SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
MOLECULAR EVOLUTION
ENTOMOLOGY
GENETIC MARKERS

Received
20 February 2014

Accepted
22 August 2014

Published
15 September 2014

Correspondence and requests for materials should be addressed to X.-W.W. (xwwang@zju.edu.cn)

Developing conversed microsatellite markers and their implications in evolutionary analysis of the *Bemisia tabaci* complex

Hua-Ling Wang[1], Jiao Yang[1], Laura M. Boykin[2], Qiong-Yi Zhao[3], Yu-Jun Wang[1], Shu-Sheng Liu[1] & Xiao-Wei Wang[1]

[1]Ministry of Agriculture Key Laboratory of Agricultural Entomology, Institute of Insect Sciences, Zhejiang University, Hangzhou 310058, China, [2]ARC Centre of Excellence, Plant Energy Biology, The University of Western Australia, 35 Stirling Highway, Crawley 6009 Western Australia, [3]Queensland Brain Institute, University of Queensland, Brisbane, Qld 4072, Australia.

The study of population genetics among the *Bemisia tabaci* complex is limited due to the lack of conserved molecular markers. In this study, 358, 433 and 322 new polynucleotide microsatellites are separately identified from the transcriptome sequences of three cryptic species of the *B. tabaci* complex. The cross species transferability of 57 microsatellites was then experimentally validated. The results indicate that these markers are conserved and have high inter-taxon transferability. Thirteen markers were employed to assess the genetic relationships among six cryptic species of the *B. tabaci* complex. To our surprise, the inferred phylogeny was consistent with that of mitochondrial COI sequences, indicating that microsatellites have the potential to distinguish species of the *B. tabaci* complex. Our results demonstrate that development of microsatellites from transcriptome data is a fast and cost-effective approach. These markers can be used to analyze the population genetics and evolutionary patterns of the *B. tabaci* complex.

The whitefly *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidea) is a species complex containing at least 34 cryptic species[1–4]. The species complex colonizes more than 600 different species of plants and causes significant damage through transmitting plant viruses and feeding on plant phloem sap[3]. These cryptic species are morphologically indistinguishable[5] and the mitochondrial cytochrome oxidase I (mtCOI) marker has been widely used to delimit different members of the complex[2,6]. To date, at least 12 distinct genetic groups have been identified from the complex based on mtCOI sequences[2,6] and all available mating studies are in favor of the species-level boundaries[3]. The 12 distinct groups relates to the break in divergence frequencies identified at around 12%. However, it is perhaps more important to consider that there are 4 major clusters that represent the complex: (1) SubSaharan Africa (the ancestral cluster); 2) Asia; 3) New World and 4) North Africa/Middle East/Asia Minor.

During the last twenty years, the Middle East Asia Minor 1 (MEAM1) and Mediterranean (MED) cryptic species of the complex have invaded many countries around the world and the invasion of MEAM1 and MED are associated with the displacement of closely related members of the complex[7,8]. Numerous efforts have been made to reveal the possible factors responsible for the invasion of MEAM1 and MED whiteflies. However, because the species of the *B. tabaci* complex are morphologically indistinguishable, the evolution of the complex and the migration and displacement process of MEAM1 and MED invasion are hard to trace.

Previously, various genetic markers have been used to study the genetic diversity/structures of different cryptic species of the *B. tabaci* complex such as the random amplification of polymorphic DNA (RAPD) PCR[9], amplified fragment length polymorphisms (AFLP)[10], restriction fragment length polymorphism (RFLP)[11], mitochondrial DNA[6], ribosomal ITS1[12] and microsatellite markers[13–16]. Among these genetic markers, microsatellites, or simple sequence repeats (SSRs), are randomly repeated motifs of DNA composed of 1–6 base pair (bp) long units[17], which can be highly polymorphic among populations and are valuable for linkage mapping, comparative genomics and gene-based association studies[18]. In addition, microsatellites are also indispensable tools that can be used to reconstruct invasion histories and colonization routes and to reveal population bottlenecks and regional dispersal patterns[19]. Owing to these advantages, microsatellite has become increasingly popular for analyses of population genetics and evolutionary mechanisms of pest invasions[20].

To date, 54 microsatellite markers are available for *B. tabaci*[14,21–23]. However, all of these microsatellites were derived from the genomic DNA and the connections between these markers and gene functions are completely unknown. Expressed Sequence Tag (EST) and transcriptome sequences contain polymorphic genetic markers and can be used to identify microsatellites[24]. Compared to the genomic DNA-derived microsatellites, EST- and transcriptome-derived microsatellites lack introns and intragenic regions and can correspond to genes with known or predicted functions[25]. In addition, those microsatellites have fewer null alleles and stutter bands[26] and have more potential statistical power in multiple comparisons[27]. Furthermore, EST- and transcriptome-derived microsatellites have high degree of transferability across species[28], and can be used in closely relative species[29]. Therefore, systematical investigation of EST- and transcriptome-derived microsatellites will facilitate evolutionary and comparative studies in the *B. tabaci* complex composed of closely related cryptic species[27].

Recently, the transcriptomes of two invasive (MEAM1, MED) and one indigenous whitefly species (Asia II 3) have been sequenced[30,31]. These studies have generated a tremendous amount of data and provided a valuable source for the identification of microsatellite markers in whiteflies. The first objective of this study is to identify microsatellites from the three transcriptomes. In addition, microsatellites located in different regions of a gene serve various functions[32]. The distribution of microsatellites on genes was also analyzed. Furthermore, PCR experiments were employed to verify these predicted microsatellites and their cross species transferability. By comparative analysis of the newly developed microsatellites, the genetic relationships of six *B. tabaci* species were revealed. This study provides a rich resource of microsatellites for the *B. tabaci* complex and will facilitate researches on whitefly genetic diversity and evolution.

