Insect Biochemistry and Molecular Biology 110 (2019) 112–120

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/ibmb

Genome of the African cassava whitefly *Bemisia tabaci* and distribution and genetic diversity of cassava-colonizing whiteflies in Africa

Wenbo Chen[a,1], Everlyne N. Wosula[b,1], Daniel K. Hasegawa[c], Clerisse Casinga[d], Rudolph R. Shirima[b], Komi K.M. Fiaboe[e], Rachid Hanna[e], Apollin Fosto[e], Georg Goergen[f], Manuele Tamò[f], George Mahuku[b], Harun M. Murithi[b], Leena Tripathi[g], Bernard Mware[g], Lava P. Kumar[h], Pheneas Ntawuruhunga[i], Christopher Moyo[j], Marie Yomeni[k], Stephen Boahen[l], Michael Edet[m], Wasiu Awoyale[m], William M. Wintermantel[n], Kai-Shu Ling[c], James P. Legg[b,**], Zhangjun Fei[a,o,*]

[a] *Boyce Thompson Institute, Cornell University, Ithaca, NY, 14853, USA*
[b] *International Institute of Tropical Agriculture, Dar es Salaam, Tanzania*
[c] *U.S. Department of Agriculture-Agricultural Research Service, U.S. Vegetable Laboratory, Charleston, SC, 29414, USA*
[d] *International Institute of Tropical Agriculture, Bukavu-Kalambo, Democratic Republic of the Congo*
[e] *International Institute of Tropical Agriculture, Yaòunde, Cameroon*
[f] *International Institute of Tropical Agriculture, Cotonou, Benin*
[g] *International Institute of Tropical Agriculture, Nairobi, Kenya*
[h] *International Institute of Tropical Agriculture, Ibadan, Nigeria*
[i] *International Institute of Tropical Agriculture, Lusaka, Zambia*
[j] *International Institute of Tropical Agriculture, Lilongwe, Malawi*
[k] *International Institute of Tropical Agriculture, Freetown, Sierra Leone*
[l] *International Institute of Tropical Agriculture, Nampula, Mozambique*
[m] *International Institute of Tropical Agriculture, Monrovia, Liberia*
[n] *U.S. Department of Agriculture-Agricultural Research Service, Crop Improvement and Protection Research, Salinas, CA, 93905, USA*
[o] *U.S. Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, 14853, USA*

ARTICLE INFO

Keywords:
Cassava whitefly
Genome assembly
SNP genotyping
Distribution
Genetic diversity

ABSTRACT

The whitefy *Bemisia tabaci*, a species complex consisting of many morphologically indistinguishable species divided into distinct clades, is one of the most globally important agricultural pests and plant virus vectors. Cassava-colonizing *B. tabaci* transmits viruses that cause cassava mosaic disease (CMD) and cassava brown streak disease (CBSD). Half of all cassava plants in Africa are affected by these viral diseases, resulting in annual production losses of more than US$ 1 billion. Here we report the draft genome of the cassava whitefly *B. tabaci* Sub-Saharan Africa - East and Central Africa (SSA-ECA), the super-abundant population that has been associated with the rapid spread of viruses causing the pandemics of CMD and CBSD. The SSA-ECA genome assembled from Illumina short reads has a total size of 513.7 Mb and a scaffold N50 length of 497 kb, and contains 15,084 predicted protein-coding genes. Phylogenetic analysis suggests that SSA-ECA diverged from MEAM1 around 5.26 million years ago. A comprehensive genetic analysis of cassava-colonizing *B. tabaci* in Africa was also conducted, in which a total of 243 whitefly specimens were collected from 18 countries representing all major cassava-growing regions in the continent and genotyped using NextRAD sequencing. Population genomic analyses confirmed the existence of six major populations linked by gene flow and inferred the distribution patterns of these populations across the African continent. The genome of SSA-ECA and the genetic findings provide valuable resources and guidance to facilitate whitefly research and the development of strategies to control cassava viral diseases spread by whiteflies.

* Corresponding author. Boyce Thompson Institute, Cornell University, Ithaca, NY, 14853, USA.
** Corresponding author.
E-mail addresses: j.legg@cgiar.org (J.P. Legg), zf25@cornell.edu (Z. Fei).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ibmb.2019.05.003
Received 27 March 2019; Received in revised form 6 May 2019; Accepted 14 May 2019
Available online 15 May 2019

1. Introduction

The whitefly *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae) is an agricultural pest of ornamental, vegetable, grain, legume, and cotton crops, causing damage directly through feeding and indirectly through the transmission of plant pathogenic viruses belonging to at least five genera (*Begomovirus*, *Crinivirus*, *Ipomovirus*, *Carlavirus* and *Torradovirus*). In Africa, the rapid geographical expansion of a pandemic of severe cassava mosaic disease (CMD) has devastated cassava crops in 12 countries in East and Central Africa (Legg et al., 2006, 2011). CMD is caused by cassava mosaic begomoviruses (Bock and Woods, 1983) that are transmitted by *B. tabaci* (Legg et al., 2011). In the past decade, cassava brown streak disease (CBSD) has similarly spread as a pandemic through East and Central Africa (Alicai et al., 2007; Legg et al., 2011). CBSD is caused by two species of *Ipomovirus* (Mbanzibwa et al., 2011) that are also transmitted by *B. tabaci* (Maruthi et al., 2005, 2017), and has emerged as a risk to food security and a major threat to the production of cassava (Rey and Vanderschuren, 2017). The two diseases affect approximately half of cassava plants in Africa, resulting in annual production losses of more than US$ 1 billion (Legg et al., 2014). Over a 5-year period, 10–250-fold increases in *B. tabaci* abundance on cassava occurred in Uganda, Kenya, Tanzania and Burundi, leading to them being referred to as the 'super-abundant' whitefly (Legg et al., 2014). The rapid increase in *B. tabaci* abundance has been associated with the spread of CMD (Legg and Ogwal, 1998; Otim-Nape et al., 1996), and has driven the outbreaks of CBSD (Legg et al., 2011).

Bemisia tabaci is considered to be a species complex consisting of many morphologically indistinguishable species divided into distinct clades (Liu et al., 2012). Species-level delimitation has been proposed based on the mitochondrial cytochrome oxidase I (mtCOI) sequence, with a $> 3.5\%$ divergence level used to separate putative species (Dinsdale et al., 2010). Sub-Saharan Africa (SSA) was considered to comprise five genetically distinct groups of cassava-colonizing *B. tabaci* (SSA1-5) (Berry et al., 2004; Esterhuizen et al., 2013). Based on the mtCOI sequence divergence, SSA1 has been further divided into five subgroups; SSA1 subgroup 1 (SSA1-SG1), SSA1-SG2, SSA1-SG3, SSA1-SG4 (Legg et al., 2014) and SSA1-SG5 (Ghosh et al., 2015). Recently, analysis of $> 7{,}000$ genome-wide single nucleotide polymorphisms (SNPs) obtained from cassava-colonizing *B. tabaci* populations from eight African countries, however, has led to the proposal for an alternative classification (Wosula et al., 2017). Although this supports the unique identity of SSA whiteflies, it suggests the occurrence of six major genetic groups which differ from the groupings based on the mtCOI sequences. SSA-East and Central Africa (SSA-ECA) is the group that co-occurs with the pandemics of severe CMD and CBSD, and is the genotype frequently occurring in super-abundant populations (Wosula et al., 2017).

