



OPEN

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

S. Elfekih<sup>1,7</sup>, W. T. Tay<sup>2,7</sup>✉, A. Polaszek<sup>3</sup>, K. H. J. Gordon<sup>2</sup>, D. Kunz<sup>4</sup>, S. Macfadyen<sup>2</sup>, T. K. Walsh<sup>2</sup>, S. Vyskočilová<sup>5</sup>, J. Colvin<sup>5</sup> & P. J. De Barro<sup>6</sup>

The *Bemisia* cassava whitefly complex includes species that cause severe crop damage through vectoring cassava viruses in eastern Africa. Currently, this whitefly complex is divided into species and subgroups (SG) based on very limited molecular markers that do not allow clear definition of species and population structure. Based on 14,358 genome-wide SNPs from 62 *Bemisia* cassava whitefly individuals belonging to sub-Saharan African species (SSA1, SSA2 and SSA4), 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 nested within SSA2, and that populations of the SSA1 species comprise well-defined south-eastern (Madagascar, Tanzania) and north-western (Nigeria, Democratic Republic of Congo, Burundi) putative sub-species. Signatures of allopatric incipient speciation, and the presence of a 'hybrid zone' separating the two putative sub-species were also detected. These findings provide insights into the evolution and molecular ecology of a highly cryptic hemipteran insect complex in African, and allow the systematic use of genomic data to be incorporated in the development of management strategies for this cassava pest.

Confidence in species identification provides insights into the evolutionary forces that drive species diversification, and it underpins the strategies for biological conservation and the management of natural resources. While morphological differences have enabled many species to be identified, cryptic species complexes present a particular challenge due to their general lack of readily distinguishable morphological characters<sup>1–3</sup>. Ecological and life history data including host plant use and mating behaviour are needed to mitigate and address these challenges<sup>3–5</sup>. This is often time-consuming and impractical especially with large numbers of samples across wide geographic range. Incorporating DNA markers, e.g., diagnostic mitochondrial DNA (mtDNA) markers<sup>6–8</sup>, genome-wide sequencing<sup>9,10</sup>, and whole-genome resequencing<sup>11</sup> can increase confidence in differentiating cryptic species in a complex and identifying interspecific hybrids.

The challenges can become critical when individual cryptic species in a complex have a serious impact on plant health or agricultural productivity: insect vectors of plant viral diseases affecting cassava (*Manihot esculenta*), the most important food crop grown in Africa<sup>12</sup>, is one example of such case. Cassava production is vulnerable to cassava mosaic disease (CMD) caused by a number of cassava mosaic geminiviruses (CMGs)<sup>13,14</sup> and CMD has been reported in Africa since the early 1970s<sup>15</sup> and more recently in South-East Asia<sup>16,17</sup>. While the vector responsible for the spread of the CMD in South-East Asia is not well-understood, the pandemic in east and central Africa was associated with the distribution of sub-Saharan African (SSA) species of the *Bemisia* whitefly complex<sup>13</sup>. Invasive whitefly species such as the Mediterranean (MED), Middle East-Asia Minor 1 (MEAM1), and SSA2, have become pests of global agriculture, spreading plant viral diseases and becoming resistant to chemical control agents<sup>18,19</sup>. In particular, the endemic African cassava *Bemisia* whitefly SSA2 species occupies regions affected by the CMD pandemic. SSA2 has also expanded into southern Europe since 2007<sup>20</sup> whereas SSA1 has not (though widespread throughout the African continent). Defining the actual CMD vectors in Africa remains a challenge due to the cryptic nature of the many species in the Aleyrodidae whitefly genus *Bemisia*<sup>3,18,21,22</sup>.

<sup>1</sup>Australian Centre for Disease Preparedness, CSIRO, Geelong, VIC, Australia. <sup>2</sup>Black Mountain Laboratories, CSIRO, Canberra, ACT, Australia. <sup>3</sup>Department of Life Sciences, Natural History Museum, London, UK. <sup>4</sup>The Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK. <sup>5</sup>Natural Resources Institute, University of Greenwich, Kent, UK. <sup>6</sup>CSIRO, Ecosciences Precinct, Brisbane, Australia. <sup>7</sup>These authors contributed equally: S. Elfekih and W.T. Tay. ✉email: weetek.tay@csiro.au

Historically, the lack of distinguishable morphological characters among *Bemisia tabaci* has resulted in geographically distinct species being reclassified as a single species<sup>23–25</sup> and underpinned the *B. tabaci* taxonomic confusion (e.g., review by De Barro et al.<sup>18</sup>; see also Tay et al.<sup>26,27</sup>). Currently, over 38 cryptic species are recognised within the *B. tabaci* (Gennadius) species complex<sup>24</sup> through molecular diagnostics of a partial (i.e., 657 bp) mitochondrial DNA cytochrome oxidase subunit I (mtCOI)<sup>18,28,29</sup>, mating experiments<sup>5,30,31</sup> (although low numbers of F1 hybrids<sup>10,32</sup> including between SSA1 and SSA2 (previously ‘Ug1’ and ‘Ug2’)<sup>33,34</sup> are known), behavioural<sup>4</sup>, and host preference<sup>3</sup> studies.

Re-analyses of partial mtCOI sequences relating to the MED species<sup>3</sup>, and of the *B. tabaci* standard partial mtCOI dataset<sup>33</sup> demonstrated that ca. 60–65% were nuclear mitochondrial DNA sequences (NUMTs) or other PCR artefacts<sup>10,29,35</sup>. These analyses, together with molecular characterisation of original voucher specimens, have begun to disentangle the species complex<sup>26,27</sup>, identified historical species ranges<sup>26,27,36</sup>, and provided phylogenetic evidence that the ‘*B. tabaci* cryptic species complex’ is a polyphyletic group derived from various SSA *Bemisia* whitefly species<sup>29</sup>. Delimitation of populations across the full African *Bemisia* species habitat range will be needed. This would help to ascertain whether any of these species, e.g., SSA1, is genetically distinct with limited gene flow from others either due to external (e.g., geographic or climatic limitation) or internal (e.g., presence/absence of secondary symbionts) factors including host association. Recognition of such taxonomic classifications and their distribution in sub-Saharan Africa will be important for management of these pest species that could pose a threat to plant health and global agricultural biosecurity<sup>37</sup>.

The advent of high-throughput sequencing (HTS) technologies has driven the availability of genome-wide data that is further transforming the *Bemisia* taxonomy. HTS has demonstrated that various species previously recognised via characterisation of mtCOI (e.g., SSA4<sup>29</sup>; Middle East Asia Minor 2 (MEAM2)<sup>10,29,35</sup>) were likely artefacts caused by the sub-optimal primers<sup>16,17</sup> and sequencing of NUMTs<sup>10,35</sup>. Analyses of mitochondrial genomes also demonstrated inconsistencies of *B. tabaci* cryptic species identification based on the standard partial mtCOI gene region, and cautioned the need to combine behavioural and molecular evidence to delimitate species status<sup>3,31</sup>.

Despite the recent advances and their importance in the African agricultural landscape, the taxonomy, population structure, and evolutionary genetics of the African cassava *Bemisia* whitefly species complex (i.e., SSA1, SSA2, SSA3, SSA6, and SSA9)<sup>29</sup> remain poorly understood. A recent study of taxonomic structure and gene flow among genetically diverse SSA1, SSA2, SSA3, and SSA4 species<sup>38</sup> described serious discrepancies between their results obtained from mtCOI and from nextRAD sequencing. This (the conflicting results of mtCOI and nextRAD sequencing) findings has strongly implies that the species delimitation of sub-Saharan African *Bemisia* species needs to be revisited since it is currently based on mtCOI. We seek to show why mtCO and nuclear DNA should result in such conflicting implications for the African cassava *Bemisia* whitefly SSA1 (which is also further refined into subgroups (SG) 1, 2, and 3 in, e.g., Legg et al.<sup>13</sup>), SSA2, and SSA4 species. With the cassava *Bemisia* whitefly being the most common whitefly on cassava crops in Africa, and capable of vectoring diseases (e.g., cassava mosaic begomoviruses, CMBs; cassava brown streak ipomoviruses, CBSIs) known to cause significant crop losses<sup>39</sup>, our findings will also provide insights into incipient speciation of this agriculturally important *Bemisia* species complex with the capacity to impact on food security in sub-Saharan Africa<sup>40,41</sup>.