Results

Identification of microsatellites from the *B. tabaci* transcriptome databases. A total of 27.653 Mbp, 44.937 Mbp and 24.468 Mbp of sequences from the MEAM1, MED and Asia II 3 transcriptomes were used for mining microsatellites with the MISA-Micro Satellite program[33] (Table 1). There were 6419, 11711 and 4115 microsatellites in MEAM1, MED and Asia II 3 respectively (Table S1), which correspond to one microsatellite per 3.837 ~ 5.946 Kbp of transcriptome sequences. The total numbers of polynucleotide repeats were 358, 433 and 322 in MEAM1, MED and Asia II 3 respectively. While most microsatellites-containing unigenes have only one microsatellite, there are 362, 299 and 190 unigenes containing multiple microsatellites. In addition, 277, 367 and 193 unigenes contain compound microsatellites in MEAM1, MED and Asia II 3, respectively (Table 1).

Of the characterized microsatellites, mononucleotide repeats were the most common, followed by dinucleotide, trinucleotide and tetranucleotide repeats (Fig. 1A). On a complementary strand, a polyA repeat is the same as a polyT repeat. Similarly, in different reading frames or on a complementary strand, (AC)n is the same as (CA)n, (TG)n and (GT)n, while (AAG)n is the same as (AGA)n, (GAA)n, (CTT)n, (TTC)n, and (TCT)n. Thus, mononucleotide, dinucleotide and trinucleotide repeats can be grouped into 2, 4 and 10 unique classes respectively[34]. In the three species of the *B. tabaci* complex, A/T motifs were the most abundant in mononucleotide repeats, and AG class were the most common in dinucleotide repeats (Fig. 1B). However, the usage of trinucleotide was different. In MEAM1, the AAG class was the most widespread followed by ATG and AAC class (Fig. 1C). In MED, the most prevalent three triplet codons were AAG class, whereas in Asia II 3, ATG is the most frequent class (Fig. 1C). A total of 45 tetra microsatellites were identified from both MEAM1 and Asia II 3. Furthermore, two pentra motifs were found in Asia II 3 and one hexa motif was found in MEAM1 and Asia II 3, respectively (Fig. 1A).

Distribution of microsatellites in 3′UTR, 5′UTR and CDS regions. The distribution of polynucleotide microsatellites in CDS, 5′UTR and 3′UTR regions was investigated. Based on the information of BLASTx homology, the position of 71, 71 and 40 polynucleotide microsatellites were respectively determined for MEAM1, MED and Asia II 3 (Table 2 & Table S2). In the CDS, 45, 48 and 27 microsatellites with polynucleotide repeats were found in MEAM1, MED and Asia II 3, respectively, which were significantly higher than that of UTRs (Table 2). Interestingly, the number of trinucleotide repeats in CDS region was also much higher than other types of microsatellites. The characteristics of the amino acids encoded by the trinucleotide repeats in CDS were then investigated. In MEAM1, MED and Asia II 3, a total of 37, 37 and 20 triplet codons were found and they encoded 10, 15 and 8 different amino acids respectively (Fig. 2A). The codons encoded aromatic amino acids took up most of the partition, followed by aliphatic and heterocyclic amino acids (Fig. 2B). The codons encoded hydrophilic amino acids were 29, 20 and 17, while encoded hydrophobic amino acids were 3, 15 and 3 in MEAM1, MED and Asia II 3, respectively.

Gene Ontology (GO) and KEGG annotation of microsatellite-containing sequences. GO assignments were used to classify the functions of the genes with microsatellites. Based on sequence homology, 272, 456 and 255 microsatellite-containing sequences from MEAM1, MED and Asia II 3 have GO annotations and can be categorized into 35 functional groups. The 35 functional groups were classified into three main categories (Biological process, Cellular component and Molecular function) (Fig. 3). GO analysis showed that 'Metabolic process', 'Cellular process' and 'Cell part' terms are dominant. Next, these genes were annotated to different KEGG pathways. A total of 318, 538, 297 microsatellite-containing sequences of MEAM1, MED and Asia II 3 were mapped to 193, 222 and 210 KEGG pathways, respectively (Table S3). Some of these genes are related to resistance to environmental stresses and insecticides, such as aldehyde oxidase[35], cytochrome P450[36] and mitogen-stress activated protein kinases 2[37] (Table S4).

Characterization of the predicted microsatellite markers. To validate the predicted microsatellites, 88 primer pairs were synthesized (24 for MEAM1, 32 for MED and 32 for Asia II 3) to amplify microsatellites from whitefly DNA. Among the 88 primer pairs, 57 (65%) generated clear PCR products and 31 (35%) did not

Table 1 | Frequency and distribution of microsatellites in three species of the *B. tabaci* complex

Descriptions	MEAM1	MED	Asia II 3
Total number of sequences examined:	57,741	168,900	52,535
Total size of examined sequences (bp):	27,653,107	44,936,957	24,468,191
Total number of identified microsatellites:	6,419	11,711	4,115
Number of microsatellites containing sequences:	6,057	11,512	3,925
Number of polynucleotide microsatellites	358	433	322
Number of sequences with >1 microsatellites:	362	299	190
Number of compound microsatellites:	277	367	193

