the challenges faced by
426 researchers to delimit species in the cryptic *B. tabaci* species complex, despite the promises
427 of molecular systematics and species delimitation via mitochondrial barcoding approaches.
428 Similar to MEAM2 (Boykin, et al. 2013; Dinsdale, et al. 2010), the *Bemisia* cryptic species
429 SSA4 (Berry, et al. 2004) was shown to be the product of NUMTs. This species
430 misidentification underpinned all subsequent interpretations of gene flow and evolutionary
431 ecology in the SSA *Bemisia* cassava whitefly species (Wosula, et al. 2017) based on genome-
432 wide SNP markers.

433 Significant genetic diversity in the *B. tabaci* species complex can obscure species and
434 sub-species identification. For instance, Elfekih et al. (2018a) detected significant diversity in
435 the MED species based on nextRAD SNP data; however, the MED-ASL species was not
436 identified until confirmed by complete mitogenome study, mating experiments (Vyskočilová,

437 et al. 2018), and host preference studies (Vyskočilová, et al. 2019). Similar studies are needed
438 to provide additional biological support for subgroup/sub-species status within SSA1.

439 In developing the DNA barcoding approach, Hebert et al. (2003) showed that the COI
440 gene is especially effective at differentiating at the species, order and phylum levels due to
441 the associated nucleotide substitutional rates for this gene, while for fine-scale identification
442 such as within species level, the use of genes with greater substitutional rates were
443 recommended. Subgroup classification within the SSA1 African cassava whitefly species
444 based on the mtDNA gene might be possible, provided that a gene region with an appropriate
445 substitutional rate is identified. Nevertheless, there is currently no evidence for maternal
446 lineage-based patterns to population dispersal behaviour, host adaptation, and fitness cost
447 differentiation in the *Bemisia tabaci* complex and the African cassava *Bemisia* species
448 complex, and therefore the subgroup delimitation based on the partial mtCOI gene is not
449 justified.

450 Previously, Legg et al. (2014) reported that the SSA1-SG1, representing the major
451 SSA1 mtCOI subgroup, was implicated in the severe CMD pandemic in East Africa. Based
452 on the subgroup classification, Wosula et al. (2017) reclassified these SSA1-SG1 individuals
453 into three SNP-based groups of SSA-ECA, SSA-CA, and SSA-WA due to perceived
454 inconsistencies (e.g., SSA1-SG3, SSA1-SG5; see Ghosh et al. (2015)) based solely on the
455 utilization of the short (i.e., *ca*. 657bp) mtCOI partial gene marker. Based on the same
456 dataset, our reanalysis found that while phylogenetic signals were detected for the separate
457 clustering of the 'WA' and 'ECA' populations, there was not a strong genetic difference
458 between the 'CA' and 'ESA' populations, and we therefore caution against the adoption of
459 the nomenclature as proposed by Wosula et al. (2017) in future studies. Furthermore,
460 admixture analysis also showed a substantial gene flow limitation between the 'WA/ECA'
461 groups and the 'CA/ESA' groups.