Recently, genome sequences of the *B. tabaci* MEAM1 and MED have been reported (Chen et al., 2016; Xie et al., 2017). However, the relatively high divergence between MEAM1/MED and the cassava whitefly (Wosula et al., 2017) has limited the use of the MEAM1/MED genomes as references for the cassava whitefly. Here we report the *de novo* assembly and annotation of the genome of the super-abundant cassava whitefly *B. tabaci* SSA-ECA, which provides a valuable resource to facilitate the development of efficient strategies to control this pest and the cassava viruses it transmits. We previously reported SNP-genotyping through NextRAD sequencing of cassava whitefly samples that were collected from eight cassava-growing countries (Wosula et al., 2017). To obtain a more comprehensive understanding of the distribution and genetic diversity of cassava-colonizing *B. tabaci* in Africa, whitefly samples were collected from cassava in 10 additional countries and four countries that either had few samples in our previous study or harbored high whitefly diversity. SNPs were called from the combined samples using the SSA-ECA genome as the reference. This study therefore presents the most comprehensive characterization of the genetics of cassava-colonizing *B. tabaci* populations in Africa to date, as well as the genome sequence of SSA-ECA, the genotype directly linked to super-abundance and the spread of the severe CMD and CBSD pandemics.

2. Materials and methods

2.1. Sampling and genome sequencing

Over 10,000 cassava-colonizing whiteflies were collected from cassava plants in a single field in Murumba Village, Chato District, northwestern Tanzania, a location on the southwestern shores of Lake Victoria which is characterized by abundant *B. tabaci* populations and rapid spread of CBSD. Adult *B. tabaci* were aspirated into vials, immediately killed and preserved in 90% ethanol before being transported to the laboratory. Using a stereoscope, $\sim 1{,}050$ haploid male whiteflies were isolated. Genomic DNA was prepared from these whiteflies using the same protocol described in Chen et al. (2016). These whiteflies were confirmed to belong to the SSA-ECA group through PCR amplification with primers specific to the mtCOI gene as well as phylogenetic analysis with ~ 90 diverse cassava-colonizing *B. tabaci* populations using genome-wide SNPs (Wosula et al., 2017). Paired-end libraries were constructed using the Illumina TruSeq DNA sample preparation kit, and mate-pair libraries were constructed using the Nextera Mate Pair Sample Preparation kit, following the manufacturer's instructions (Illumina, San Diego, CA). All libraries were sequenced on an Illumina HiSeq 2500 system.

2.2. Genome assembly

Raw Illumina reads were processed to remove duplicated read pairs, which were defined as having identical bases in the first 100 bp of both left and right reads, and only one read pair from the duplicates was kept. Illumina adaptor and low-quality sequences were removed from the reads using Trimmomatic (Bolger et al., 2014). *De novo* assembly was performed with Platanus (Kajitani et al., 2014), using high-quality cleaned reads from both paired-end and mate-pair libraries. The assembled scaffolds were compared against the NCBI non-redundant nucleotide (nt) database using BLASTN with an e-value cutoff of 1e-5. Scaffolds with over 90% of their lengths similar to only bacterial or viral sequences were considered to be contaminants and therefore discarded. To further improve the assembly, we used Pilon (Walker et al., 2014) to correct base errors, fix mis-assemblies and fill gaps.

2.3. Transcriptome sequencing

To improve gene prediction, RNA-Seq was conducted on *B. tabaci* SSA-ECA from three different developmental stages (adults, nymphs and eggs) collected from cassava plants in a single location in northwestern Tanzania (Murumba village, Chato District). Total RNA was extracted using the Ambion TRIzol Reagent (Thermo Fisher, USA) following the manufacturer's instructions. Strand-specific RNA-Seq libraries were constructed using the protocol described in Zhong et al. (2011) and sequenced on the Illumina HiSeq 2500 system. Raw reads were processed by trimming adaptor and low-quality sequences using Trimmomatic (Bolger et al., 2014). The cleaned reads were aligned to the assembled SSA-ECA genome using HISAT2 (Kim et al., 2015), followed by reference-guided assembly using StringTie (Pertea et al., 2015). The assembled transcripts were used to improve protein-coding gene predictions in the *B. tabaci* SSA-ECA genome.

2.4. Annotation of repetitive elements

A MITE (miniature inverted-repeat transposable element) library and a *de novo* repeat library were constructed by scanning the assembled *B. tabaci* SSA-ECA genome using MITE-Hunter (Han and

Wessler, 2010) and RepeatModeler (http://www.repeatmasker.org/RepeatModeler), respectively. The identified repeats in the libraries were subsequently compared against the NCBI non-redundant (nr) protein database using BLAST with an e-value cutoff of 1e-5, and those having hits to known protein sequences were removed. Finally, the *de novo* repeat library and the MITE library were used to identify repeat sequences in the assembled SSA-ECA genome by scanning the genome using RepeatMasker (http://www.repeatmasker.org/) and the RepeatRunner (http://www.yandell-lab.org/software/repeatrunner.html) subroutine in the MAKER annotation pipeline (Cantarel et al., 2008).

2.5. Gene prediction

Protein-coding genes were predicted from the repeat-masked genome assembly of *B. tabaci* SSA-ECA using MAKER (Cantarel et al., 2008), which synthesizes results from *ab initio* gene predictions with experimental gene evidence to produce final consensus gene models. The evidence used in this study included full coding sequences (CDS) of *B. tabaci* collected from NCBI, transcripts assembled from our RNA-Seq data, completed proteomes of fruit fly, pea aphid and whitefly MEAM1, and proteins from Swiss-Prot. All these sequences were aligned to the SSA-ECA genome using Spaln (Gotoh, 2008). MAKER was used to run a battery of trained gene predictors including Augustus (Stanke and Waack, 2003), BRAKER (Hoff et al., 2015) and Snap (Korf, 2004), and then integrated with the experimental gene evidence, to produce evidence-based gene predictions.

To functionally annotate the predicted genes, their protein sequences were searched against different protein databases including UnitProt (TrEMBL/SwissProt) and two insect proteomes (whitefly MEAM1 and fruit fly) using BLAST with an e-value cutoff of 1e-4. The protein sequences were also compared against the InterPro domain database (Mitchell et al., 2015). GO annotation was performed with Blast2GO (Conesa et al., 2005).

2.6. Comparative genomics

We compared the predicted *B. tabaci* SSA-ECA genes with those of MEAM1 and 14 other arthropods (*Apis mellifera*, *Camponotus floridanus*, *Tribolium castaneum*, *Bombyx mori*, *Danaus plexippus*, *Anopheles gambiae*, *Drosophila melanogaster*, *Pediculus humanus*, *Diaphorina citri*, *Acyrthosiphon pisum*, *Rhodnius prolixus*, *Nilaparvata lugens*, *Daphnia pulex* and *Tetranychus urticae*). The proteome sequences of all 16 arthropod species were used to construct orthologous groups using OrthoMCL (Li et al., 2003). The 668 single-copy genes shared by all 16 species were used to reconstruct their phylogenetic relationships. Briefly, protein sequences of the single-copy genes were aligned with MUSCLE (Edgar, 2004), and the positions in the alignment containing gaps in more than 20% of the sequences were removed by trimAl (Capella-Gutierrez et al., 2009). A phylogenetic tree was then constructed using the maximum-likelihood method implemented in PhyML (Guindon et al., 2009), with the JTT model for amino acid substitutions and the aLRT method for branch support. The alignment was also used to estimate divergence times among lineages using mcmctree in the PAML package (Yang, 2007). Fossil calibrations were set according to previous studies (Benton and Donoghue, 2006; Donoghue and Benton, 2007).