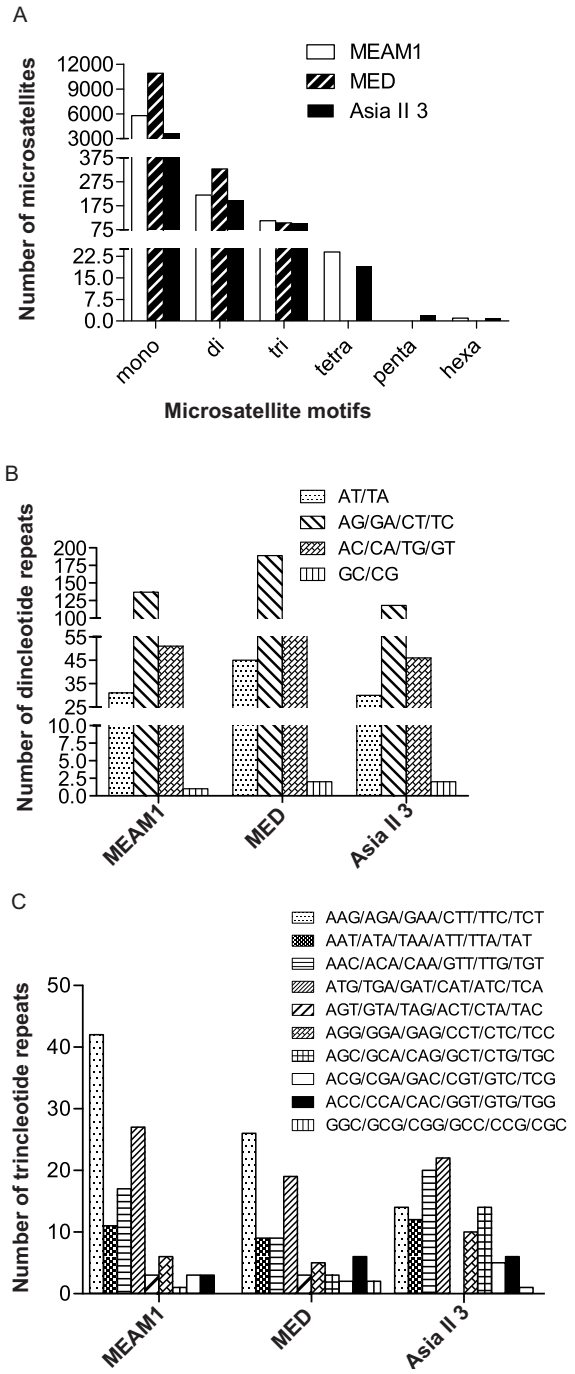

Figure 1 | Distribution of microsatellites in the three whitefly species. (A) Distribution of repeat loci. (B) Distribution of different dinucleotide repeats. (C) Distribution of different trinucleotide repeats.

yield good amplification. The frequencies of microsatellite amplification in MEAM1, MED and Asia II 3 were 70.83% (24 designed, 17 effective), 53.13% (32 designed, 17 effective) and 71.88% (32 designed, 23 effective), respectively. The failure to amplify microsatellites may be caused by non-specific primers, the presence of large introns in the genomic DNA or inappropriate PCR conditions. Among the 57 markers that generated clear PCR products, 42 markers showed polymorphisms; whereas 15 generated a single PCR product (monomorphism). Details (repeat motifs, PCR primers, allele sizes, gene ID, accession number and possible change of amino acids) of the 57 markers are shown in Table S5.

Table 2 | The numbers of polynucleotide microsatellites in 3′UTRs, 5′UTRs and CDS

Species/regions	P2	P3	P4	Total
MEAM1				
5′UTRs	6	1	0	7
3′UTRs	13	6	0	19
CDS	8	37	0	45
MED				
5′UTRs	6	2	0	8
3′UTRs	13	2	0	15
CDS	11	37	0	48
Asia II 3				
5′UTRs	3	1	0	4
3′UTRs	5	4	0	9
CDS	6	20	1	27

P2, P3 represent the types of microsatellites.
P2: dinucleotide; P3: trinucleotide.

The cross-taxa transferability of microsatellites. Next, the cross-amplification of the 57 primer pairs were investigated by PCR and capillary electrophoresis[38]. The result showed that 42 (73.68%) amplified fragments from all of the three whitefly species (Table S6), suggesting that these microsatellites are highly conserved and may act as genetic markers for the *B. tabaci* complex. In addition, 8 primer pairs could amplify fragments from the two invasive whiteflies and 1 pair amplified from both MED and Asia II 3, while the remaining 6 pairs of primers only amplified fragments from one of the three species. Next, the functions of these 42 microsatellite-containing genes were determined through Blast search, of which 33 showed significant similarities to known genes (Table S5). The alleles of these 42 microsatellites were compared among MEAM1, MED and Asia II 3 whiteflies. Of the 42 microsatellites, 26 contained the same alleles in MEAM1 and MED, whereas only 17 markers shared in MEAM1 and Asia II 3 and 16 markers shared in MED and Asia II 3 (Table S6), which is consistent with the fact that MEAM1 and MED have closer relationships among the three cryptic species. Interestingly, the sodium channel gene (Gene ID 99267), which is associated with xenobiotic resistance[39], was found to contain different alleles in invasive and native whitefly species (Table S6).

Characteristics of the 13 microsatellites in six species of the *B. tabaci* complex. Thirteen polymorphic microsatellites (Table S5 marked in yellow) were then employed to assess the polymorphism and heterozygosity among six laboratory colonies of the *B. tabaci* complex (8 individuals for each species) (Table 3). A total of 93 alleles were identified from laboratory colonies of the six species based on the 13 microsatellite markers. When compared the performance of 13 microsatellite markers independently among the 6 cryptic species, the number of alleles (N_A) ranged from 3 to 15, with an average of 7.2 alleles per locus. The observed (*Ho*) and expected (H_E) heterozygosities ranged from 0.022 to 0.581 and 0.297 to 0.888 respectively. When compared the performance of these loci among the 6 cryptic species, the China 1 cryptic species exhibited the highest N_A (3.077), while MEAM1 and Asia II 7 displayed the lowest N_A (2.231) (Table 3). The MEAM1 showed the lowest observed (*Ho*) and expected (H_E) heterozygosities. There are null alleles for two loci (291416, 27966) in MEAM1, two loci (31541, 22561) in MED, one locus (25869) in Asia II 3, three loci (102573, 29116, 99267) in Asia II 7 and two loci (36306, 102573) in Asia II 6. The Hardy-Weinberg equilibrium test was done for each locus in each population, 25 of 50 groups were significantly deviated from Hardy-Weinberg equilibrium (Table 3). Significant genotypic linkage disequilibrium was not detected. With regard to the polymorphism (PIC) of the 13 microsatellite-containing genes, the