462 Although adopting the subgroup status could help resolve virus transmission capacity
463 and provide insights into understanding pest introduction pathways (Anderson, et al. 2018;
464 Tabachnick 1996), there is currently insufficient evidence to link either the SSA1-NW (i.e.,
465 'WA/ECA') or SSA1-SW (i.e., 'CA/ESA') populations/subgroups to historical CMD
466 outbreaks. Legg et al. (2002) described a distinct genetic cluster (*B. tabaci* Ug2) assumed to
467 be associated with the CMD epidemic in the 1990's, mainly due to its geographic distribution
468 in relation to the front of the epidemic. However, transmission leading to outbreaks of CMD
469 in the sub-Saharan Africa region is complex, with one factor being the sharing and
470 propagation of diseased cassava cuttings between farmers. There was a recombination event
471 between two geminiviruses that produced a more virulent hybrid (Colvin, et al. 2004),
472 however we have little historical information about how the viruses interacted with the
473 different *Bemisia* cryptic species in terms of disease dynamics (Polston, et al. 2014). The role
474 of the SSA2 species in the CMD epidemic is also unclear given that the SSA2's habitat range
475 (east and west Africa) overlaps the SSA1 species (Legg, et al. 2002; Mugerwa, et al. 2018;
476 Wosula, et al. 2017). The exact relationship between SSA1/SSA2 and outbreaks of CMD
477 (Manani, et al. 2017) should be re-investigated, taking into consideration the current
478 nextRAD SNP data as genomic evidence that supports the presence of hybrid SSA1/SSA2
479 individuals. Furthermore, the vector competencies of other common cassava whitefly species
480 (such as *Bemisia afer*) to different cassava viruses also needs to be fully understood.
481 Stringent delimitation of population groups across the full habitat range of the SSA1
482 species will be needed to ascertain whether the SSA1 as a whole is represented by genetically
483 distinct groups where limited gene flow could be further identified to external (e.g.,
484 geographic, climatic) and/or metagenomic (e.g., secondary symbionts) factors. The genetic
485 differentiation between the two SSA1 genetic groups suggests they can be considered as
486 subspecies of SSA1. Recognition of such biological classifications and their distribution in

487 sub-Saharan Africa will be important for management of these agriculturally and biosecurity
488 significant species, as for others posing a threat to global health and biosecurity (Paini, et al.
489 2016).

490 The South-Eastern SSA1 sub-populations belonging to Madagascar (SSA-ESA) and
491 Tanzania (SSA-CA) are geographically isolated from the North-Western sub-populations
492 sampled from Nigeria (SSA-WA), Democratic Republic of Congo and Burundi (SSA-ECA).
493 The mountain range separating these two SSA1 categories (SE and NW sub-populations)
494 poses a geographical barrier to gene flow, and representing a potential factor that could
495 promote incipient allopatric speciation. For instance, the East African Rift Valley is a major
496 geographical barrier providing the right environment for incipient species and recent
497 radiations to thrive and evolve (Gottelli, et al. 2004). Specifically, the western branch of the
498 Rift Valley, i.e., the Albertine Rift covering part of Tanzania, Burundi, Rwanda, Uganda and
499 Democratic Republic of Congo, is a chain of mountains situated at 1500–3500 meters above
500 sea level, and formed by volcanic activity that started during the plio-pleistocene. In addition
501 to the volcanic activity, climate change during this glacial period led to habitat fragmentation
502 and therefore species diversification in the newly formed biogeographical zones (Huhndorf,
503 et al. 2007). In addition to allopatric forces, species diversification in the whitefly cryptic
504 species system has been explained by host plant associations and the ability of this system to
505 feed on a wide range of host plants including acquiring new ones (Matsubayashi, et al. 2010).
506 Using RNAseq, Malka et al. (2018) provided insights into how the variation in level of
507 expression of detoxification genes might have conferred plasticity to the complex and that
508 likely also contributed to its diversification.

509 *B. tabaci sensu lato* on cassava in sub-Saharan Africa is a species complex going
510 through a continuous diversification process. The examination of the genetic diversity via
511 mitochondrial and genome-wide markers validated discrimination between SSA1 and SSA2,

512 and the latter being the same as SSA4. This was mainly facilitated by careful data analysis,
513 and taking into consideration artefacts caused by pseudogenes (NUMTs). The SSA1 species
514 shows signatures of incipient speciation with sub-populations divided into two
515 biogeographical niches; North Western (NW) and South Eastern (SE) regions with a hybrid
516 zone between Tanzania, Uganda and Democratic Republic of Congo. We conclude that
517 accurate species identification that takes into consideration all the evolutionary forces driving
518 the diversification in this system will be critical in the implementation of any future
519 management strategy for this invasive and economically important insect pest complex.
520 Further comparative analyses on the endosymbiont metacommunities and host association
521 can shed more light onto the major causes underlying the gradual process of species
522 diversification and adaptive radiations in the cassava whitefly system.
523