2.7. Genotyping of cassava-colonizing whiteflies with NextRAD sequencing

We previously collected 95 cassava-colonizing whitefly samples from eight African countries (Burundi, Cameroon, Central African Republic (CAR), Democratic Republic of Congo (DRC), Madagascar, Nigeria, Rwanda and Tanzania) (Wosula et al., 2017). In this study, an additional 190 adult whitefly specimens were collected randomly between 2015 and 2018 from 10 additional countries (Benin, Ghana, Kenya, Liberia, Malawi, Mozambique, Sierra Leone, Togo, Uganda and Zambia) and from four countries (Cameroon, DRC, Nigeria, Tanzania) that either had few samples in our previous study or harbored high whitefly diversity. Detailed information on locations of sampling within countries and time of collection is provided in Table S1. The whiteflies were aspirated live from cassava plants and then preserved in 95% ethanol prior to storage at −20 °C. DNA extraction was carried out following the procedures described in Wosula et al. (2017).

Genomic DNA of the 190 whiteflies was used to construct NextRAD libraries by SNPsaurus, LLC (http://snpsaurus.com/) as described in Russello et al. (2015). Raw reads from NextRAD sequencing (190 samples combined with 95 samples that were sequenced in Wosula et al., 2017) were first processed to remove adaptor and low-quality sequences. The cleaned reads were then aligned to the SSA-ECA genome, and only uniquely mapped reads were used for SNP calling using TASSEL5 (Bradbury et al., 2007). The resulting raw SNPs were filtered by the following criteria: 1) individual samples with missing data rate > 65% were excluded; 2) SNPs with missing data in > 20% of the samples or minor allele frequency (MAF) < 0.05 were removed; 3) SNPs with genotype quality (GQ) < 30 or SNPs with another SNP of < 5 bp away were excluded.

The final filtered SNPs (63,770) were used to construct a maximum-likelihood phylogenetic tree using phyML (Guindon et al., 2010) with default parameters. STRUCTURE (v2.3.4) (Hubisz et al., 2009) was used to perform a model-based clustering for inferring population structure. Twenty independent runs for each K value ranging from 1 to 10 were performed with a burn-in length of 10,000 followed by 10,000 iterations, where K is the assumed number of populations. The best K was deduced from the distribution of ΔK. The optimal K was implemented in a final run with 100,000 burn-in and 100,000 iterations. Principal component analysis (PCA) was performed using PLINK (v1.9) (Chang et al., 2015), and the result was illustrated using the ggplot2 package in R.

2.8. Gene flow analysis

We performed gene flow analysis using the D-statistics (Patterson et al., 2012), which were calculated using the AdmixTools v5.1 (Patterson et al., 2012). This program makes use of the tree structure of (out group, x; y, SSA4), where 'out group' included *Bemisia afer* and sweetpotato *B. tabaci* whitefly (non-cassava haplotype). Under the assumption of the model, there is no gene flow between 'out group' and SSA4, but there is potential gene flow between either population x and y or x and SSA4, which results in negative or positive D, respectively. D = 0 indicates a lack of gene flow between the two populations. Significant deviation from 0 of D is estimated by the Z-score, which is considered to be significant if |Z-score| > 4.

To infer directionality of gene flow between different populations, we performed the partitioned *D*-statistic test, which is based on a five-taxon tree (((P1, P2), ($P3_1$, $P3_2$)), O), where $P3_1$ and $P3_2$ are two lineages within the P3 clade (Eaton and Ree, 2013). The test can infer directionality through its measurement of introgression of shared ancestral alleles, D12. If gene flow occurs from $P3_1$ into P2, then the derived P3 alleles which arise in the ancestor of $P3_1$ and $P3_2$, and are thus shared by both taxa, would also appear in P2. In contrast, if gene flow occurs only in the opposite direction, from P2 into $P3_1$, then P2 will not contain alleles that are shared by the two P3 taxa, and thus the partitioned test would find a non-significant D12. The partitioned *D*-statistic test was also calculated using the AdmixTools v5.1 (Patterson et al., 2012).

2.9. GIS mapping

Geo-referenced coordinates for whitefly samples were used to generate maps using ArcGIS 10.1 (ESRI, Redlands, California, USA). Maps were produced illustrating the geographic distributions of cassava-colonizing *B. tabaci* whiteflies across sampled countries in Africa.

3. Results and discussion

3.1. Genome of the cassava whitefly B. tabaci SSA-ECA

The *B. tabaci* SSA-ECA genome was sequenced using the Illumina technology. A total of three paired-end libraries with insert sizes of 350 bp, 550 bp and 850 bp and three mate-pair libraries with insert sizes of 3 Kb, 8 Kb and 15 Kb were constructed and sequenced, which yielded about 255 Gb of raw sequence data. After removing duplicates, adaptor and low-quality sequences, a total of 126.5 Gb cleaned sequences were obtained (Table S2). K-mer distribution of the cleaned sequences displayed a very weak peak (Fig. S1), indicating the possible heterogeneity of the sequenced sample, which would hinder the *de novo* assembly of the genome. Therefore, we further used Quake (Kelley et al., 2010) to reduce the heterogeneity in the sequences by correcting the low-frequency reads with higher-frequency reads. After this correction, a more obvious peak was observed in the k-mer distribution plot (Fig. S1), supporting reduced heterogeneity in the corrected reads. Finally, we obtained 65 Gb of cleaned and corrected high-quality sequences that were used in the assembly.

The final assembled genome of *B. tabaci* SSA-ECA contained 71,393 scaffolds with an N50 length of 497 kb and a total length of 513.7 Mb, both of which were shorter than the published genome of *B. tabaci* MEAM1 (Chen et al., 2016), which is not unexpected given the heterogeneous nature of SSA-ECA used for genome sequencing, while the N50 length of the assembled SSA-ECA scaffolds was similar to that of the MED scaffolds (Table 1). We evaluated the completeness of the gene content in the draft genome assembly of SSA-ECA using BUSCO (Simao et al., 2015), which indicated that the assembly captured 95.3% of the core eukaryotic genes and 90.2% were completely covered by the assembly. Our results suggest that although a relatively high portion (~25%) of the SSA-ECA genome was missing in the assembly, the majority of the gene space was successfully assembled.

A total of 39.6% of the assembled genome was annotated as repeat elements (Table S3), which was slightly less than that in the MEAM1 genome (Table 1). The most predominant repeat elements were MITEs, which occupied 24.8% of the genome. In addition, 11.9% of genome was annotated as unknown repeats, which were unable to be classified into any known families.

A total of 15,084 protein-coding genes were predicted in the SSA-ECA genome, which is similar to the number of genes predicted in MEAM1. The gene structure of SSA-ECA, including the average lengths of CDS and exons, and the number of exons per gene, is also similar to that of MEAM1 and MED (Table 1). Among the 15,084 predicted SSA-ECA genes, 73.9% had hits to proteins in the UniProt (TrEMBL/Swiss-Prot) database, 42.3% were annotated with GO terms, 68.7% contained InterPro domains, and 87.2%, 79.5% and 55.0% of the SSA-ECA predicted genes shared detectable homology with MEAM1, MED and fruit fly genes, respectively.

Table 1
Genome assembly statistics of African cassava whitefly SSA-ECA.

	SSA-ECA	MED	MEAM1
Genome assembly			
Assembled genome size (Mb)	513.7	658.2	615.0
Gap (%)	4.9	3.0	2.3
Scaffold N50 (bp)	497,869	436,791	3,232,964
Scaffold L50	218	421	56
Contig N50 (bp)	10,224	44,366	29,876
Contig L50	12,381	4,288	5,762
Minimum length (bp)	500	501	500
Genomic features			
GC content (%)	39.3	39.5	39.6
Repeat (%)	39.6	40.0	43.2
Number of protein-coding genes	15,084	20,786	15,664
Mean coding sequence length (bp)	1,339.57	1,505.5	1,469.80
Mean number of exons per gene	6.19	6.0	6.73
Mean exon length (bp)	365.97	351.0	421.1
Mean intron length (bp)	2,310.58	1,776.0	1,875.50
BUSCO evaluation - % present (complete)			
Genome[a]	85.8 (78.0)	82.3 (78.0)	86.7 (84.1)
Protein[a]	95.3 (90.2)	92.1 (88.3)	96.8 (94.4)

[a] BUSCO was run against the assembled genome sequences and the predicted proteins, respectively. Numbers shown are percentages of core eukaryotic genes covered by the assemblies or proteins, with percentages of genes that were completely covered shown in parenthesis.