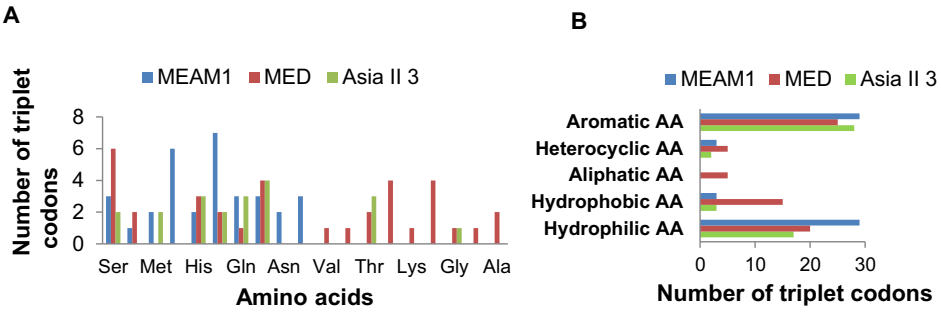

Figure 2 | The characteristics of trinucleotide repeats in CDS region of three species. (A) Distribution of different amino acid codons in the three species. (B) Distribution of amino acid codons which encoded different types of amino acids in the three species.

glycine-rich cell wall structural protein precursor (Gene ID: 29116), sodium channel (Gene ID: 99267) and NADH dehydrogenase subunit 4L (Gene ID: 76476) had high polymorphism (Table 3). The CCAAT/enhancer binding protein (Gene ID: 102573) showed relatively low polymorphism (Table 3). The polymorphism information content (PIC) ranged from 0.246 to 0.866, with an average of 0.612 in the six cryptic species, which indicates the effectiveness of microsatellites markers for detecting polymorphism[40]. In addition, the genetic diversity of the 13 microsatellite loci was compared among the six whitefly species. The level of gene diversity from the highest to the lowest was Asia II 3, MED, China 1, Asia II 6, MEAM1 and Asia II 7 (Table S7). Fig. 4 displays the gene diversity of the 13 microsatellite loci in every comparison. For example, in the comparison of MED and MEAM1, 7 of the 13 loci in MEAM1 showed a decrease of gene diversity compared to that of MED. To further testify these markers, 4 microsatellites were randomly picked to sequence and analyze the alleles per locus among the six cryptic species (Fig. 5). Complex mutational patterns (single base mutation, change of repeat units and indels within flanking region) were observed in these transcriptome-derived microsatellites (Fig. 5), which is normal in insects[41].

The implications of microsatellites in evolutionary analysis. To date, at least 12 distinct genetic groups have been identified from the complex based on mtCOI sequences which consist of 4 major clusters that represent the complex: 1) SubSaharan Africa (the ancestral cluster); 2) Asia; 3) New World and 4) North Africa/Middle East/Asia Minor. The 6 species covered belong to 2 of the 4 major clusters, i.e. Asia and North Africa/Middle East/Asia Minor. The Neighbor Joining method was used for cluster analysis of the six species (Figure 6 A & B). Interestingly, the genetic relationship based on the 13 microsatellite loci is in agreement with the phylogeny of the partial mtCOI sequences[2]. Principal coordinates analysis (PCA) were done using GENALEX 6 software[42] and the results revealed that the China 1, Asia II 3, Asia II 6 and Asia II 7 clustered on the left quadrant of the plot (Figure 6 B & C). When considered the first and third factors, the results revealed clearly that there are three genetic groups among the 6 cryptic species. The PCA analysis supports the phylogenetic clusters. In conclusion, these microsatellites may be used as markers to describe the genetic diversity of the *B. tabaci* complex.

Discussion

To our knowledge, this is the first investigation of the frequency and distribution of *B. tabaci* microsatellites using transcriptome data (Table 1). We found that mononucleotide repeats were the most common, followed by dinucleotide repeats, trinucleotide and tetra-