524 **Acknowledgements**

525 SE was funded by the CSIRO Office of the Chief Executive (OCE) post-doctoral fellowship
526 (R-4546-1) and a European Molecular Biology Organization (EMBO) fellowship ASTF-
527 6889. WTT, SE, KG DK and TW were supported by CSIRO H&B Genes of Biosecurity
528 Importance fund (R-8681-1). SM, AP, WTT, SV, JC, PDB were supported by the Natural
529 Resources Institute, University of Greenwich from a grant provided by the Bill & Melinda
530 Gates Foundation (Grant Agreement OPP1058938).
531

532 **Literature cited**

533 Alexander DH, Novembre J, Lange K 2009. Fast model-based estimation of ancestry in
534 unrelated individuals. Genome Research 19: 1655-1664. doi: 10.1101/gr.094052.109
535 Anderson CJ, et al. 2018. Hybridization and gene flow in the mega-pest lineage of moth,
536 *Helicoverpa*. Proc Natl Acad Sci U S A 115: 5034-5039. doi: 10.1073/pnas.1718831115

537 Anderson CJ, Tay WT, McGaughran A, Gordon K, Walsh TK 2016. Population structure and
538 gene flow in the global pest, *Helicoverpa armigera*. Mol Ecol 25: 5296-5311. doi:
539 10.1111/mec.13841
540 Behere GT, Tay WT, Russell DA, Batterham P 2008. Molecular markers to discriminate
541 among four pest species of *Helicoverpa* (Lepidoptera: Noctuidae). Bull Entomol Res 98:
542 599-603. doi: 10.1017/S0007485308005956
543 Berry SD, et al. 2004. Molecular evidence for five distinct *Bemisia tabaci* (Homoptera :
544 Aleyrodidae) geographic haplotypes associated with cassava plants in sub-Saharan
545 Africa. Annals of the Entomological Society of America 97: 852-859. doi: Doi
546 10.1603/0013-8746(2004)097[0852:Meffdb]2.0.Co;2
547 Bolger AM, Lohse M, Usadel B 2014. Trimmomatic: a flexible trimmer for Illumina
548 sequence data. Bioinformatics 30: 2114-2120. doi: 10.1093/bioinformatics/btu170
549 Boykin LM, Bell CD, Evans G, Small I, De Barro PJ 2013. Is agriculture driving the
550 diversification of the *Bemisia tabaci* species complex (Hemiptera: Sternorrhyncha:
551 Aleyrodidae)?: Dating, diversification and biogeographic evidence revealed. BMC Evol
552 Biol 13: 228. doi: 10.1186/1471-2148-13-228
553 Boykin LM, Savill A, De Barro P 2017. Updated mtCOI reference dataset for the *Bemisia*
554 *tabaci* species complex F1000Res 6: 1835. doi: 10.12688/f1000research.12858.1
555 Chen WB, et al. 2016. The draft genome of whitefly *Bemisia tabaci* MEAM1, a global crop
556 pest, provides novel insights into virus transmission, host adaptation, and insecticide
557 resistance. BMC Biology 14: 110. doi: 10.1186/s12915-016-0321-y
558 Colvin J, Omongo CA, Maruthi MN, Otim-Nape GW, Thresh JM 2004. Dual begomovirus
559 infections and high *Bemisia tabaci* populations: two factors driving the spread of a
560 cassava mosaic disease pandemic. Plant Pathology 53: 577-584. doi: 10.1111/j.1365-
561 3059.2004.01062.x