Table 2
Summary statistics of predicted detoxification-related genes.

	SSA-ECA	MEAM1	MED
Cytochrome P450 (CYP)	130	130	153
UDP-glucuronosyltransferase (UGT)	67	81	63
Glutathione S-transferase (GST)	25	22	21
Carboxylesterase (COE)	42	51	51
ABC transporter (ABC)	48	50	59
Cathepsin	84	111	167
Phosphatidylethanolamine-binding protein (PEBP)	98	202	149
Aromatic peroxygenase (APO)	15	20	17
Total	509	667	680

3.2. Detoxification of xenobiotic compounds

Several enzyme families have been widely reported to be implicated in detoxification of xenobiotic compounds such as plant secondary metabolites and insecticides. These families include cytochrome P450 (CYP), UDP-glucuronosyltransferase (UGT), glutathione S-transferase (GST), ABC transporter (ABC), and carboxylesterase (COE). As shown in Table 2, the SSA-ECA genome contains 130 predicted CYPs, the same number as detected in MEAM1, but fewer than the amount (153) found in MED. The SSA-ECA genome also encodes 67 predicted UGTs, which is similar to MED (63) but less than MEAM1 (81). Furthermore, a total of 25 GSTs, 42 COE and 48 ABCs were predicted in the SSA-ECA genome, all of which are similar to or fewer than the corresponding amounts identified in MEAM1 and MED.

Recently, several additional gene families have been found to be involved in insecticide resistance in *B. tabaci*, including cathepsin, phosphatidylethanolamine-binding protein (PEBP) and aromatic peroxygenase (APO) (Chen et al., 2016). A total of 84 cathepsin and 98 PEBP genes were predicted in the SSA-ECA genome, which is substantially less than the amounts found in MEAM1 and MED (Table 2). In addition, 15 APO genes were predicted in SSA-ECA, similar to the amounts detected in MEAM1 and MED.

Overall, SSA-ECA has a reduced set of predicted genes related to detoxification of xenobiotic compounds compared to MEAM1 or MED. This is consistent with the fact that the SSA-ECA has a much narrower host range (largely restricted to cassava) than the more polyphagous MEAM1 and MED.

3.3. Comparative genomics

We constructed orthologous groups by comparing the complete set of SSA-ECA protein-coding genes with those of MEAM1 and 14 other arthropods. A total of 2,535 orthologous groups were shared by all 16 species, including 668 single-copy orthologous genes. These single-copy genes were used to reconstruct their phylogenetic relationships, which were consistent with previous studies (Chen et al., 2016; Xie et al., 2017; Xue et al., 2014). The divergence time between SSA-ECA and MEAM1 was estimated to be around 5.26 million years ago (Mya) (Fig. 1A). This is different from the divergence time of 50–70 Mya between sub-Saharan Africa and other *B. tabaci* species complex reported by Boykin et al. (2013), who used a 657-bp fragment at the 3'

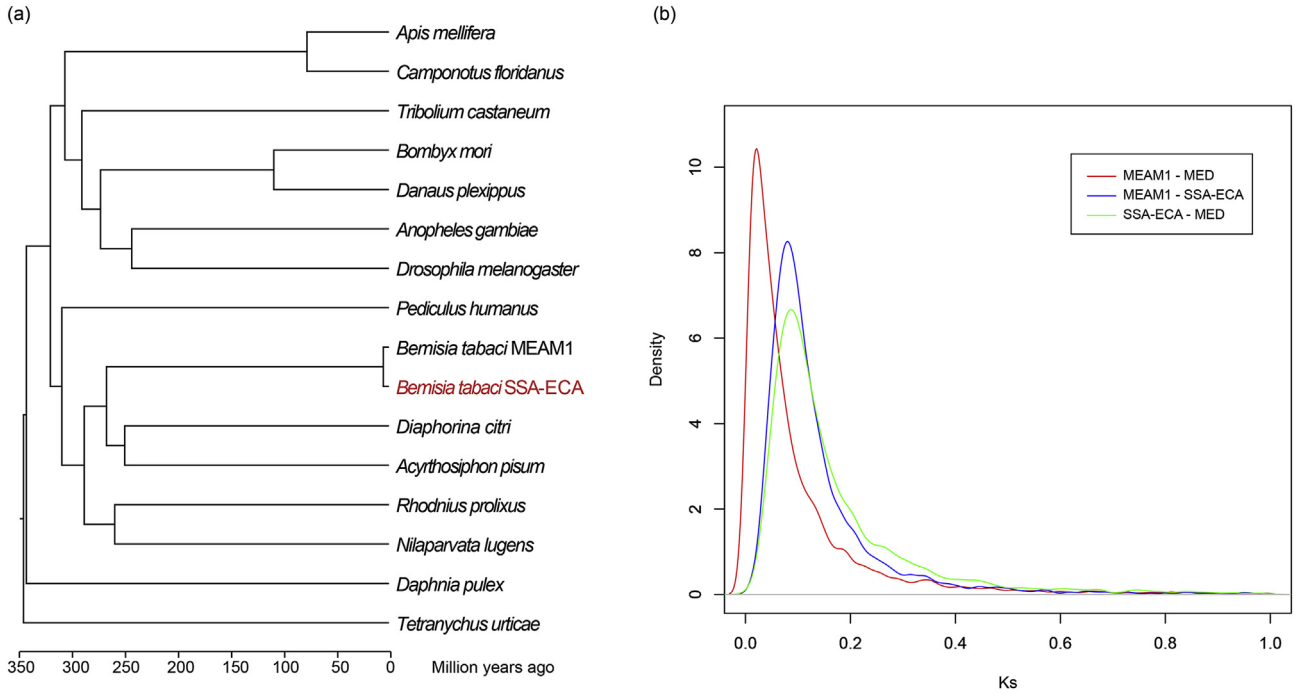

Fig. 1. Phylogenomics of cassava whitefly *Bemisia tabaci* SSA-ECA. (a) Phylogenetic relationships of SSA-ECA and 15 other arthropod species. *Daphnia pulex* and *Tetranychus urticae* were used as the outgroup taxa. Branch length represents the divergence time. (b) Distribution of synonymous substitution rate (Ks) of orthologous gene pairs between MEAM1 and MED, SSA-ECA and MEAM1, and SSA-ECA and MED.

end of the single mtCOI gene to infer the divergence time. We argue that a comparative genomic approach provides a more accurate picture of the evolutionary history of the *Bemisia* complex species. In addition, the low synonymous nucleotide substitution rates (*Ks*) of orthologous gene pairs between SSA-ECA and MEAM1 (peaked at 0.079; see below) further supported the relatively recent divergence between these two whiteflies.

We further constructed orthologous groups among SSA-ECA, MEAM1 and MED whiteflies, which resulted in the identification of 8,166 groups shared by all three whiteflies. Ks of orthologous gene pairs between MEAM1 and SSA-ECA peaked at 0.079, which was similar to that between SSA-ECA and MED (0.088), but greater than that between MEAM1 and MED (0.021) (Fig. 1B).