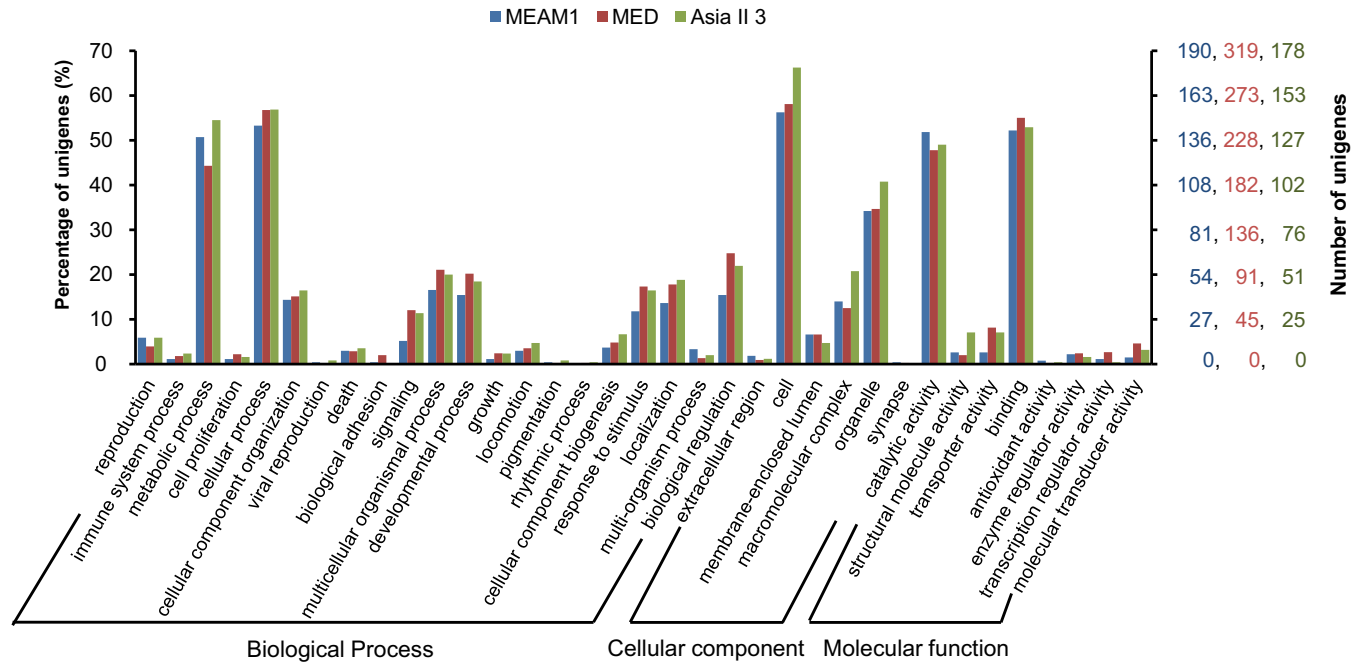

Figure 3 | GO classifications of microsatellite-containing sequences from the three species. The right y-axis represents the number of microsatellite-containing genes in a category. The left y-axis represents the percentage of a specific category of microsatellite-containing genes in that main category.

Table 3 | The characteristics of the newly developed microsatellite markers for six cryptic species of the *B. tabaci* complex

Locus no.		MEAM1	MED	Asia II 3	China 1	Asia II 7	Asia II 6	Overall
19488	N_A	1	2	3	1	1	1	4
	H_E	0.000	0.429	0.692	0.000	0.000	0.000	0.536
	H_O	0.000	0.500	0.714	0.000	0.000	0.000	0.219
	F_{IS}		−0.200	−0.035				−0.077
	PIC							0.491
98023	N_A	1	1	3	1	1	2	5
	H_E	0.000	0.000	0.567	0.000	0.000	0.125	0.715
	H_O	0.000	0.000	0.250	0.000	0.000	0.125	0.067
	F_{IS}			0.576*			0.000	0.475
	PIC							0.651
36306	N_A	2	2	3	3	3	3	3
	H_E	0.356	0.200	0.714	0.464	0.750	0.714	0.522
	H_O	0.000	0.200	0.500	0.500	0.500	0.000	0.269
	F_{IS}	1.000**	0.000	0.333	−0.091	0.368**	1.000**	0.512
	PIC							0.458
102573	N_A	1	5	3	3	1	3	7
	H_E	0.000	0.593	0.242	0.385	0.000	0.433	0.297
	H_O	0.000	0.429	0.250	0.143	0.000	0.000	0.140
	F_{IS}		0.294	−0.037	0.647**		1.000***	0.529
	PIC							0.246
31541	N_A	5	5	3	3	2	4	11
	H_E	0.857	0.733	0.275	0.242	0.125	0.455	0.612
	H_O	0.500	0.500	0.286	0.250	0.125	0.500	0.342
	F_{IS}	0.455	0.333***	−0.044	−0.037	0.000	−0.111	0.187
	PIC							0.586
25869	N_A	2	1	5	4	2	3	8
	H_E	0.458	0.000	0.889	0.867	1.000	0.833	0.797
	H_O	0.625	0.000	0.000	0.333	1.000	1.000	0.333
	F_{IS}	−0.400		1.000***	0.667	−1.000	−0.333	0.388
	PIC							0.633
76476	N_A	1	3	4	3	3	1	8
	H_E	0.000	0.567	0.692	0.508	0.604	0.000	0.797
	H_O	0.000	0.250	0.750	0.625	0.857	0.000	0.413
	F_{IS}		0.576***	−0.091	−0.250	−0.469		−0.038
	PIC							0.761
30306	N_A	2	2	1	4	3	3	6
	H_E	0.500	0.485	0.000	0.582	0.342	0.633	0.778
	H_O	0.000	0.000	0.000	0.286	0.125	0.000	0.068
	F_{IS}	1.000**	1.000**		0.529**	0.650**	1.000***	0.850
	PIC							0.731
29116	N_A	5	4	4	7	4	4	10
	H_E	0.708	0.800	0.642	0.901	0.792	0.675	0.855
	H_O	0.375	0.500	0.500	0.429	0.375	1.000	0.532
	F_{IS}	0.488**	0.391***	0.233**	0.544**	0.544***	−0.534	0.306
	PIC							0.827
27966	N_A	2	2	1	2	2	1	6
	H_E	0.458	0.400	0.000	0.233	0.000	0.000	0.667
	H_O	0.125	0.250	0.000	0.000	0.000	0.000	0.065
	F_{IS}	0.741*	0.391		1.000***			0.398
	PIC							0.688
99267	N_A	4	4	4	5	5	6	15
	H_E	0.561	0.659	0.747	0.717	0.533	0.846	0.888
	H_O	0.667	0.286	0.857	0.375	0.625	0.714	0.581
	F_{IS}	−0.212	0.586	−0.161	0.494	−0.186	0.167***	0.151
	PIC							0.866
35350	N_A	1	5	2	3	1	1	6
	H_E	0.000	0.783	0.233	0.567	0.000	0.000	0.567
	H_O	0.000	0.750	0.000	0.000	0.000	0.000	0.125
	F_{IS}		0.046	1.000***	1.000***			0.544
	PIC							0.522
22561	N_A	2	2	2	1	1	1	5
	H_E	0.125	0.233	0.264	0.000	0.000	0.000	0.544
	H_O	0.125	0.000	0.000	0.000	0.000	0.000	0.022
	F_{IS}	0.000	1.000***	1.000***				0.800
	PIC							0.495
Mean N_A		2.231	2.923	2.293	3.077	2.231	2.539	
Mean H_E		0.310	0.453	0.458	0.420	0.359	0.363	
Mean H_O		0.186	0.282	0.316	0.226	0.326	0.257	