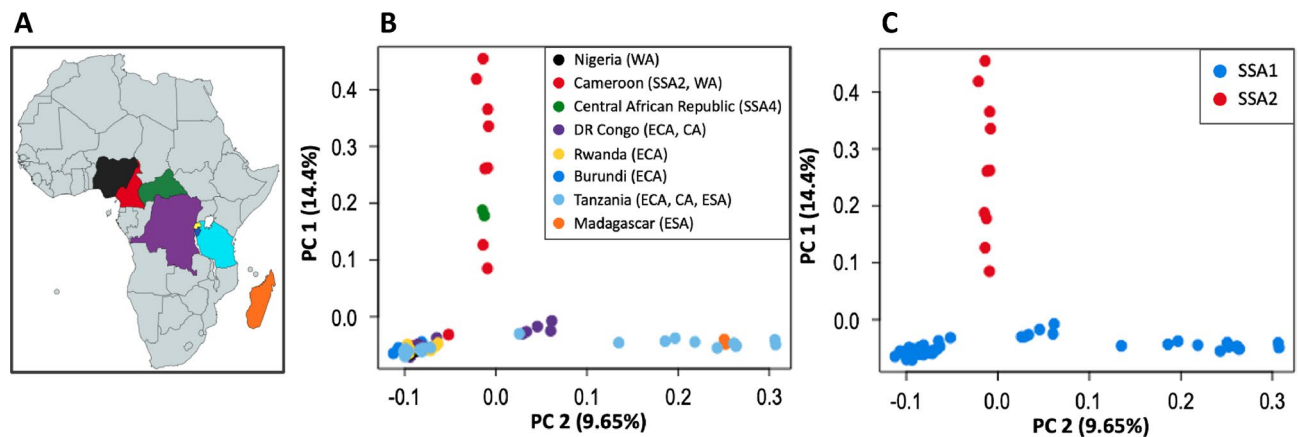
## Results

Wosula et al.<sup>38</sup> reported a total of 25 partial mtCOI sequences from their 72 SSA samples studied. Of these 25 sequences, only 11 represented unique (non-redundant) mtCOI haplotypes [seven SSA1: MF417579, MF417581, MF417585, MF417587, MF417590, MF417592, MF417601; one SSA3: MF417591; one SSA2: MF417597; two ‘putative SSA4’ MF417589, MF417593 (with ca. 94% nucleotide identity to SSA3)].

**Integrity of African cassava *Bemisia* SSA mtCOI sequences.** Berry et al.<sup>42</sup> reported 11 ‘Sub-Saharan III: Western Africa-Cameroon/Cassava’ sequences which were renamed as SSA4<sup>18</sup>. Of these 11 SSA4 sequences, only two short sequences (i.e., Cam Ayos 1 WO2, Cam Ayos 2 WO3) did not have INDELS or exhibit any premature stop codons at the amino acid level (Suppl. Fig. 1). The remaining nine sequences had random INDELS that resulted in frame-shift mutations and premature stop codons, which suggested that they are likely to be NUMTs. The Cam Ayos 2 WO3 sequence also had nucleotide substitutions that resulted in amino acid residue changes in highly conserved mtCOI regions, suggesting that this was likely also a NUMT sequence. The remaining single, short but ‘functional’ Cam Ayos 1 WO2 sequence (AF344246) that phylogenetically grouped<sup>42</sup> with the 10 NUMT-related SSA4 sequences suggested that as a whole, these SSA4 sequences originally reported by Berry et al.<sup>42</sup> were all likely NUMTs. The fact that SSA4 sequences were phylogenetically clustered to SSA2 suggested they were also potentially of nuclear origin, similar to that reported for *B. tabaci* MEAM2 NUMT<sup>10,35</sup> that originated from the nuclear genome of MEAM1<sup>10,35</sup>.

Further evidence that the partial mtCOI gene lacks the power to differentiate between evolutionary very closely related cryptic species (e.g., the ‘subgroups’ (SG)<sup>13,43</sup> taxonomic status within SSA1) can be seen by examining nucleotide substitution patterns. For example, re-assessment of the two SSA1-SG5 sequences (MF417585, MF417586) showed that they differed from the SSA1-SG3 sequence JQ286482 by only three C/T substitutions that resulted in 0.56% nucleotide difference (Suppl. Fig. 2A). The basis for suggesting SSA1-SG5 and SSA1-SG3 to be different ‘subgroups’ within the SSA1 species was therefore unclear. For example, sequences with similar (Suppl. Fig. 2B) or with greater (Suppl. Fig. 2C) numbers of nucleotide substitutions were placed within same subgroups, or potentially as the same species<sup>31</sup> (Suppl. Fig. 2D).

**Species delimitation and introgression.** The genome-wide SNP-based Principal Component Analysis (PCA) discriminated between SSA1 and SSA2 sampled from eight sub-Saharan African nations (Fig. 1A), sug-



**Figure 1.** Sampling sites and principal component analysis of SSA *Bemisia* samples. Principal component analysis (PCA) was based on 14,358 genome-wide SNPs across the 62 sub-Saharan African cassava *Bemisia* samples from eight sub-Saharan countries. (A) African countries where whitefly individuals were sampled from. Country names and colours as (B). African map created using mapchart <mapchart.net>. (B) Country of origins for these SSA1 and SSA2 *Bemisia* species are shown by unique colour codes. Parentheses after country names are the ‘genetic groups’ as proposed by Wosula et al.<sup>38</sup>. (C) PCA showing clustering of individuals by their SSA1 or SSA2 species status from this study’s re-analyses of genome-wide SNPs.

gesting that SSA1 and SSA2 are genetically distinct and likely to be different species as supported by mating studies<sup>31,34</sup>. Samples from SSA4 clustered with SSA2 (Fig. 1B,C), suggesting that mtCOI-based SSA4 species identification was likely misidentification of SSA2 due to NUMTs<sup>29</sup>. Removal of NUMTs resolved the mtCOI discrepancies, and reanalyses of genome-wide SNP data provided insights into SSA1-species-wide population substructure, identified likely geographic contact zones, and signatures of infra- and inter-species hybridization.

Maximum-Likelihood (ML) phylogeny inferred using RAxML based on 14,358 genome-wide SNPs (Fig. 2A) also supported the PCA result and broadly divided these 62 SSA individuals into three clusters that consisted of: (i) SSA2 (i.e., individuals previously identified as SSA4 or SSA2 by mtCOI; 100% node confidence), (ii) SSA1-SE (i.e., ‘SSA-CA/SSA-ESA’ populations; node confidence 89%), and (iii) SSA1-NW (i.e., ‘SSA-ECA/SSA-WA’ populations; node confidence 96%). The phylogenomic analysis supported the SSA1 and SSA2 species status as consistent with results from partial mtCOI gene sequences (Fig. 2B,C). Two clades within SSA1, denoted as SSA1-SE (South-Eastern) and SSA1-NW (North-Western) (Fig. 2C), respectively, was evident, reflecting the geographic distribution of the populations identified within them (SSA1-SE individuals from DR Congo, Tanzania, Madagascar; SSA1-NW samples from Cameroon, Nigeria, DR Congo, Rwanda, Burundi, Tanzania). Within SSA1-NW, the ‘SSA-ECA’ and ‘SSA-WA’ populations<sup>38</sup> are clustered separately with 99% and 96% confidence values. Within SSA1-SE, the ‘SSA-ESA’ and ‘SSA-CA’ populations<sup>38</sup> are partially resolved, sharing only 34% confidence and suggesting insufficient genomic evidence to support their separate genetic group classification<sup>38</sup>.

Admixture analysis identified three genetic clusters (Fig. 3) that followed the partial mtCOI gene-based species diagnosis (see also Suppl. Fig. 3), whereby the SSA2 species was clearly shown to be different from the SSA1 species. The SSA1 species was further shown to consist of two distinct genetic backgrounds corresponding to SSA1-NW and SSA1-SE (containing ‘blue’ SSA-WA and SSA-ECA, and ‘red’ SSA-CA and SSA-ESA, respectively, as defined by Wosula et al.<sup>38</sup>). Some SSA2 and all the SSA1-CA samples showed evidence for low levels of admixture.