3.4. Genetic diversity of cassava-colonizing B. tabaci in Africa

SNP-genotyping with NextRAD sequencing was used to examine the comprehensive and geographically extensive set of cassava-colonizing *B. tabaci* in Africa. SNP-genotyping was done using 243 *B. tabaci* cassava whiteflies which produced quality sequences out of the combined 285 samples. Maximum-likelihood phylogenetic analysis (Fig. 2) and PCA (Fig. 3A) revealed the existence of six major haplogroups (SSA2, SSA4, SSA-CA, SSA-ESA, SSA-WA and SSA-ECA). This is consistent with our previous study despite the fact that the earlier study had only 7,453 SNPs, with the MEAM1 genome used as the reference (Wosula et al., 2017), compared to this study, which identified 63,770 SNPs using the SSA-ECA genome as the reference. Haplogroup SSA2 comprised 35 (15%) samples, and they were from Cameroon, Kenya, Ghana, DRC and Sierra Leone. Although SSA2 is a single haplogroup, it appears to have three sub-groups comprising samples from Cameroon (Central Africa), Ghana and Sierra Leone (West Africa), and eastern DRC and Kenya (East Africa), respectively. This study confirms that SSA2 from Kenya belongs to the same haplogroup as SSA2 from Cameroon, and demonstrates that SSA2 is present in East, Central and West Africa, but not southern Africa. Consequently, it is the most geographically dispersed group. Furthermore, across most of its range, it appears to be a secondary population that does not predominate. This suggests that it may have a specific biological niche that enables it to succeed when co-occurring with other haplogroups. SSA4 comprised 18 (7.4%) samples, and they were from Cameroon, DRC and CAR. This population seemed to be restricted to Central Africa. Haplogroup SSA-CA consisted of 13 (5.3%) samples from eastern DRC and western Tanzania. This population was restricted to the regions around Lake Tanganyika in Tanzania (Kigoma) and DRC (Katanga and Tanganyika provinces). SSA-ESA comprised 56 (23.0%) samples collected from Tanzania, Kenya, Mozambique, Malawi and Zambia. There were two subgroups in this population, with samples from Tanzania and Kenya grouping in a separate cluster from those of Malawi and Mozambique. This population was restricted to southern Africa and coastal East Africa, although three samples from Kigoma (northwestern Tanzania) were also identified in this haplogroup. SSA-WA comprised 52 (21.4%) samples from Nigeria, Sierra Leone, Togo, Benin, Ghana, Liberia and CAR. This population predominated in West Africa. Haplogroup SSA-ECA comprised 69 (28.4%) samples collected from Tanzania, Uganda, Kenya, Burundi, Rwanda and eastern DRC. This population was predominant in the regions of East and Central Africa currently affected by the CBSD pandemic, and previously the CMD pandemic.

Population structure analysis revealed that the six major haplogroups have their origins from five ancestral populations (K = 5) (Fig. 3B). Haplogroups SSA-ECA, SSA-WA, SSA-ESA and SSA2 have a relatively homogenous genetic background while SSA4 and SSA-CA are extensively admixed comprising signatures from all five ancestral populations. SSA2 (including samples from Kenya) is a population that is distinct from SSA-ECA and therefore the latter is unlikely to be a hybrid of SSA2 and some other population as had been hypothesized (Legg et al., 2002). Gene flow analyses revealed a similar pattern to that reported by Wosula et al. (2017). There was significant gene flow from SSA-WA to SSA-ECA, SSA-ECA to SSA-CA, SSA-CA to SSA4, and SSA-ESA to SSA-CA (Tables S4 and S5). Significant gene flow also occurred between SSA2 and SSA4 although the direction of flow was not determined. Further analyses excluding SSA2 from Cameroon showed no evidence of gene flow between SSA2 from Kenya and SSA-ECA. This

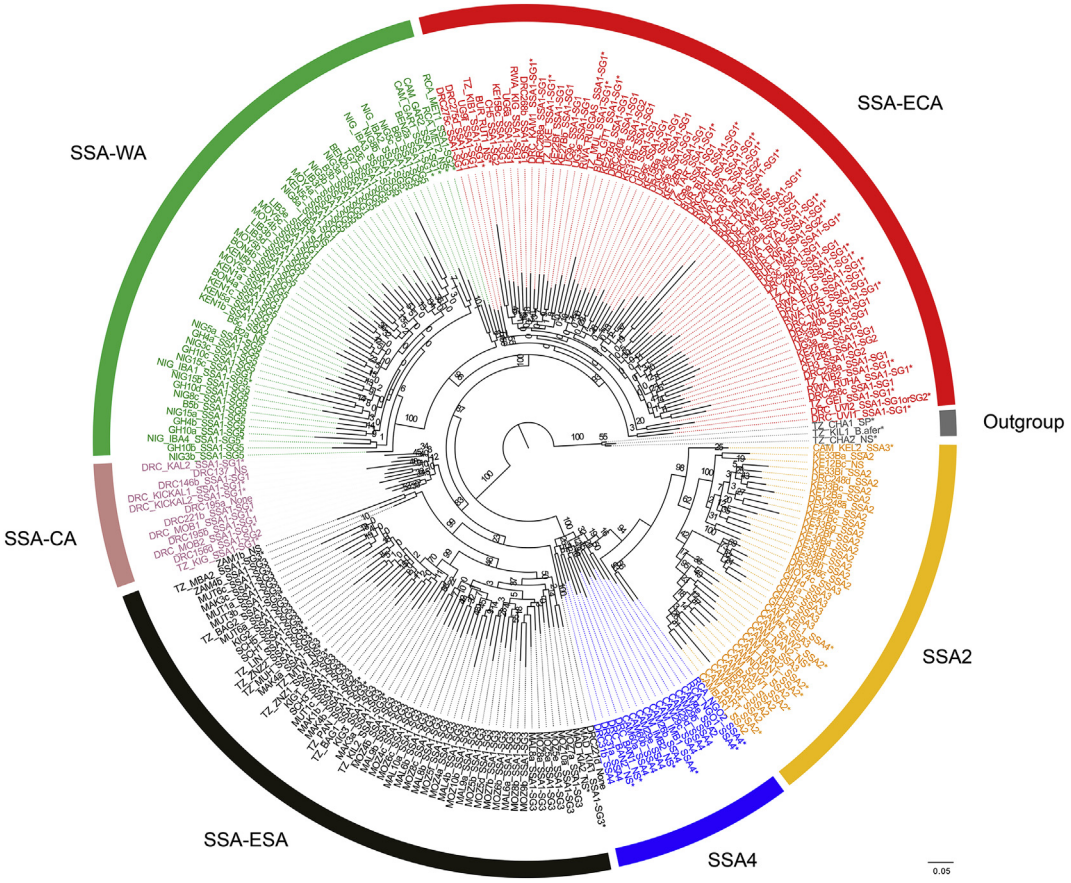

Fig. 2. Maximum-likelihood phylogenetic tree constructed based on SNPs (63,770) generated by NextRAD sequencing of *Bemisia tabaci* (cassava and non-cassava haplotypes) and *B. afer* adults sampled between 2009 and 2018 from eighteen countries in Africa. Samples designated (*) are published (Wosula et al., 2017).

confirms that SSA-ECA is a distinct population that displaced SSA2 in East Africa. The existence of six haplogroups as previously reported in Wosula et al. (2017), with samples covering almost all of the major cassava-growing countries in Africa, indicates that this is likely to represent most if not all of the major genetic groupings of *B. tabaci* occurring on cassava in the continent. Gene flow among these populations could lead to the emergence of novel haplogroups that have favorable traits such as rapid reproduction rate and increased adaptation to new environments; this could exacerbate virus epidemics through increased spread, severity and even emergence of new strains. Introgression of new alleles through hybridization has been shown to result in increased adaptive competitiveness and geographic range expansion of some hybridizing species (Wellenreuther et al., 2018). Since gene flow occurs from SSA-ECA, the superabundant population associated with severe virus epidemics (CMD and CBSD) in Eastern Africa, to SSA-CA, a population currently confined to DRC, introgression of alleles putatively associated with increased fitness of SSA-ECA could potentially lead to the rapid expansion of SSA-CA populations that would drive virus epidemics in new regions. An alternative scenario might be the displacement of one or more cryptic species by others, such as the example from Uganda in which SSA-ECA displaced SSA2. Displacement of local *B. tabaci* cryptic species by invasive groups, for example MEAM1 and MED, has been reported to have triggered severe virus epidemics or the emergence of new virus strains or viruses in affected crops and even the emergence of new viruses in previously unaffected host plants (Islam et al., 2018).