562 De Barro PJ, Liu SS, Boykin LM, Dinsdale AB 2011. *Bemisia tabaci*: a statement of species
563 status. Annu Rev Entomol 56: 1-19. doi: 10.1146/annurev-ento-112408-085504
564 Delatte H, et al. 2005. A new silverleaf-inducing biotype Ms of *Bemisia tabaci* (Hemiptera :
565 Aleyrodidae) indigenous to the islands of the south-west Indian Ocean. Bulletin of
566 Entomological Research 95: 29-35. doi: 10.1079/Ber2004337
567 Dinsdale A, Cook L, Riginos C, Buckley YM, De Barro P 2010. Refined Global Analysis of
568 *Bemisia tabaci* (Hemiptera: Sternorrhyncha: Aleyrodoidea: Aleyrodidae) Mitochondrial
569 Cytochrome Oxidase 1 to Identify Species Level Genetic Boundaries. Annals of the
570 Entomological Society of America 103: 196-208. doi: 10.1603/An09061
571 Eaton DAR 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses.
572 Bioinformatics 30: 1844-1849. doi: 10.1093/bioinformatics/btu121
573 Elfekih S, et al. 2018a. Genome-wide analyses of the *Bemisia tabaci* species complex reveal
574 contrasting patterns of admixture and complex demographic histories. PLoS One 13:
575 e0190555. doi: 10.1371/journal.pone.0190555
576 Elfekih S, et al. 2015. Evolutionary genomics of *Bemisia tabaci* and characterization of its
577 endosymbiont metacommunities using nextRAD sequencing. International Plant and
578 Animal Genome Asia 2015 July 12-15th, 2015: Singapore.
579 Elfekih S, et al. 2016a. Genome-Wide SNPs Decipher Global Incursion Pathways in the
580 *Bemisia tabaci* Species Complex. International Plant and Animal Genome Conferences.
581 2016a 9-13 January 2016: San Diego, CA, U. S. A.
582 Elfekih S, et al. 2016b. Genome-wide scans unravel fine-scale invasion routes in the *Bemisia*
583 *tabaci* species complex. 2nd International Whitefly Symposium, . 2016b 14-19 February
584 2016 Arusha, Tanzania. p38.

585 Elfekih S, Tay WT, Gordon K, Court LN, De Barro PJ 2018b. Standardized molecular
586 diagnostic tool for the identification of cryptic species within the *Bemisia tabaci*
587 complex. Pest Manag Sci 74: 170-173. doi: 10.1002/ps.4676
588 FAOSTAT 2017 [cited 2018 Jan-2018]. Available from: http://www.fao.org/faostat/en/ -
589 data/QC/visualize
590 Fumagalli M, et al. 2013. Quantifying Population Genetic Differentiation from Next-
591 Generation Sequencing Data. Genetics 195: 979-+. doi: 10.1534/genetics.113.154740
592 Ghosh S, Bouvaine S, Maruthi MN 2015. Prevalence and genetic diversity of endosymbiotic
593 bacteria infecting cassava whiteflies in Africa. BMC Microbiology 15: 93. doi:
594 10.1186/s12866-015-0425-5
595 Ghosh S, Bouvaine S, Richardson SCW, Ghanim M, Maruthi MN 2018. Fitness costs
596 associated with infections of secondary endosymbionts in the cassava whitefly species
597 *Bemisia tabaci*. Journal of Pest Science 91: 17-28. doi: 10.1007/s10340-017-0910-8
598 Gottelli D, Marino J, Sillero-Zubiri C, Funk SM 2004. The effect of the last glacial age on
599 speciation and population genetic structure of the endangered Ethiopian wolf (*Canis*
600 *simensis*). Molecular Ecology 13: 2275-2286. doi: 10.1111/j.1365-294X.2004.02226.x
601 Hadjistylli M, Roderick GK, Gauthier N 2015. First report of the Sub-Saharan Africa 2
602 species of the *Bemisia tabaci* complex in the Southern France. Phytoparasitica 43: 679-
603 687. doi: 10.1007/s12600-015-0480-3
604 Hebert PDN, Cywinska A, Ball SL, DeWaard JR 2003. Biological identifications through
605 DNA barcodes. Proceedings of the Royal Society B-Biological Sciences 270: 313-321.
606 doi: 10.1098/rspb.2002.2218
607 Huhndorf MH, Peterhans JCK, Loew SS 2007. Comparative phylogeography of three
608 endemic rodents from the Albertine Rift, east central Africa. Molecular Ecology 16: 663-
609 674. doi: 10.1111/j.1365-294X.2007.03153.x