Interestingly,  $F_{st}$  analysis (Table 1) supported SSA1-NW and SSA1-SE as potential putative sub-species of SSA1, with intermediate  $F_{st}$  estimates observed for all pair-wise comparisons at the inter-specific (i.e., SSA2 vs. SSA1;  $F_{st} = 0.286–0.334$ ) level. This was particularly evident for SSA1-NW versus the SSA1-SE ‘ESA’ population, with  $F_{st}$  of 0.226–0.243.  $F_{st}$  differentiation was less clearly defined for the SSA-CA population, whose  $F_{st}$  scores versus SSA1-NW were closer to those between populations within the two major SSA1 groups. The overall differentiation between SSA1-NW and SSA1-SE suggested likely existence of putative sub-species within SSA1, the observations concerning the SSA-CA population suggest gene flow between neighbouring populations. Treatment of SSA1 as single species against the SSA2/SSA4 species showed significant population substructure ( $F_{st}$  range 0.286–0.332). 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 belonging to two well-defined groups (i.e., SSA-NW, SSA-SE) akin to ‘putative sub-species’, showed that SSA-WA and SSA-ECA overall shared intermediate levels of gene flow with the SSA-ESA population.  $F_{st}$  estimates are lowest between the SSA-CA population and the SSA-WA, SSA-ECA, and SSA-ESA populations, suggesting greater gene flow at the contact zone.

Within the putative SSA1-NW sub-species, the SSA-WA population is predominantly found in the lowlands (ca. 300–1500 m), whereas the SSA-ECA/SSA-CA populations were found at higher altitudes (from ca. 600–≥ 1500 m). Individuals belonging to the SSA-ESA group were predominantly from coastal landscape of Tanzania and Madagascar (Fig. 4). A gene flow contact zone was detected at regions represented by five DR Congo individuals and one Tanzanian individual (i.e., the SSA-CA population). Inter-specific hybridization

**Figure 2.** Phylogeny of SSA populations. (A) RAxML phylogeny with 1000 bootstrap replications based on 14,358 nextRAD genome-wide SNPs from 62 SSA *Bemisia* individuals and three ‘non-cassava individuals’<sup>38</sup> as outgroup. Branch colour schemes (red, green, blue) are as used in Admixture analysis (see Fig. 3). Sample names and branch tip colour schemes (filled circles) followed Wosula et al.<sup>38</sup>, and where available, mtCOI GenBank accession numbers corresponding to selected samples, and their proposed mtCOI SSA1-subgroups (SG)<sup>13</sup> status (i.e., SG1, SG2, SG3, SG5<sup>44</sup>) are also listed. Individuals that have partial mtCOI gene sequence but have been excluded from the nextRAD genome-wide SNPs phylogeny analysis due to poor SNP data coverage are boxed. The *Bemisia* SSA2-SSA4 clade is indicated by green branches, and the SSA lineages as proposed by Wosula et al.<sup>38</sup> [i.e., CA (pink), ESA (black), WA (green), ECA (red)] are here identified as belonging to the SSA1 species and together, made up the ‘SSA1-SE’ and ‘SSA1-NW’ putative sub-species. Note that only nodes with  $\geq 75\%$  bootstrap support are shown. Enlarged section of branch breaks separating SSA1-NW, SSA1-SE, and SSA2 are shown by the insert box. The long branch length (0.43) separating the outgroups and the SSA1/SSA2 clades has been shortened as indicated by ‘//’. (B) Radial phylogeny with sample names and lineage colours as for (A). (C) Detail of the RAxML phylogeny in (B) showing only the SSA1 (i.e., represented by all ‘SE’ and ‘NW’ individuals) and SSA2 (represented by SSA2 and ‘SSA4’ individuals) clades. Proposed genetic groups for the SSA1 species by Wosula et al.<sup>38</sup> are also indicated (i.e., ‘CA’ and ‘ESA’, and ‘WA’ and ‘ECA’ within the putative SSA1-SE and SSA1-NW sub-species, respectively). Country codes are: Tanzania (TZ), Democratic Republic of Congo (DRC), Madagascar (MAD), Cameroon (CAM), Central African Republic (RCA), Nigeria (NIG), Rwanda (RWA), Burundi (BUR).

was detected in two SSA2 individuals (Fig. 3, green/blue genetic backgrounds) and in one DR Congo individual from SSA-CA (red/green/blue in Fig. 3).

The Treemix ML tree supported the SSA4 species as part of SSA2, and the migration edges ( $m = 2$ ) revealed gene flow in both directions, between SSA1 and SSA2 species from central Africa (i.e., SSA-CA), and supports detection of hybrid individuals identified in the Admixture analysis (Fig. 3). The weight of the migration edge going from SSA-CA to SSA2 is also higher than the one going the opposite direction (Fig. 5). TreeMix ML tree also showed a lack of branch depth between each cluster of SSA2/SSA4, ECA/WA, and CA/ESA, thereby provided further support that entities within the three clusters were identical.

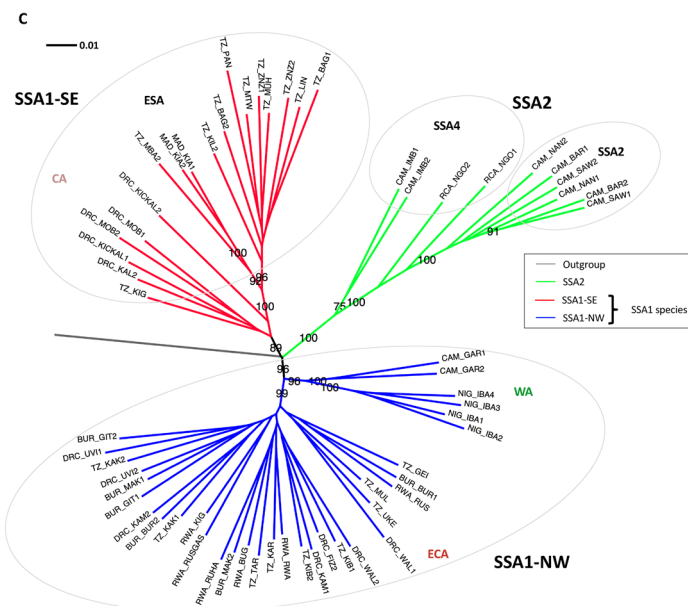
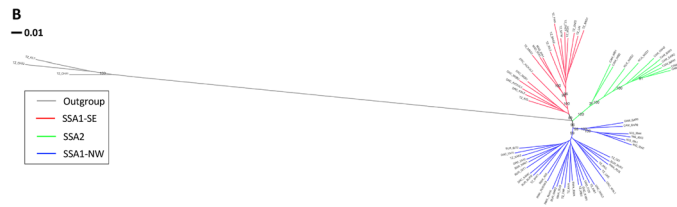
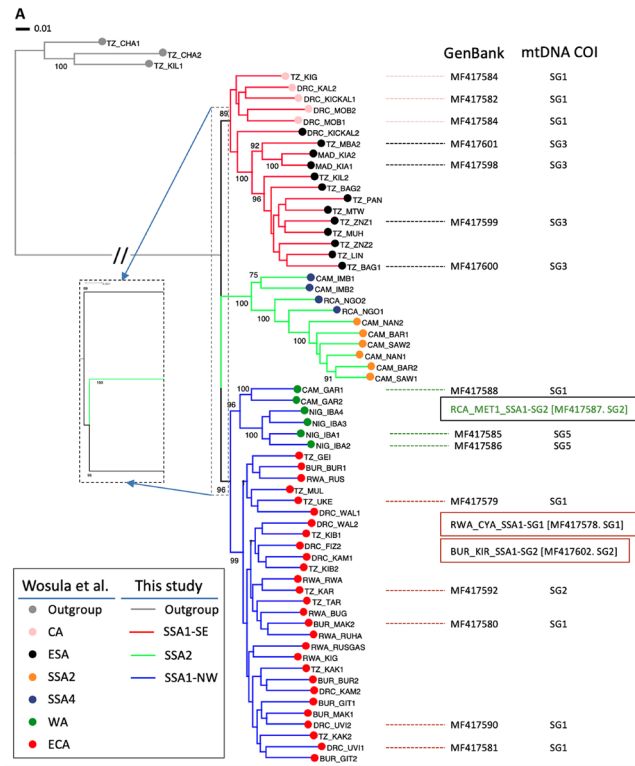
## Discussion