3.5. Geographical distribution of cassava B. tabaci

Based on SNP-genotyping, this study demonstrates that SSA2 has the most extensive distribution, ranging from Sierra Leone in West Africa to Kenya in the East. This range covers a huge diversity of ecological zones, indicating that this haplogroup must have a wide adaptation to diverse vegetation, climate and agroecological types. SSA-ECA and SSA-WA are found in the humid forest and derived savanna zones of Eastern Africa and West Africa, respectively. SSA4 and SSA-CA are restricted haplogroups, with the former confined in Cameroon and DRC and the latter in DRC (humid forest and savanna zones). SSA-ESA is found in savanna and semi-arid zones of Eastern and Southern Africa (Fig. 4). In view of its extensive distribution, SSA2 overlaps with three other haplogroups: SSA-ECA in East Africa, SSA4 in Central Africa and SSA-WA in West Africa (Fig. 4). The overlapping of populations in various zones could present opportunities for hybridization considering that all the six cassava whitefly populations are linked through gene flow. SSA2 overlaps in range with three other populations but gene flow occurs only with SSA4. The range of SSA-ECA, the population associated with severe virus epidemics in cassava, overlaps with three other populations but gene flow is only into SSA-CA. Su et al. (2017) demonstrated that hybrids of termites *Coptotermes gestroi* and *C. formosanus* possess temperature tolerance of both species and survived better at 15–35 °C. Although this study presents evidence for gene flow, and earlier studies demonstrated successful mating between different cassava-colonizing *B. tabaci* populations (Maruthi et al., 2001), further experiments will be required to prove mating compatibility and to determine the fitness outcomes for the progeny of crosses between the six major cassava-colonizing *B. tabaci* haplogroups. Results from such experiments could be used as the basis for modelling future changes in populations of each of these groups and their hybrid offspring, as well as the resulting virus spread and consequent impact on cassava production across the continent.

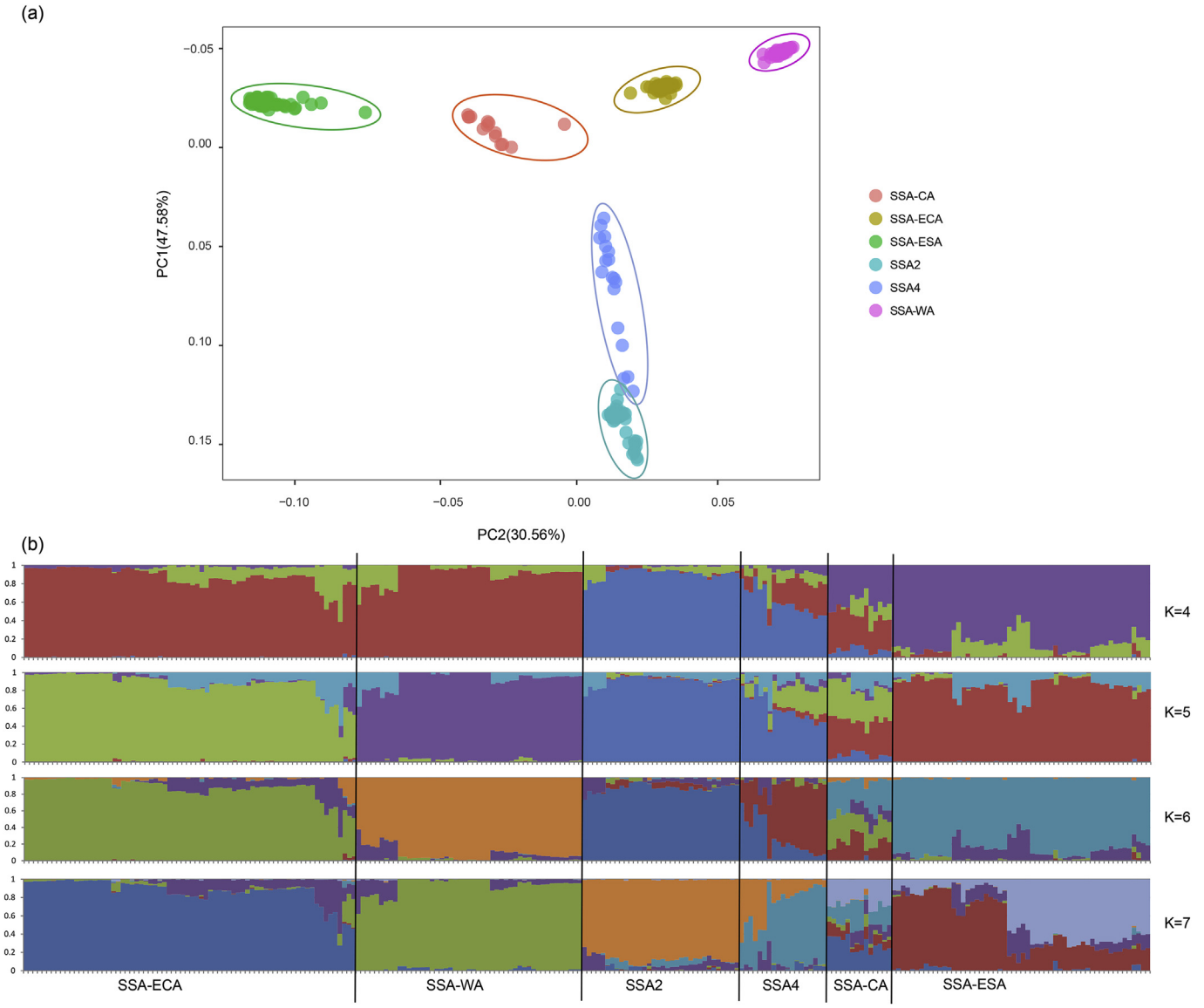

Fig. 3. Population structure of cassava-colonizing *Bemisia tabaci* in Africa. (a) Principal component analysis of 243 *B. tabaci* whiteflies (cassava haplotypes) collected from eighteen African countries. (b) Structure analysis of *B. tabaci* (cassava haplotypes) in Africa. The estimated optimal K is 5. The y axis quantifies subgroup membership, and the x axis shows different whitefly individuals.

4. Conclusion

Here we present the first draft genome assembly and annotation of the super-abundant cassava whitefly *Bemisia tabaci* SSA-ECA, and report that six major distinct populations of cassava *B. tabaci* exist in the major cassava-growing regions of Africa. These populations are all linked through gene flow and overlap in various agroecological zones. The combined effect of gene flow and overlapping distributions may lead to the emergence of novel populations that could further drive the virus epidemics in cassava either through increased spread or the emergence of novel strains. The genome assembly presented here will provide a valuable resource for studying the factors driving the super-abundance of cassava whiteflies, and for understanding the mechanisms underlying whitefly-virus interactions, transmission and disease development in cassava. It will also facilitate the development of novel strategies for whitefly and whitefly-transmitted virus control on cassava in Africa and other agriculturally important crops throughout the world. In addition, the SSA-ECA genome sequence allows for the development of robust, rapid and accurate SNP-based molecular tools for routine monitoring of the cassava-colonizing whiteflies in Africa.

Availability of supporting data

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession PGTP00000000. The version described in this paper is version PGTP01000000. The raw genome and RNA-Seq sequences have been deposited in the NCBI Short Sequence Archive (SRA) under accessions SRP125415 and SRP125413, respectively. The genome sequence and annotation are also available in the whitefly genome database (http://www.whiteflygenomics.org).