Table 3 | Continued

Locus no.	MEAM1	MED	Asia II 3	China 1	Asia II 7	Asia II 6	Overall
F_{IS}	0.423	0.395	0.333	0.489	0.132	0.306	

N_A: Number of alleles detected; H_E: Expected heterozygosity; H_O: Observed heterozygosity; F_{IS}: Inbreeding index; Hardy–Weinberg tests are indicated together with F_{IS} values.
*$P < 0.05$;
**$P < 0.01$;
***$P < 0.001$. Non-significant P values are not indicated. *PIC*: Polymorphic index content.

nucleotide repeats in all of the three species. These results confirm the theory that the microsatellite abundance decreases with the increase of motif repeat number and repeat length[43]. These microsatellites provide a valuable resource for the development of genetic markers in *B. tabaci*. In addition, these markers offer a chance to classify the functions of these microsatellite-containing genes. Interestingly, some microsatellite-containing genes were found in pathways related to environmental stress responses (Table S4). These markers may open a new avenue for the research on the *B. tabaci* pesticide resistance by estimating the frequencies of alleles in genes related to resistance across populations[44].

Many studies have demonstrated that microsatellites derived from transcribed sequences harbor higher transferability because of their conservative features[45,46]. In this study, among the 57 microsatellite markers, 42 primer pairs (73.68%) could amplify fragments from the three species. These results illustrate that the transcriptome-derived markers have high inter taxon transferability[47]. Therefore, these markers can be used to amplify microsatellites from the other closely related species that do not yet have markers and can be considered as anchor markers for comparative mapping and evolutionary studies across species[48].

For 13 microsatellites among the colonies of six species (MEAM1, MED, Asia II 3, Asia II 6, China 1 and Asia II 7) of the *B. tabaci* complex, the number of alleles (N_A) ranged from 3 to 15, with an average of 7.2 alleles per locus. Previously, Valle, Lourenção, Zucchi and Pinheiro[23], Tsagkarakou and Roditakis[22] and De Barro, Scott, Graham, Lange and Schutze[21] found that the number of alleles per locus ranged from 1 to 2, 2 to 13 and 6 to 44, respectively. The difference in polymorphisms may be attributed to different population sizes or cryptic species used. Polymorphism and heterozygos-