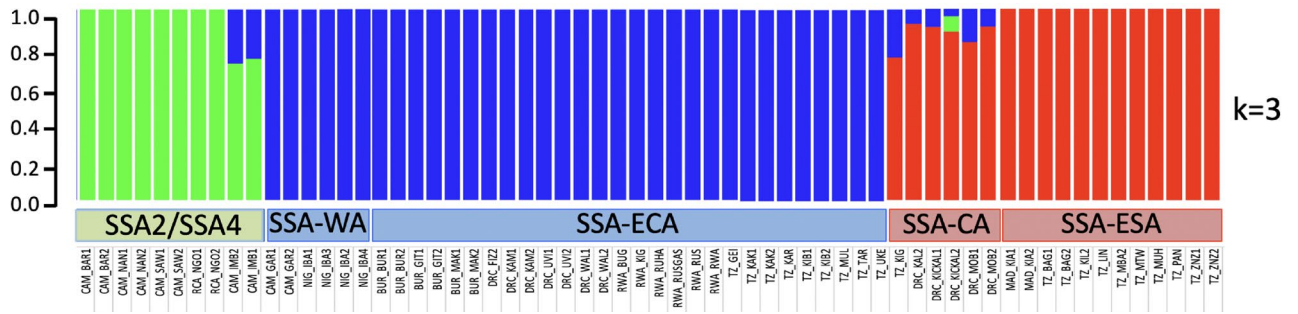
We combined genome-wide SNPs and mtCOI markers to gain evolutionary, ecological and population genomic insights into the sub-Saharan African cassava *Bemisia* whitefly species complex. By taking into consideration the impact of NUMTs, we showed no conflict in the same nuclear and mtCOI dataset in defining *Bemisia* species status. A thorough filtering and reanalysis of genome-wide SNP markers also detected signatures of incipient speciation and admixture in the *Bemisia* SSA1 species. While various evolutionary genetic studies based on the partial mtCOI gene have proposed the presence of SSA1 species subgroups<sup>13,46</sup>, we caution against the use of the mitochondrial ‘subgroup’ classification based on genome-wide SNP markers. Mating experiments were inconclusive in separating subgroups of SSA1 especially between SG1 and SG2<sup>31</sup>, while for the first time, genome-wide SNP loci showed that the SSA1 species could potentially be defined as to consist of a north-western (NW) and a south-eastern (SE) putative sub-species, with a contact zone separated by the eastern highland and central/western lowland regions.

Wosula et al.<sup>38</sup> adopted the nextRAD genome-wide SNP approach first demonstrated by Elfekih et al.<sup>10,47–49</sup> for *Bemisia* species to investigate gene flow patterns and diversity in the SSA1, SSA2, and SSA4 African cassava *Bemisia* whitefly species complex, but overlooked the impact of NUMTs<sup>10,32,35</sup> in published *B. tabaci* phylogenetic studies based on partial mtCOI sequences. This led to the conclusion that the mtCOI gene lacked efficacies at differentiating African *Bemisia* cryptic species. Increasingly, reanalyses of the *B. tabaci* cryptic species complex partial mtCOI gene dataset have detected the presence of significant NUMTs and pseudogenes<sup>3,29,35</sup>. These studies highlighted on-going challenges associated with species delimitation in the cryptic *B. tabaci* species complex. Similar to MEAM2<sup>28,50</sup>, our analysis supported that the SSA4<sup>42</sup> was likely NUMT artefacts as previously proposed<sup>29</sup>. The misidentification of SSA4 underpinned all subsequent genomic interpretations of gene flow and evolutionary ecology in the cassava SSA *Bemisia* whitefly species<sup>38</sup>.

Significant genetic diversity in the *B. tabaci* species complex can obscure species/putative sub-species identification. For instance, high genetic diversity in the MED species based on nextRAD SNP data has been reported<sup>10</sup>; however, the MED-ASL as a separate species was only further supported by including also complete mitogenome, mating experiments<sup>3</sup>, and host preference studies<sup>5</sup>. Genomics and mating experiments similar to the MED-ASL studies<sup>3,5</sup> showed that the SSA1-SG1 and SG2 were likely the same species, with SG3 potentially being different<sup>31</sup> (i.e., the ‘WA’ clade; Fig. 2B,C). Similar studies are needed to provide additional biological support for various SSA *Bemisia* species subgroups<sup>51</sup> including the putative SSA1-NW and SSA1-SE sub-species.

In developing the DNA barcoding approach, Hebert et al.<sup>52</sup> showed that the COI gene is overall effective for species level delimitation due to the associated nucleotide substitutional rates for this gene (but see Hanemaaijer et al.<sup>53</sup>), while for fine-scale identification such as within species level, the use of genes with greater mutational rates were recommended. Partial mtCOI gene therefore lacks the power to differentiate between evolutionary very closely related cryptic species status, for which a genomic approach should be considered<sup>3,31</sup>. Subgroup classification within the SSA1 African cassava *Bemisia* whitefly species based on the mtDNA gene might be possible, provided that a gene region with an appropriate substitutional rate is identified. Nevertheless, there is currently no evidence for maternal lineage-based patterns to population dispersal behaviour, host adaptation, and fitness

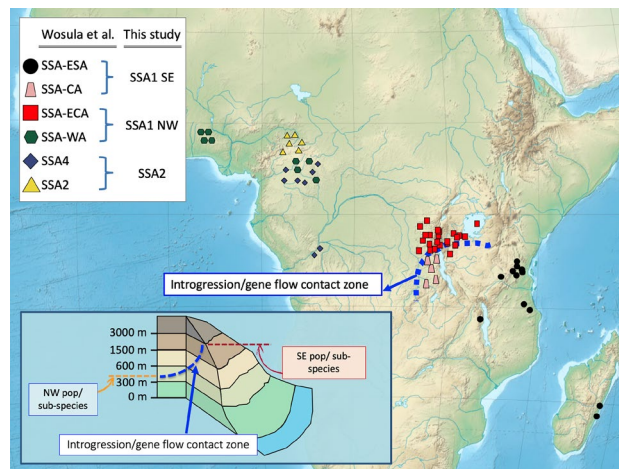




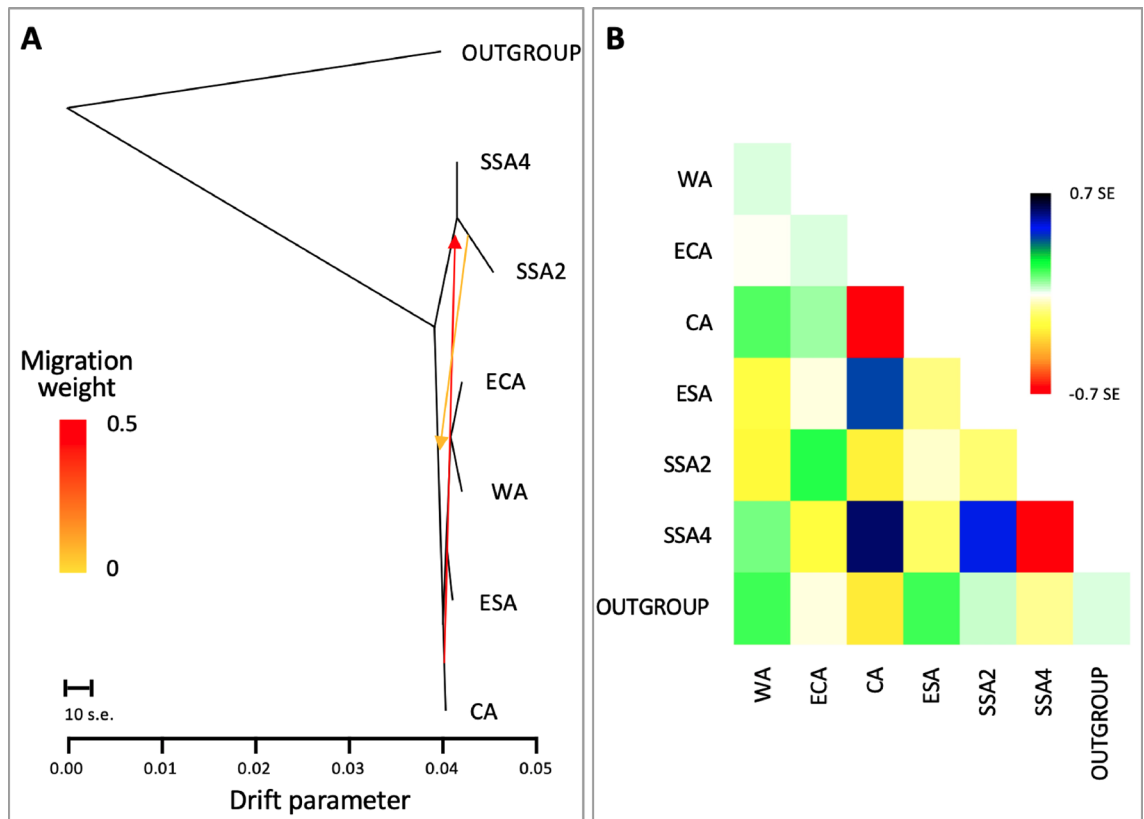
**Figure 3.** Admixture analysis at  $K = 3$  of 62 individuals of sub-Saharan African *Bemisia* cryptic species based on 14,358 SNPs. Individuals were clustered into the SSA2 species group, as well as two distinct clusters (red, blue) that belonged to the SSA1 species. The two SSA1 clusters in the Admixture panel included the SSA-WA and SSA-ECA groups (34 blue individuals; blue = SSA1-NW), and six SSA-CA (blue-red/blue-red-green) and 12 SSA-ESA (red; SSA1-SE) individuals. Evidence of introgression between the red and blue SSA1 clusters were predominantly detected in SSA-CA individuals. Three SSA2/SSA1 inter-specific hybrids (CAM\_IMB2\_SSA4, CAM\_IMB1\_SSA4, DRC\_KICKAL2\_CA) were also detected. Note: Figure colour scheme matches the phylogeny branch colours of Fig. 2A–C.