Author contributions

Z.F, J.P.L., K.S.L. and W.M.W. designed and managed the project. W.C. assembled and annotated the genome, performed the evolution and population genetics analysis, and drafted the manuscript. E.N.W. carried out the field collections, processed samples and drafted the manuscript. D.K.H. processed DNA and RNA samples for sequencing. C.C., R.R.S., K.K.M.F, R.H., A.F., G.G., M.T., G.M., H.M.M., L.T., B.M., L.P.K, P.N., C.M., M.Y., S.B., M.E. and W.A. carried out field collections. All authors read, revised, and approved the final manuscript.

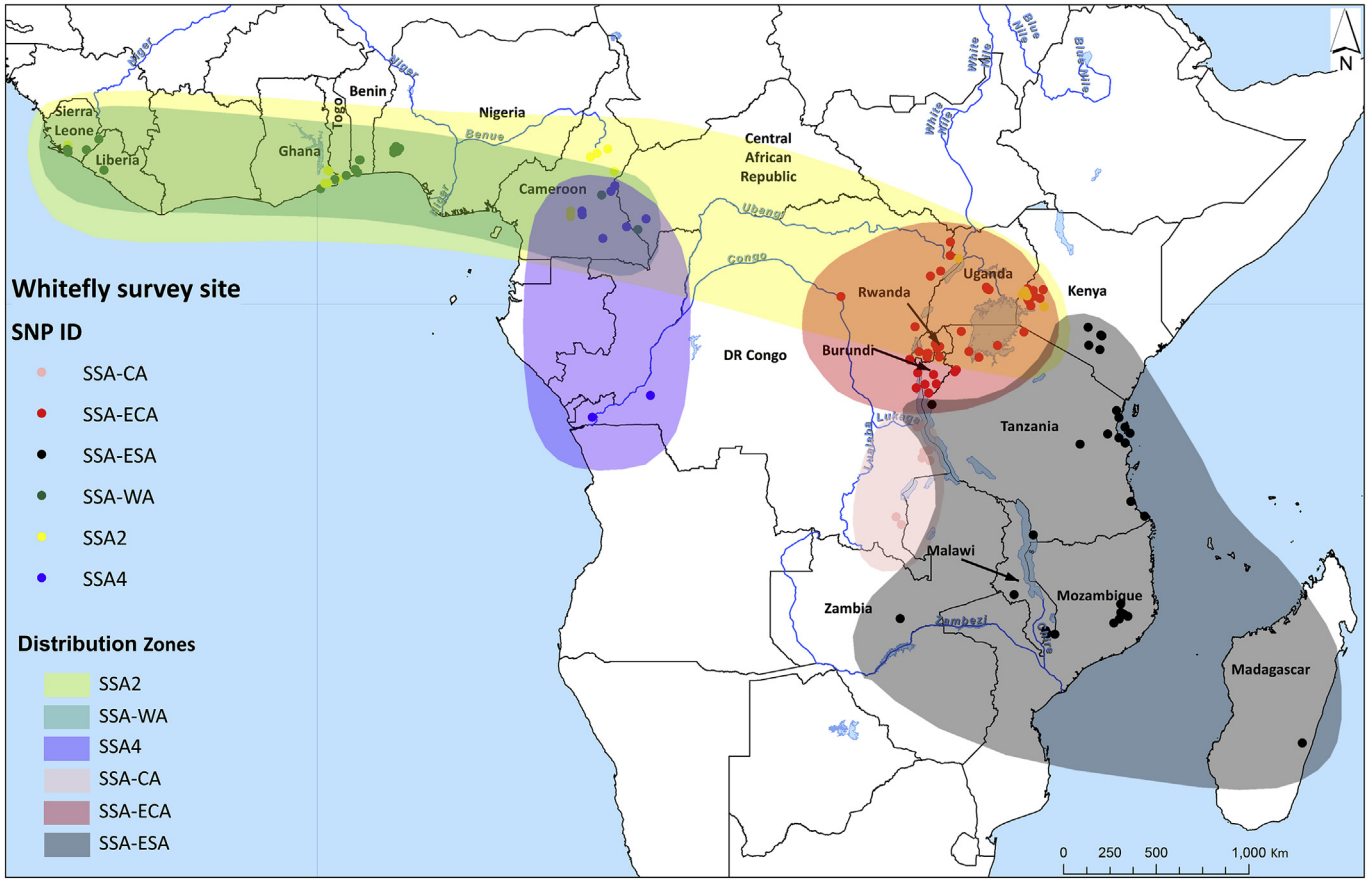

Fig. 4. Geographic distribution of cassava-colonizing *Bemisia tabaci* in Africa based on SNP genotyping.

Acknowledgments

This work was supported by grants from the USDA-ARS Office of International Research Programs from a grant provided by the USAID Feed-the-Future program (58-0210-3-012) to Z.F., J.P.L., K.S.L., W.M.W. and L.T., the USDA ARS Area-wide project as a part of the i5K initiative to K.S.L. and W.M.W., and the CGIAR Research Program for Roots, Tubers and Bananas to J.P.L. and E.N.W.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmb.2019.05.003.

References

Alicai, T., Omongo, C., Maruthi, M., Hillocks, R., Baguma, Y., Kawuki, R., Bua, A., Otim-Nape, G., Colvin, J., 2007. Re-emergence of cassava brown streak disease in Uganda. Plant Dis. 91, 24–29.

Benton, M.J., Donoghue, P.C., 2006. Paleontological evidence to date the tree of life. Mol. Biol. Evol. 24, 26–53.

Berry, S.D., Fondong, V.N., Rey, C., Rogan, D., Fauquet, C.M., Brown, J.K., 2004. 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.

Bock, K., Woods, R., 1983. Etiology of African cassava mosaic disease. Plant Dis. 67, 994–995.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Boykin, L.M., Bell, C.D., Evans, G., Small, I., De Barro, P.J., 2013. 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.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S., 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633–2635.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., Yandell, M., 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, 1.

Chen, W., Hasegawa, D.K., Kaur, N., Kliot, A., Pinheiro, P.V., Luan, J., Stensmyr, M.C., Zheng, Y., Liu, W., Sun, H., Xu, Y., Kruse, A., Yang, X., Kontsedalov, S., Lebedev, G., Fisher, T.W., Nelson, D.R., Hunter, W.B., Brown, J.K., Jander, G., Cilia, M., Douglas, A.E., Ghanim, M., Simmons, A.M., Wintermantel, W.M., Ling, K.S., Fei, Z., 2016. 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. 14, 110.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Dinsdale, A., Cook, L., Riginos, C., Buckley, Y., Barro, P.D., 2010. 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.

Donoghue, P.C., Benton, M.J., 2007. Rocks and clocks: calibrating the tree of life using fossils and molecules. Trends Ecol. Evol. 22, 424–431.

Eaton, D.A.R., Ree, R.H., 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: orobanchaceae). Syst. Biol. 62, 689–706.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Esterhuizen, L.L., Mabasa, K.G., Van Heerden, S.W., Czosnek, H., Brown, J., Van Heerden, H., Rey, M.E., 2013. Genetic identification of members of the *Bemisia tabaci* cryptic species complex from South Africa reveals native and introduced haplotypes. J. Appl. Entomol. 137, 122–135.

Ghosh, S., Bouvaine, S., Maruthi, M.N., 2015. Prevalence and genetic diversity of endosymbiotic bacteria infecting cassava whiteflies in Africa. BMC Microbiol. 15, 93.

Gotoh, O., 2008. Direct mapping and alignment of protein sequences onto genomic sequence. Bioinformatics 24, 2438–2444.