610 Katoh K, Standley DM 2013. MAFFT multiple sequence alignment software version 7:
611 improvements in performance and usability. Mol Biol Evol 30: 772-780. doi:
612 10.1093/molbev/mst010
613 Korneliussen TS, Albrechtsen A, Nielsen R 2014. ANGSD: Analysis of Next Generation
614 Sequencing Data. Bmc Bioinformatics 15: 356. doi: 10.1186/s12859-014-0356-4
615 Kunz D, et al. 2019a. Draft mitochondrial DNA genome of a 1920 Barbados cryptic *Bemisia*
616 *tabaci* 'New World' species (Hemiptera: Aleyrodidae). Mitochondrial DNA Part B-
617 Resources 4: 1183-1184. doi: 10.1080/23802359.2019.1591197
618 Kunz D, Tay WT, Elfekih S, Gordon KHJ, De Barro PJ 2019b. Take out the rubbish -
619 Removing NUMTs and pseudogenes from the *Bemisia tabaci* cryptic species mtCOI
620 database. Sci Rep Submitted.
621 Lee W, Park J, Lee GS, Lee S, Akimoto S 2013. Taxonomic status of the *Bemisia tabaci*
622 complex (Hemiptera: Aleyrodidae) and reassessment of the number of its constituent
623 species. PLoS One 8: e63817. doi: 10.1371/journal.pone.0063817
624 Legg JP, French R, Rogan D, Okao-Okuja G, Brown JK 2002. A distinct *Bemisia tabaci*
625 (Gennadius) (Hemiptera : Sternorrhyncha : Aleyrodidae) genotype cluster is associated
626 with the epidemic of severe cassava mosaic virus disease in Uganda. Molecular Ecology
627 11: 1219-1229. doi: DOI 10.1046/j.1365-294X.2002.01514.x
628 Legg JP, et al. 2014. Spatio-temporal patterns of genetic change amongst populations of
629 cassava *Bemisia tabaci* whiteflies driving virus pandemics in East and Central Africa.
630 Virus Res 186: 61-75. doi: 10.1016/j.virusres.2013.11.018
631 Li H, Durbin R 2010. Fast and accurate long-read alignment with Burrows-Wheeler
632 transform. Bioinformatics 26: 589-595. doi: 10.1093/bioinformatics/btp698
633 Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:
634 2078-2079. doi: 10.1093/bioinformatics/btp352

635 Liu SS, et al. 2007. Asymmetric mating interactions drive widespread invasion and
636 displacement in a whitefly. Science 318: 1769-1772. doi: 10.1126/science.1149887
637 Macfadyen S, et al. 2018. Cassava whitefly, *Bemisia tabaci* (Gennadius) (Hemiptera:
638 Aleyrodidae) in East African farming landscapes: a review of the factors determining
639 abundance. Bull Entomol Res 108: 565-582. doi: 10.1017/S0007485318000032
640 Malka O, et al. 2018. Species-complex diversification and host-plant associations in *Bemisia*
641 *tabaci*: A plant-defence, detoxification perspective revealed by RNA-Seq analyses.
642 Molecular Ecology 27: 4241-4256. doi: 10.1111/mec.14865
643 Manani DM, Ateka EM, Nyanjom SRG, Boykin LM 2017. Phylogenetic Relationships
644 among Whiteflies in the *Bemisia tabaci* (Gennadius) Species Complex from Major
645 Cassava Growing Areas in Kenya. Insects 8: 25. doi: 10.3390/insects8010025
646 Martin JH, Mound LA 2007. An annotated check list of the world's whiteflies (Insecta :
647 Hemiptera : Aleyrodidae). Zootaxa: 1-84.
648 Matsubayashi KW, Ohshima I, Nosil P 2010. Ecological speciation in phytophagous insects.
649 Entomologia Experimentalis Et Applicata 134: 1-27. doi: 10.1111/j.1570-
650 7458.2009.00916.x
651 Minato N, et al. 2019. Surveillance for Sri Lankan cassava mosaic virus (SLCMV) in
652 Cambodia and Vietnam one year after its initial detection in a single plantation in 2015.
653 PLoS One 14: e0212780. doi: 10.1371/journal.pone.0212780
654 Mugerwa H, et al. 2018. African ancestry of New World, *Bemisia tabaci*-whitefly species.
655 Sci Rep 8: 2734. doi: 10.1038/s41598-018-20956-3
656 Paini DR, et al. 2016. Global threat to agriculture from invasive species. Proceedings of the
657 National Academy of Sciences of the United States of America 113: 7575-7579. doi:
658 10.1073/pnas.1602205113