Species	SSA1				
Sub-species	SSA2	SSA-NW	SSA-SE		
Populations	SSA2/SSA4	SSA-WA	SSA-ECA	SSA-CA	SSA-ESA
SSA2/SSA4					
SSA-WA	0.291				
SSA-ECA	0.334	0.171			
SSA-CA	0.286	0.166	0.158		
SSA-ESA	0.332	0.243	0.226	0.160	

**Table 1.** Population substructure based on the  $F_{st}$  of Weir and Cockerham<sup>45</sup>.



**Figure 4.** A map showing sampling sites of the SSA *Bemisia* cryptic species. The sample colour schemes and sampling sites followed Fig. 6B of Wosula et al.<sup>38</sup> (green: SSA-WA; red: SSA-ECA; pink: SSA-CA; black: SSA-ECA; Yellow: SSA2; blue: SSA4). Note: The SSA4 species is likely NUMT artefacts of the SSA2 species (refer to main text). The African topological map colour represents elevation from sea-level (see insert map of non-incremental approximation of elevation). The gene flow contact zone between populations on the NW–SE transect is indicated by the thick blue dashed line. Map modified and prepared using Microsoft PowerPoint for Mac, Version 16.39 (20071300). Sourced: [https://commons.wikimedia.org/wiki/File:Africa\\_topography\\_map.png](https://commons.wikimedia.org/wiki/File:Africa_topography_map.png).



**Figure 5.** Gene flow between SSA species. (A) TreeMix ML tree with two migration edges ( $m = 2$ ), showing strong gene flow from SSA1 to SSA2/SSA4 (red arrow), and weaker gene flow from SSA2 to SSA1. Clustering between ECA and WA, and between ESA and CA, provided support for their close relationships that could potentially indicate their respective putative sub-species status (i.e., SSA1-NW, SSA1-SE, respectively). Two basal branches leading to SSA1 populations (i.e., ECA, WA, ESA, CA) and SSA2/SSA4 further supported the populations analysed as consisted of two species (i.e., SSA1, SSA2). The scale bar indicates ten times the average standard error (s.e.). (B) The residual fit was plotted from the ML tree, which is the covariance matrix between each pair of populations. Residuals above zero are an indication of the populations most likely to be admixed.

cost differentiation in the *B. tabaci* complex and the African cassava *Bemisia* species complex, and therefore the subgroup delimitation based on the partial mtCOI gene is not justified.

The SSA1-SG1, representing the major SSA1 mtCOI subgroup, was previously implicated in the CMD pandemic in East Africa<sup>13</sup>. Wosula et al.<sup>38</sup> reclassified SSA1-SG1 individuals (e.g., SSA1-SG3, SSA1-SG5<sup>44</sup>) into three SNP-based groups of SSA-ECA, SSA-CA, and SSA-WA, due to discrepancies between the 657 bp mtCOI partial gene and nuclear SNP markers. Based on the same nuclear and mitochondrial dataset, we showed that while separate phylogenetic signals were detected for the ‘WA’ and ‘ECA’ populations, a strong genetic difference between the ‘CA’ and ‘ESA’ populations was lacking, and we therefore advocate following established good practice<sup>51</sup> to refrain prematurely adopting other nomenclature (i.e., those proposed by Wosula et al.<sup>38</sup>).

Although refined higher-level species status could help resolve virus transmission capacity and provide insights into understanding pest introduction pathways<sup>11,54</sup>, there is currently insufficient evidence in African cassava *Bemisia* whiteflies to link either the WA/ECA or CA/ESA populations to historical CMD outbreaks. The *B. tabaci* SSA2 (i.e., ‘Ug2’<sup>33</sup>) was assumed to be associated with the 1990s CMD epidemic, mainly due to its geographic distribution in relation to the front of the epidemic<sup>55</sup>. However, transmission underpinning outbreaks of CMD in the sub-Saharan Africa region is complex, with one factor being the sharing and propagation of diseased cassava cuttings between farmers. There was a recombination event between two geminiviruses that produced a more virulent hybrid<sup>56</sup>, however we have little historical information about how the viruses interacted with the different *Bemisia* cryptic species in terms of disease dynamics<sup>57</sup>. Evidence for bidirectional interspecific hybridization detected between SSA1 and SSA2 in this study, therefore may have implications to understanding the evolutionary genetics associated with plant virus transmission capacity by the cassava SSA *Bemisia* whitefly species complex. The role of the SSA2 species in the CMD epidemic nevertheless remained unclear, with the SSA2’s habitat range (east and west Africa) overlapping the SSA1 species<sup>22,38,55,58–62</sup>. The exact relationship between SSA1/SSA2 and outbreaks of CMD<sup>63</sup> should be re-visited, taking into consideration current genomic evidence for the presence of SSA1/SSA2 hybrids. Furthermore, vector competencies of other common cassava whitefly species (e.g., *B. afer*) to different cassava viruses also need further investigation, an ideally include a genomic component.

*B. tabaci* sensu lato on cassava in sub-Saharan Africa is a species complex going through a continuous diversification process. The putative SSA1-SE sub-species that the Madagascan (SSA-ESA) and Tanzanian (SSA-CA) populations belong to are geographically isolated from the Nigerian (SSA-WA), DR Congo, and Burundi (SSA-ECA) populations with a hybrid zone between Tanzania, Uganda and DR Congo. The East African Rift Valley is a major geographical barrier to gene flow and provides the environment to promote incipient speciation and to enable recent radiations to thrive and evolve<sup>64</sup>, and could underpin incipient allopatric speciation in the SSA1. The “rift effect” acting as a gene flow barrier has been described in other insects (e.g., the maize stalk borer *Busseola fusca*<sup>65</sup>, the malaria vector mosquitoes *Anopheles gambiae*<sup>66,67</sup>) and vertebrates (e.g., birds<sup>68</sup>, frogs<sup>69</sup>). Specifically, the western branch of the Rift Valley, i.e., the Albertine Rift covering part of Tanzania, Burundi, Rwanda, Uganda and DR Congo, is a mountain chain situated at 1500–3500 m above sea level (Fig. 4) formed by volcanic activity since the plio-pleistocene. In addition to volcanic activities, climate change during this glacial period led to habitat fragmentation that promoted species diversification in the newly formed biogeographical zones<sup>70</sup>. Species diversification in the African *Bemisia* whitefly cryptic species system has also been explained by host plant associations and their ability to feed on diverse host plants and acquiring new ones<sup>71</sup>. Using RNAseq, Malka et al.<sup>72</sup> provided insights into how the variation in detoxification gene expression levels might have conferred plasticity that likely also contributed to species diversification.

Finally, we note that our genome-wide SNP phylogenetic analysis showed weak support (bootstrap value 34%, data not shown) for closer evolutionary relationship between SSA2 and SSA1-NW, than between the two putative SSA1 sub-species. This closer phylogenetic relationship between SSA2 and SSA1-NW is likely to be unreliable given the weak bootstrap support, greater gene flow patterns between the SSA1 sub-species (Table 1), and the congruence of species status based on mtCOI molecular diagnostics. Our study therefore further highlights the need for careful selection of outgroups that takes into consideration the evolutionary biological questions to be addressed (e.g., whether the target species were more closely related (i.e., between SSA1, SSA2, SSA3, SSA6, SSA9)<sup>29</sup> or evolutionary diverged (i.e., between SSA1 and AsiaII\_1)<sup>29</sup>). Data obtained for this population genomic reanalysis involved only limited number of target and outgroups individuals and species that represented significant evolutionary divergent lineages from the SSA1/SSA2 species<sup>38</sup>. The challenge of obtaining sufficient sequencing coverage that is necessary for filtering to obtain a robust set of genome-wide SNP markers for the *Bemisia* cryptic species complex<sup>10,47–49</sup> known to have a large and complex genome organization<sup>73,74</sup> will likely further enhance and perpetuate their on-going status as one of the world’s most damaging and challenging group of agricultural insect pest complex.