Guindon, S., Delsuc, F., Dufayard, J.F., Gascuel, O., 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. 537, 113–137.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010.

New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
Han, Y., Wessler, S.R., 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 38 e199-e199.
Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., Stanke, M., 2015. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32, 767–769.
Hubisz, M.J., Falush, D., Stephens, M., Pritchard, J.K., 2009. Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Resour. 9, 1322–1332.
Islam, W., Akutse, K.S., Qasim, M., Khan, K.A., Ghramh, H.A., Idrees, A., Latif, S., 2018. *Bemisia tabaci*-mediated facilitation in diversity of begomoviruses: evidence from recent molecular studies. Microb. Pathog. 123, 162–168.
Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., 2014. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24, 1384–1395.
Kelley, D.R., Schatz, M.C., Salzberg, S.L., 2010. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 11 R116.
Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360.
Korf, I., 2004. Gene finding in novel genomes. BMC Bioinf. 5, 59.
Legg, J., French, R., Rogan, D., Okao-Okuja, G., Brown, J.K.J.M.E., 2002. 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.
Legg, J., Jeremiah, S., Obiero, H., Maruthi, M., Ndyetabula, I., Okao-Okuja, G., Bouwmeester, H., Bigirimana, S., Tata-Hangy, W., Gashaka, G., 2011. Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa. Virus Res. 159, 161–170.
Legg, J., Ogwal, S., 1998. Changes in the incidence of African cassava mosaic virus disease and the abundance of its whitefly vector along south–north transects in Uganda. J. Appl. Entomol. 122, 169–178.
Legg, J., Owor, B., Sseruwagi, P., Ndunguru, J., 2006. Cassava mosaic virus disease in East and Central Africa: epidemiology and management of a regional pandemic. Adv. Virus Res. 67, 355–418.
Legg, J., Shirima, R., Tajebe, L.S., Guastella, D., Boniface, S., Jeremiah, S., Nsami, E., Chikoti, P., Rapisarda, C., 2014. Biology and management of Bemisia whitefly vectors of cassava virus pandemics in Africa. Pest Manag. Sci. 70, 1446–1453.
Li, L., Stoeckert Jr., C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189.
Liu, S.-s., Colvin, J., De Barro, P.J., 2012. Species concepts as applied to the whitefly *Bemisia tabaci* systematics: how many species are there? J. Integ. Agric. 11, 176–186.
Maruthi, M., Hillocks, R., Mtunda, K., Raya, M., Muhanna, M., Kiozia, H., Rekha, A., Colvin, J., Thresh, J., 2005. Transmission of cassava brown streak virus by *Bemisia tabaci* (Gennadius). J. Phytopathol. 153, 307–312.
Maruthi, M.N., Colvin, J., Seal, S., 2001. Mating compatibility, life-history traits, and RAPD-PCR variation in *Bemisia tabaci* associated with the cassava mosaic disease pandemic in East Africa. Entomol. Exp. Appl. 99, 13–23.
Maruthi, M.N., Jeremiah, S.C., Mohammed, I.U., Legg, J.P., 2017. The role of the whitefly, *Bemisia tabaci* (Gennadius), and farmer practices in the spread of cassava brown streak ipomoviruses. J. Phytopathol. 165, 707–717.
Mbanzibwa, D.R., Tian, Y.P., Tugume, A.K., Mukasa, S.B., Tairo, F., Kyamanywa, S., Kullaya, A., Valkonen, J.P., 2011. Simultaneous virus-specific detection of the two cassava brown streak-associated viruses by RT-PCR reveals wide distribution in East Africa, mixed infections, and infections in Manihot glaziovii. J. Virol. Methods 171, 394–400.
Mitchell, A., Chang, H.Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.Y., Bateman, A., Punta, M., Attwood, T.K., Sigrist, C.J., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D.A., Wu, C.H., Orengo, C., Sillitoe, I., Mi, H., Thomas, P.D., Finn, R.D., 2015. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 43, D213–D221.
Otim-Nape, G., Thresh, J., Fargette, D., 1996. *Bemisia tabaci* and cassava mosaic virus disease in Africa. In: Gerling, D., Meyer, R.T. (Eds.), Bemisia: 1995, Taxonomy, Biology, Damage, Control and Management. Intercept Publishers, Andover, UK, pp. 319–350.
Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D., 2012. Ancient admixture in human history. Genetics 192, 1065–1093.
Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295.
Rey, C., Vanderschuren, H.V., 2017. Cassava mosaic and brown streak diseases: current perspectives and beyond. Annu. Rev. Virol. 4, 429–452.
Russello, M.A., Waterhouse, M.D., Etter, P.D., Johnson, E.A., 2015. From promise to practice: pairing non-invasive sampling with genomics in conservation. PeerJ 3, e1106.
Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.
Stanke, M., Waack, S., 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19, 215–225.
Su, N.-Y., Chouvenc, T., Li, H.-F., 2017. Potential hybridization between two invasive termite species, *Coptotermes formosanus* and *C. gestroi* (Isoptera: rhinotermitidae), and its biological and economic implications. Insects 8, 14.
Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963.
Wellenreuther, M., Munoz, J., Chavez-Rios, J.R., Hansson, B., Cordero-Rivera, A., Sanchez-Guillen, R.A., 2018. Molecular and ecological signatures of an expanding hybrid zone. Ecol. Evol. 8, 4793–4806.
Wosula, E.N., Chen, W., Fei, Z., Legg, J.P., 2017. Unravelling the genetic diversity among cassava *Bemisia tabaci* whiteflies using NextRAD sequencing. Genome Biol. Evol. 9, 2958–2973.
Xie, W., Chen, C., Yang, Z., Guo, L., Yang, X., Wang, D., Chen, M., Huang, J., Wen, Y., Zeng, Y., Liu, Y., Xia, J., Tian, L., Cui, H., Wu, Q., Wang, S., Xu, B., Li, X., Tan, X., Ghanim, M., Qiu, B., Pan, H., Chu, D., Delatte, H., Maruthi, M.N., Ge, F., Zhou, X., Wang, X., Wan, F., Du, Y., Luo, C., Yan, F., Preisser, E.L., Jiao, X., Coates, B.S., Zhao, J., Gao, Q., Xia, J., Yin, Y., Liu, Y., Brown, J.K., Zhou, X.J., Zhang, Y., 2017. Genome sequencing of the sweetpotato whitefly *Bemisia tabaci* MED/Q. GigaScience 6, 1–7.
Xue, J., Zhou, X., Zhang, C.-X., Yu, L.-L., Fan, H.-W., Wang, Z., Xu, H.-J., Xi, Y., Zhu, Z.-R., Zhou, W.-W., Pan, P.-L., Li, B.-L., Colbourne, J.K., Noda, H., Suetsugu, Y., Kobayashi, T., Zheng, Y., Liu, S., Zhang, R., Liu, Y., Luo, Y.-D., Fang, D.-M., Chen, Y., Zhan, D.-L., Lv, X.-D., Cai, Y., Wang, Z.-B., Huang, H.-J., Cheng, R.-L., Zhang, X.-C., Lou, Y.-H., Yu, B., Zhuo, J.-C., Ye, Y.-X., Zhang, W.-Q., Shen, Z.-C., Yang, H.-M., Wang, J., Wang, J., Bao, Y.-Y., Cheng, J.-A., 2014. Genomes of the rice pest brown planthopper and its endosymbionts reveal complex complementary contributions for host adaptation. Genome Biol. 15, 521.
Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.
Zhong, S., Joung, J.G., Zheng, Y., Chen, Y.R., Liu, B., Shao, Y., Xiang, J.Z., Fei, Z., Giovannoni, J.J., 2011. High-throughput illumina strand-specific RNA sequencing library preparation. Cold Spring Harb. Protoc. 2011, 940–949.