659 Patil BL, Fauquet CM 2009. Cassava mosaic geminiviruses: actual knowledge and
660 perspectives. Mol Plant Pathol 10: 685-701. doi: 10.1111/j.1364-3703.2009.00559.x
661 Pickrell JK, Pritchard JK 2012. Inference of Population Splits and Mixtures from Genome-
662 Wide Allele Frequency Data. PLoS Genetics 8: e1002967. doi:
663 10.1371/journal.pgen.1002967
664 Polston JE, De Barro P, Boykin LM 2014. Transmission specificities of plant viruses with the
665 newly identified species of the *Bemisia tabaci* species complex. Pest Management
666 Science 70: 1547-1552. doi: 10.1002/ps.3738
667 Reich D, Thangaraj K, Patterson N, Price AL, Singh L 2009. Reconstructing Indian
668 population history. Nature 461: 489-U450. doi: 10.1038/nature08365
669 Stamatakis A 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses
670 with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690. doi:
671 10.1093/bioinformatics/btl446
672 Tabachnick WJ 1996. *Culicoides variipennis* and bluetongue-virus epidemiology in the
673 United States. Annual Review of Entomology 41: 23-43. doi: DOI
674 10.1146/annurev.en.41.010196.000323
675 Tay WT, et al. 2017a. The Trouble with MEAM2: Implications of Pseudogenes on Species
676 Delimitation in the Globally Invasive *Bemisia tabaci* (Hemiptera: Aleyrodidae) Cryptic
677 Species Complex. Genome Biology and Evolution 9: 2732-2738. doi:
678 10.1093/gbe/evx173
679 Tay WT, et al. 2017b. Novel molecular approach to define pest species status and tritrophic
680 interactions from historical *Bemisia* specimens. Scientific Reports 7: 429. Doi:
681 10.1038/s41598-017-00528-7
682 Tay WT, Evans GA, Boykin LM, De Barro PJ 2012. Will the Real *Bemisia tabaci* Please
683 Stand Up? PLoS One 7: e50550. DOI: 10.1371/journal.pone.0050550

684 Vyskočilová S, Seal S, Colvin J 2019. Relative polyphagy of "Mediterranean" cryptic
685 *Bemisia tabaci* whitefly species and global pest status implications. Journal of Pest
686 Science 92: 1071-1088. doi: 10.1007/s10340-019-01113-9
687 Vyskočilová S, Tay WT, van Brunschot S, Seal S, Colvin J 2018. An integrative approach to
688 discovering cryptic species within the *Bemisia tabaci* whitefly species complex. Sci Rep
689 8: 10886. doi: 10.1038/s41598-018-29305-w
690 Walsh TK, et al. 2019. Mitochondrial DNA genomes of five major *Helicoverpa* pest species
691 from the Old and New Worlds (Lepidoptera: Noctuidae). Ecol Evol 9: 2933-2944. doi:
692 10.1002/ece3.4971
693 Wang HL, et al. 2016. First Report of Sri Lankan cassava mosaic virus Infecting Cassava in
694 Cambodia. Plant Disease 100: 1029-1029. doi: 10.1094/Pdis-10-15-1228-Pdn
695 Weir BS, Cockerham CC 1984. Estimating F-Statistics for the Analysis of Population-
696 Structure. Evol. 38: 1358-1370. doi: Doi 10.2307/2408641
697 Wosula EN, Chen WB, Fei ZJ, Legg JP 2017. Unravelling the Genetic Diversity among
698 Cassava *Bemisia tabaci* Whiteflies Using NextRAD Sequencing. Genome Biology and
699 Evolution 9: 2958-2973. doi: 10.1093/gbe/evx219
700 Zheng XW, et al. 2012. A high-performance computing toolset for relatedness and principal
701 component analysis of SNP data. Bioinformatics 28: 3326-3328. doi:
702 10.1093/bioinformatics/bts606
703