## Material and methods

**mtCOI data handling.** The mtCOI sequences of Wosula et al.<sup>38</sup> were downloaded from GenBank (MF417578–MF417602), including all published reference sequences used to identify the SSA1 ‘genetic groups’. These partial mtCOI sequences were imported into Geneious R11.1 and aligned with reference sequences (Suppl. Fig. 1) using the MAFFT alignment program<sup>75</sup> with the following settings (Algorithm option set at ‘Auto’; Scoring matrix at 200PAM/K = 2; Gap open penalty and Offset values at 1.53 and 0.123, respectively). We examined all sequences for: (i) insertion/deletions (indels), (ii) presence of premature stop codons, and (iii) presence of unexpected amino acid substitutions at evolutionary conserved regions in the partial COI gene region<sup>29</sup> as signatures of pseudogenes (NUMTs).

**nextRAD data processing.** The raw data from nextRAD sequencing<sup>38</sup> was accessed from GenBank (SRP103541). The fastq sequences were screened using FastQC v.0.11.4 for quality control. The reads were trimmed by quality and reads longer than 151 bp was trimmed to this length using Trimmomatic<sup>76</sup> version 0.36.

**Mapping quality.** We mapped the reads from each individual sample to the *B. tabaci* MEAM1 genome<sup>73</sup> using Burrows-Wheeler Aligner (BWA) v.0.7.12<sup>77</sup>. The alignments used the BWA-MEM algorithm with default settings. The SAM files were converted to BAM output using Samtools<sup>78</sup> v. 1.3.1. The BAM files were subsequently sorted, indexed and checked for quality and mapping percentages per scaffold.

**SNP genotyping.** SNPs were first called using a de novo approach in PyRAD v. 0.9<sup>79</sup>. PyRAD is a pipeline designed for RADseq datasets that aims to capture variation at the species/clade level. It allows clustering of highly divergent sequences and takes into consideration indel variation. We also performed SNP calling in PyRAD, using reads mapped to the MEAM1 reference genome<sup>73</sup>.

**Genetic clusters and tests for introgression.** Genetic clusters within the dataset that consists of 62 SSA samples were investigated by Principal Component Analysis (PCA) using the SNPRelate R package<sup>80</sup>. We used ADMIXTURE v.1.3.0<sup>81</sup> on the 62-sample dataset to identify genetic clusters and estimate the genetic ancestry of the SSA samples. We ran the program with K ancestral clusters varying from 1 to 10 with each ‘K’ value repeated 100 times. A cross-validation test was performed to determine the optimal K value.

**Phylogenetic analyses.** The topology of the phylogenetic relationships between individuals/populations/sub-species and species within the dataset (62 SSA samples and three outgroup individuals (TZ\_CHA1, TZ\_CHA2, TZ\_KIL1)) was examined using a maximum likelihood (ML) approach. The SSA species status were first identified using the mtCOI marker, then, a phylogeny was reconstructed using the nextRAD genome-wide SNPs excluding samples with low genotype quality (mapping quality < 10) to minimize biases that could potentially be introduced by missing data. We performed five replicates of phylogenetic reconstruction in RAxML v.7.2.8<sup>82</sup> using the GTR substitution model and GTRGAMMA as the GAMMA model of rate heterogeneity, with 1000 bootstrap replications. TreeMix<sup>83</sup> v. 1.12 was used to identify genetic mixing, the history of population splits



and admixture patterns. We ran simulations in order to infer the best-fit scenario for the genetic relationships between these cassava *Bemisia* whitefly populations involving up to 5 migrations events with 100 bootstrap replications (e.g., based on previous studies using this type of analysis<sup>49,84,85</sup>). The likelihoods were compared using the following command in R: `tail -n 1 Data/data.frq.strat.tree.*.lik | cut -d \: -f 2 > Data/data`. The decision to select the number of migration  $m = 2$  is based on the residual fit matrix generated by the TreeMix program. For each migration ( $m = 0, m = 1 \dots, m = 5$ ), we generate a residual plot. When this plot is saturated at a given  $m$ , the best fit scenario and the best-fit ML tree that corresponds to the specific  $m$  number are obtained. If it is  $m = 2$ , then any number bigger than 2 would not add any more information. Programming scripts used in this study are available as Supplemental Data 1–7.

Received: 24 April 2020; Accepted: 17 March 2021

Published online: 12 April 2021

## References

- Hebert, P. D., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. T. species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astrartes fuligator*. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14812–14817. <https://doi.org/10.1073/pnas.0406166101> (2004).
- Saez, A. G. & Lozano, E. Body doubles. *Nature* **433**, 111. <https://doi.org/10.1038/433111a> (2005).
- Vyskočilová, S., Tay, W. T., van Brunschot, S., Seal, S. & Colvin, J. An integrative approach to discovering cryptic species within the *Bemisia tabaci* whitefly species complex. *Sci. Rep.* **8**, 10886. <https://doi.org/10.1038/s41598-018-29305-w> (2018).
- Liu, S. S. *et al.* Asymmetric mating interactions drive widespread invasion and displacement in a whitefly. *Science* **318**, 1769–1772. <https://doi.org/10.1126/science.1149887> (2007).
- Vyskočilová, S., Seal, S. & Colvin, J. Relative polyphagy of “Mediterranean” cryptic *Bemisia tabaci* whitefly species and global pest status implications. *J. Pest Sci.* **92**, 1071–1088. <https://doi.org/10.1007/s10340-019-01113-9> (2019).
- Behere, G. T., Tay, W. T., Russell, D. A. & Batterham, P. Molecular markers to discriminate among four pest species of *Helicoverpa* (Lepidoptera: Noctuidae). *Bull. Entomol. Res.* **98**, 599–603. <https://doi.org/10.1017/S0007485308005956> (2008).
- Elfekih, S., Tay, W. T., Gordon, K., Court, L. N. & De Barro, P. J. Standardized molecular diagnostic tool for the identification of cryptic species within the *Bemisia tabaci* complex. *Pest Manag. Sci.* **74**, 170–173. <https://doi.org/10.1002/ps.4676> (2018).
- Walsh, T. K. *et al.* Mitochondrial DNA genomes of five major *Helicoverpa* pest species from the Old and New Worlds (Lepidoptera: Noctuidae). *Ecol. Evol.* **9**, 2933–2944. <https://doi.org/10.1002/ece3.4971> (2019).
- Anderson, C. J., Tay, W. T., McLaughran, A., Gordon, K. & Walsh, T. K. Population structure and gene flow in the global pest, *Helicoverpa armigera*. *Mol. Ecol.* **25**, 5296–5311. <https://doi.org/10.1111/mec.13841> (2016).
- Elfekih, S. *et al.* Genome-wide analyses of the *Bemisia tabaci* species complex reveal contrasting patterns of admixture and complex demographic histories. *PLoS ONE* **13**, e0190555. <https://doi.org/10.1371/journal.pone.0190555> (2018).
- Anderson, C. J. *et al.* Hybridization and gene flow in the mega-pest lineage of moth, *Helicoverpa*. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5034–5039. <https://doi.org/10.1073/pnas.1718831115> (2018).
- FAOSTAT. <http://www.fao.org/faostat/en/#data/QC/visualize> (2017).
- Legg, J. P. *et al.* Spatio-temporal patterns of genetic change amongst populations of cassava *Bemisia tabaci* whiteflies driving virus pandemics in East and Central Africa. *Virus Res.* **186**, 61–75. <https://doi.org/10.1016/j.virusres.2013.11.018> (2014).
- Patil, B. L. & Fauquet, C. M. Cassava mosaic geminiviruses: Actual knowledge and perspectives. *Mol. Plant Pathol.* **10**, 685–701. <https://doi.org/10.1111/j.1364-3703.2009.00559.x> (2009).
- Macfadyen, S. *et al.* Cassava whitefly, *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae) in East African farming landscapes: A review of the factors determining abundance. *Bull. Entomol. Res.* **108**, 565–582. <https://doi.org/10.1017/S0007485318000032> (2018).
- Minato, N. *et al.* Surveillance for Sri Lankan cassava mosaic virus (SLCMV) in Cambodia and Vietnam one year after its initial detection in a single plantation in 2015. *PLoS ONE* **14**, e0212780. <https://doi.org/10.1371/journal.pone.0212780> (2019).
- Wang, H. L. *et al.* First Report of Sri Lankan cassava mosaic virus Infecting Cassava in Cambodia. *Plant Dis.* **100**, 1029–1029. <https://doi.org/10.1094/Pdis-10-15-1228-Pdn> (2016).
- De Barro, P. J., Liu, S. S., Boykin, L. M. & Dinsdale, A. B. *Bemisia tabaci*: A statement of species status. *Annu. Rev. Entomol.* **56**, 1–19. <https://doi.org/10.1146/annurev-ento-112408-085504> (2011).
- Hopkinson, J. *et al.* Insecticide resistance status of *Bemisia tabaci* MEAM1 (Hemiptera: Aleyrodidae) in Australian cotton production valleys. *Austral Entomol.* **59**, 202–214 (2020).
- Hadjistylli, M., Roderick, G. K. & Gauthier, N. First report of the Sub-Saharan Africa 2 species of the *Bemisia tabaci* complex in the Southern France. *Phytoparasitica* **43**, 679–687. <https://doi.org/10.1007/s12600-015-0480-3> (2015).
- Lee, W., Park, J., Lee, G. S., Lee, S. & Akimoto, S. Taxonomic status of the *Bemisia tabaci* complex (Hemiptera: Aleyrodidae) and reassessment of the number of its constituent species. *PLoS ONE* **8**, e63817. <https://doi.org/10.1371/journal.pone.0063817> (2013).
- Mugerwa, H. *et al.* African ancestry of New World, *Bemisia tabaci*-whitefly species. *Sci. Rep.* **8**, 2734. <https://doi.org/10.1038/s41598-018-20956-3> (2018).
- Martin, J. H. An identification guide to common whitefly pest species of the world (Homopt Aleyrodidae). *Int. J. Pest Manag.* **33A**, 298–322. <https://doi.org/10.1080/09670878709371174> (1987).
- Martin, J. H. & Mound, L. A. An annotated check list of the world's whiteflies (Insecta: Hemiptera: Aleyrodidae). *Zootaxa* **1492**, 1–84 (2007).
- Mound, L. A. Host-correlated variation in *Bemisia tabaci* (Gennadius). *Proc. R. Entomol. Soc. Lond.* **A38**, 171–180 (1963).
- Tay, W. T. *et al.* Novel molecular approach to define pest species status and tritrophic interactions from historical *Bemisia* specimens. *Sci. Rep.* <https://doi.org/10.1038/s41598-017-00528-7> (2017).
- Tay, W. T., Evans, G. A., Boykin, L. M. & De Barro, P. J. Will the real *Bemisia tabaci* please stand up?. *PLoS ONE* <https://doi.org/10.1371/journal.pone.0050550> (2012).
- Dinsdale, A., Cook, L., Riginos, C., Buckley, Y. M. & De Barro, P. Refined global analysis of *Bemisia tabaci* (Hemiptera: Sternorrhyncha: Aleyrodoidea: Aleyrodidae) mitochondrial cytochrome oxidase 1 to identify species level genetic boundaries. *Ann. Entomol. Soc. Am.* **103**, 196–208. <https://doi.org/10.1603/An09061> (2010).
- Kunz, D., Tay, W. T., Elfekih, S., Gordon, K. H. J. & De Barro, P. J. Take out the rubbish - Removing NUMTs and pseudogenes from the *Bemisia tabaci* cryptic species mtCOI database. *bioRxiv*. <https://doi.org/10.1101/724765> (2019).
- Wongnikong, W., van Brunschot, S. L., Hereward, J. P., De Barro, P. J. & Walter, G. H. Testing mate recognition through reciprocal crosses of two native populations of the whitefly *Bemisia tabaci* (Gennadius) in Australia. *Bull. Entomol. Res.* **110**, 328–339. <https://doi.org/10.1017/S0007485319000683> (2020).
- Mugerwa, H., Wang, H.-L., Sseruwagi, P., Seal, S. & Colvin, J. Whole-genome single nucleotide polymorphism and mating compatibility studies reveal the presence of distinct species in sub-Saharan Africa *Bemisia tabaci* whiteflies. *Insect Sci.* <https://doi.org/10.1111/1744-7917.12881> (2020).