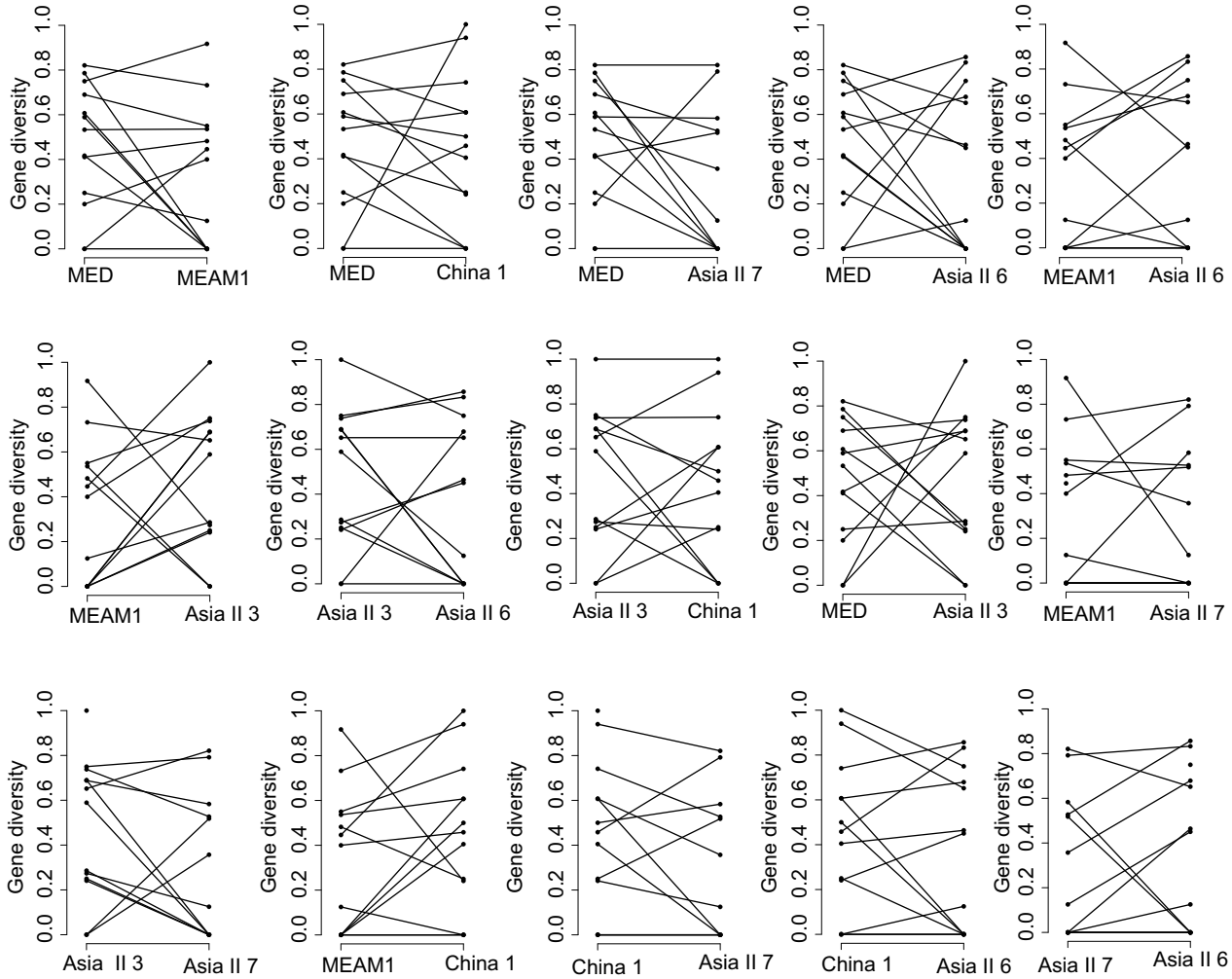

Figure 4 | Comparison of gene diversity among the six cryptic species of the *B. tabaci* complex. Each point represents the gene diversity of one locus and the lines connect points from different species.

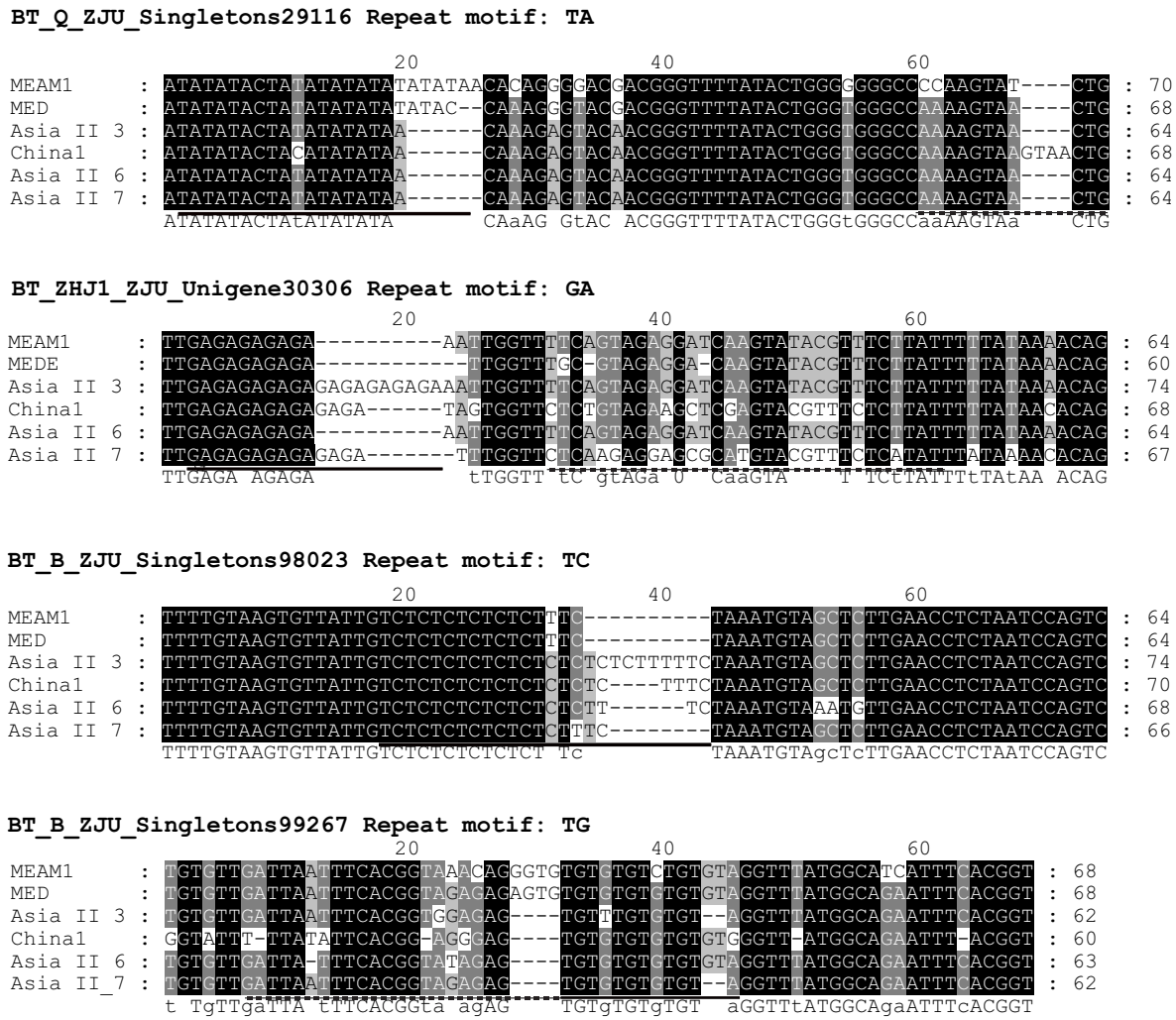

Figure 5 | **Mutational patterns of selected microsatellites in *B. tabaci*.** The black lines indicate the repeat motifs and the black-dotted lines represent the indels in the flanking region.

ity are critical characteristics of microsatellite markers. The high allele numbers and heterozygosity (Table 3) of these transcriptome-derived markers suggests that they can be used to assess the genetic profile of the *B. tabaci* complex. A total of 25 loci significantly deviated from Hardy-Weinberg equilibrium (Table 3). The source of deficiencies includes the recurrent inbreeding, subpopulation structure (Wahlund effect) and/or null alleles. The big differences between H_o and H_E also were found among the genomics-based microsatellites developed for *B. tabaci* by De Barro[49]. In addition, the marked differences between H_o and H_E also were observed in the genomics-based microsatellites of *B. tabaci* collected in different environment[14,16]. Some loci did not have heterozygosity and the F_{IS} values were all positive, suggesting that the loss of heterozygosity can happen in some microsatellites. This is probably due to the fact that our samples were obtained from lab populations. Additional experiments with field populations are warranted to reveal the polymorphism and heterozygosity of different species of the *B. tabaci* complex.