32. Delatte, H. *et al.* A new silverleaf-inducing biotype Ms of *Bemisia tabaci* (Hemiptera: Aleyrodidae) indigenous to the islands of the south-west Indian Ocean. *Bull. Entomol. Res.* **95**, 29–35. <https://doi.org/10.1079/Ber2004337> (2005).
33. Boykin, L. M., Savill, A. & De Barro, P. Updated mtCOI reference dataset for the *Bemisia tabaci* species complex. *F1000Research* **6**, 1835. <https://doi.org/10.12688/f1000research.12858.1> (2017).
34. Liu, S. S., Colvin, J. & De Barro, P. J. Species concepts as applied to the whitefly *Bemisia tabaci* systematics: How many species are there?. *J Integr Agr* **11**, 176–186. [https://doi.org/10.1016/S2095-3119\(12\)60002-1](https://doi.org/10.1016/S2095-3119(12)60002-1) (2012).
35. Tay, W. T. *et al.* The trouble with MEAM2: Implications of pseudogenes on species delimitation in the globally invasive *Bemisia tabaci* (Hemiptera: Aleyrodidae) cryptic species complex. *Genome Biol. Evol.* **9**, 2732–2738. <https://doi.org/10.1093/gbe/evx173> (2017).
36. Kunz, D. *et al.* Draft mitochondrial DNA genome of a 1920 Barbados cryptic *Bemisia tabaci* “New World” species (Hemiptera: Aleyrodidae). *Mitochondrial DNA B* **4**, 1183–1184. <https://doi.org/10.1080/23802359.2019.1591197> (2019).
37. Paine, D. R. *et al.* Global threat to agriculture from invasive species. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7575–7579. <https://doi.org/10.1073/pnas.1602205113> (2016).
38. Wosula, E. N., Chen, W. B., Fei, Z. J. & Legg, J. P. Unravelling the genetic diversity among cassava *Bemisia tabaci* whiteflies using NextRAD sequencing. *Genome Biol. Evol.* **9**, 2958–2973. <https://doi.org/10.1093/gbe/evx219> (2017).
39. Thresh, J. M., Fargette, D. & Otim-Nape, G. W. Effects of African cassava mosaic geminivirus on the yield of cassava. *Trop. Sci.* **34**, 26–42 (1994).
40. Legg, J. *et al.* A global alliance declaring war on cassava viruses in Africa. *Food Secur.* **6**, 231–248. <https://doi.org/10.1007/s12571-014-0340-x> (2014).
41. Legg, J. P. *et al.* Biology and management of *Bemisia* whitefly vectors of cassava virus pandemics in Africa. *Pest Manag. Sci.* **70**, 1446–1453. <https://doi.org/10.1002/ps.3793> (2014).
42. Berry, S. D. *et al.* Molecular evidence for five distinct *Bemisia tabaci* (Homoptera: Aleyrodidae) geographic haplotypes associated with cassava plants in sub-Saharan Africa. *Ann. Entomol. Soc. Am.* **97**, 852–859. [https://doi.org/10.1603/0013-8746\(2004\)097\[0852:Meffib\]2.0.Co;2](https://doi.org/10.1603/0013-8746(2004)097[0852:Meffib]2.0.Co;2) (2004).
43. Mugerwa, H., Rey, M. E. C., Tairo, F., Ndunguru, J. & Sseruwagi, P. Two sub-Saharan Africa 1 populations of *Bemisia tabaci* exhibit distinct biological differences in fecundity and survivorship on cassava. *Crop Prot.* **117**, 7–14. <https://doi.org/10.1016/j.cropro.2018.11.011> (2019).
44. Ghosh, S., Bouvaine, S. & Maruthi, M. N. Prevalence and genetic diversity of endosymbiotic bacteria infecting cassava whiteflies in Africa. *BMC Microbiol.* <https://doi.org/10.1186/s12866-015-0425-5> (2015).
45. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population-structure. *Evolution* **38**, 1358–1370. <https://doi.org/10.2307/2408641> (1984).
46. Ghosh, S., Bouvaine, S., Richardson, S. C. W., Ghanim, M. & Maruthi, M. N. Fitness costs associated with infections of secondary endosymbionts in the cassava whitefly species *Bemisia tabaci*. *J. Pest Sci.* **91**, 17–28. <https://doi.org/10.1007/s10340-017-0910-8> (2018).
47. Elfekih, S. *et al.* Evolutionary genomics of *Bemisia tabaci* and characterization of its endosymbiont metacommunities using nextRAD sequencing. *International Plant and Animal Genome Asia, Singapore* 23–25 July 2015 (2015).
48. Elfekih, S. *et al.* Genome-wide SNPs Decipher Global Inursion pathways in the *Bemisia tabaci* species complex. *International Plant and Animal Genome Conferences San Diego*, 9–13 January 2016 (2016).
49. Elfekih, S. *et al.* Genome-wide scans unravel fine-scale invasion routes in the *Bemisia tabaci* species complex. *2nd International Whitefly Symposium, Arusha, Tanzania*. p38, 14–19 February 2016 (2016).
50. Boykin, L. M., Bell, C. D., Evans, G., Small, I. & De Barro, P. J. Is agriculture driving the diversification of the *Bemisia tabaci* species complex (Hemiptera: Sternorrhyncha: Aleyrodidae)? Dating, diversification and biogeographic evidence revealed. *BMC Evol. Biol.* **13**, 228. <https://doi.org/10.1186/1471-2148-13-228> (2013).
51. Boykin, L. M. *et al.* Review and guide to a future naming system of African *Bemisia tabaci* species. *Syst. Entomol.* **43**, 427–433. <https://doi.org/10.1111/syen.12294> (2018).
52. Hebert, P. D. N., Cywinska, A., Ball, S. L. & DeWaard, J. R. Biological identifications through DNA barcodes. *Proc. R. Soc. B Biol. Sci.* **270**, 313–321. <https://doi.org/10.1098/rspb.2002.2218> (2003).
53. Hanemaaijer, M. J. *et al.* Mitochondrial genomes of *Anopheles arabiensis*, *An. gambiae* and *An. coluzzii* show no clear species division [version 2; peer review: 2 approved]. *F1000Research* **7**, 347. <https://doi.org/10.12688/f1000research.13807.2> (2019).
54. Tabachnick, W. J. *Culicoides variipennis* and bluetongue-virus epidemiology in the United States. *Annu. Rev. Entomol.* **41**, 23–43. <https://doi.org/10.1146/annurev.en.41.010196.000323> (1996).
55. Legg, J. P., French, R., Rogan, D., Okao-Okuja, G. & Brown, J. K. A distinct *Bemisia tabaci* (Gennadius) (Hemiptera: Sternorrhyncha: Aleyrodidae) genotype cluster is associated with the epidemic of severe cassava mosaic virus disease in Uganda. *Mol. Ecol.* **11**, 1219–1229. <https://doi.org/10.1046/j.1365-294X.2002.01514.x> (2002).
56. Colvin, J., Omongo, C. A., Maruthi, M. N., Otim-Nape, G. W. & Thresh, J. M. Dual begomovirus infections and high *Bemisia tabaci* populations: Two factors driving the spread of a cassava mosaic disease pandemic. *Plant Pathol.* **53**, 577–584. <https://doi.org/10.1111/j.1365-3059.2004.01062.x> (2004).
57. Polston, J. E., De Barro, P. & Boykin, L. M. Transmission specificities of plant viruses with the newly identified species of the *Bemisia tabaci* species complex. *Pest Manag. Sci.* **70**, 1547–1552. <https://doi.org/10.1002/ps.3738> (2014).
58. Ally, H. M. *et al.* What has changed in the outbreaking populations of the severe crop pest whitefly species in cassava in two decades?. *Sci. Rep.* <https://doi.org/10.1038/s41598-019-50259-0> (2019).
59. Kalyebi, A. *et al.* Within-season changes in land use impact pest abundance in smallholder African cassava production systems. *Insects* (2021) (Revised Submitted).
60. Kalyebi, A. *et al.* African cassava whitefly, *Bemisia tabaci*, cassava colonization preferences and control implications. *PLoS ONE* **13**, e0204862. <https://doi.org/10.1371/journal.pone.0204862> (2018).
61. Macfadyen, S. *et al.* Landscape factors and how they influence whitefly pests in cassava fields across East Africa. *Landsc. Ecol.* **36**, 45–67. <https://doi.org/10.1007/s10980-020-01099-1> (2021).
62. Tay, W. T. *et al.* A high-throughput amplicon sequencing approach for population-wide species diversity and composition survey. *bioRxiv* <https://doi.org/10.1101/2020.10.12.336545> (2020).
63. Manani, D. M., Ateka, E. M., Nyanjom, S. R. G. & Boykin, L. M. Phylogenetic relationships among whiteflies in the *Bemisia tabaci* (Gennadius) species complex from major cassava growing areas in Kenya. *Insects* <https://doi.org/10.3390/insects8010025> (2017).
64. Gottelli, D., Marino, J., Sillero-Zubiri, C. & Funk, S. M. The effect of the last glacial age on speciation and population genetic structure of the endangered Ethiopian wolf (*Canis simensis*). *Mol. Ecol.* **13**, 2275–2286. <https://doi.org/10.1111/j.1365-294X.2004.02226.x> (2004).
65. Sezonlin, M. *et al.* Phylogeography and population genetics of the maize stalk borer *Busseola fusca* (Lepidoptera, Noctuidae) in sub-Saharan Africa. *Mol. Ecol.* **15**, 407–420. <https://doi.org/10.1111/j.1365-294X.2005.02761.x> (2006).
66. Lehmann, T. *et al.* The rift valley complex as a barrier to gene flow for *Anopheles gambiae* in Kenya. *J. Hered.* **90**, 613–621. <https://doi.org/10.1093/jhered/90.6.613> (1999).
67. Schmidt, H. *et al.* Transcontinental dispersal of *Anopheles gambiae* occurred from West African origin via serial founder events. *Commun. Biol.* **2**, 473. <https://doi.org/10.1038/s42003-019-0717-7> (2019).

68. Mairal, M. *et al.* Geographic barriers and Pleistocene climate change shaped patterns of genetic variation in the Eastern Afrotropical biodiversity hotspot. *Sci. Rep.* **7**, 45749. <https://doi.org/10.1038/srep45749> (2017).
69. Freilich, X. *et al.* Comparative Phylogeography of Ethiopian anurans: Impact of the Great Rift Valley and Pleistocene climate change. *BMC Evol. Biol.* **16**, 206. <https://doi.org/10.1186/s12862-016-0774-1> (2016).
70. Huhndorf, M. H., Peterhans, J. C. K. & Loew, S. S. Comparative phylogeography of three endemic rodents from the Albertine Rift, east central Africa. *Mol. Ecol.* **16**, 663–674. <https://doi.org/10.1111/j.1365-294X.2007.03153.x> (2007).
71. Matsubayashi, K. W., Ohshima, I. & Nosil, P. Ecological speciation in phytophagous insects. *Entomol. Exp. Appl.* **134**, 1–27. <https://doi.org/10.1111/j.1570-7458.2009.00916.x> (2010).
72. Malka, O. *et al.* Species-complex diversification and host-plant associations in *Bemisia tabaci*: A plant-defence, detoxification perspective revealed by RNA-Seq analyses. *Mol. Ecol.* **27**, 4241–4256. <https://doi.org/10.1111/mec.14865> (2018).
73. Chen, W. B. *et al.* The draft genome of whitefly *Bemisia tabaci*MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. *BMC Biol.* <https://doi.org/10.1186/s12915-016-0321-y> (2016).
74. Xie, W. *et al.* The invasive MED/Q *Bemisia tabaci* genome: A tale of gene loss and gene gain. *BMC Genomics* **19**, 68. <https://doi.org/10.1186/s12864-018-4448-9> (2018).
75. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. <https://doi.org/10.1093/molbev/mst010> (2013).
76. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170> (2014).
77. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595. <https://doi.org/10.1093/bioinformatics/btp698> (2010).
78. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> (2009).
79. Eaton, D. A. R. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844–1849. <https://doi.org/10.1093/bioinformatics/btu121> (2014).
80. Zheng, X. W. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328. <https://doi.org/10.1093/bioinformatics/bts606> (2012).
81. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664. <https://doi.org/10.1101/gr.094052.109> (2009).
82. Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690. <https://doi.org/10.1093/bioinformatics/btl446> (2006).
83. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* <https://doi.org/10.1371/journal.pgen.1002967> (2012).
84. Decker, J. E. *et al.* Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet.* **10**, e1004254. <https://doi.org/10.1371/journal.pgen.1004254> (2014).
85. Gompert, Z. *et al.* Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Mol. Ecol.* **23**, 4555–4573. <https://doi.org/10.1111/mec.12811> (2014).

## Acknowledgements

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

## Author contributions

Project conceptualisation (W.T.T., A.P., S.M., S.V., K.G., J.C., P.D.B.). Data analysis (S.E., W.T.T., K.G., D.K., S.V., T.W.). Manuscript preparation (S.E., W.T.T., K.G., S.M., T.W., J.C., A.P., S.V., D.K., P.D.B.). Funding sourced by (P.D.B., J.C., S.M.).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-87107-z>.

**Correspondence** and requests for materials should be addressed to W.T.T.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021