Due to the lack of morphological characteristics, the systematics of *B. tabaci* species complex depends largely on molecular approaches[2–4,6]. To date, the evolutionary history of *B. tabaci* complex was inferred exclusively from the partial sequence of COI gene[2,6,50]. Other molecular markers are in great need to further illustrate the complicated evolutionary relationships of the *B. tabaci* species complex. For the 4 major clusters that represent the *B. tabaci* species complex, the most divergent cluster is SubSaharan Africa followed by New World. We have used 11 newly developed microsatellites to study the SubSaharan Africa1 and the results revealed that they can work on the cryptic species (data unpublished). Owing to the absence of the specimen, we did not study the New World, however, the most divergent cluster is SubSaharan Africa followed by New World among the 4 major clusters. Therefore, our newly developed microsatellites are conserved and many of them can be used to analyze other cryptic species of the *B. tabaci* complex. The genetic relationship of 6 species derived by 13 novel conserved microsatellites markers is in accordance with the phylogeny of the partial COI sequences. It provides additional evidence for the evolutionary relationship of the 6 cryptic species. In addition, our results also prove that microsatellites can be considered as suitable markers for future evolutionary analysis in other species complex.

Conclusion

In this study, we characterize a large number of transcriptome-derived microsatellites from members of the *B. tabaci* complex. Functional categorization of these microsatellites-containing genes provides valuable information about the potential functions of these microsatellites. In addition, 57 microsatellite markers were experimentally validated. Many of these microsatellites can be used for population analysis and are transferable across different species of the *B. tabaci* complex. Complex mutational patterns were observed in these transcriptome-derived microsatellites. What's more, by ana-

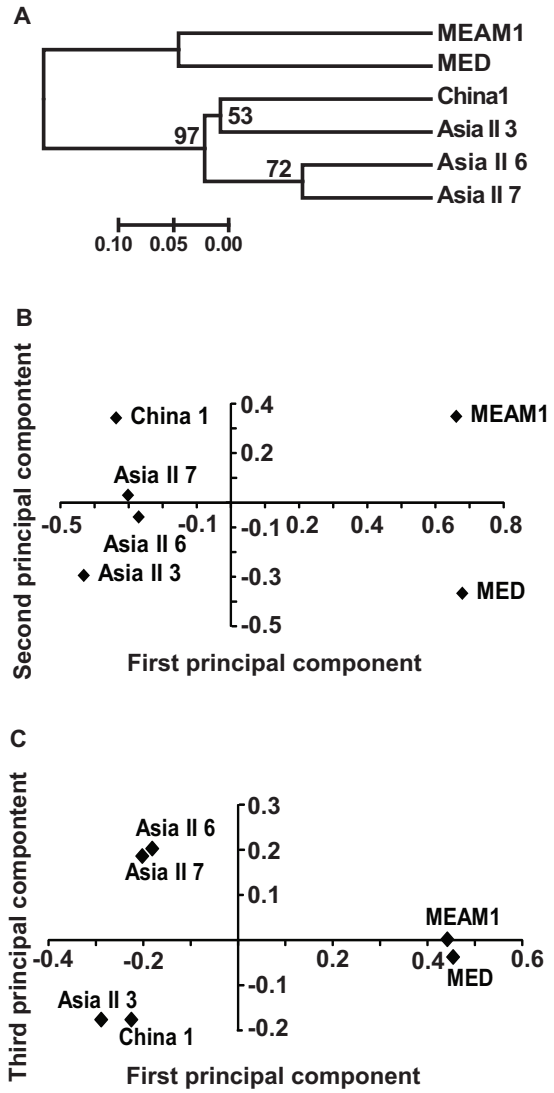

Figure 6 | Phylogenetic and Principle components analysis of the six cryptic species. (A) Phylogeny of the six cryptic species. POPTREE2 software[59] were used to construct the phylogenetic tree and the numbers on the node represent bootstrap support. The scale bar represent Nei's standard genetic distance D[60]. (B) Two-dimensional scatter plot of the first and second factors for 6 cryptic species. The first and second PCs account for 60.21% and 20.52% of the variation, respectively. (C) Two-dimensional scatter plot of the first and third factors for 6 cryptic species. The first and third PCs summarize 60.21% and 13.75% of the variation, respectively.

lyzing the genetic relationships of six *B. tabaci* species basing on newly developed microsatellites, the results indicated that these microsatellites may be used as markers to describe the genetic diversity of the *B. tabaci* complex. These markers enrich the existing microsatellite markers of *B. tabaci* and can be used to analyze the genetic diversity and evolutionary pattern of this whitefly complex.

Methods

Mining the microsatellites from the transcriptome database. The MISA-Micro Satellite identification tool (http://pgrc.ipk-gatersleben.de/misa/) was used to search for microsatellites in the transcriptome data of MEAM1, MED and Asia II 3 whiteflies. Microsatellites were defined as being mononucleotide repeats >= 10 repeats and di-, tri-, tetra-, penta- and hexanucleotide repeats >= 6 repeats[33]. Criteria for compound microsatellites was an interval of bases <= 100 of the motif length.

The location of microsatellites. Coding sequences (CDS) of each gene were first determined by BLASTx against the Swissprot database using a threshold of 1×10^{-5}. CDS with unexpected stop codon in the Blast hit region were removed. Start codon positions were determined by examination of the in-frame ATG codon present 30 bp upstream or downstream of the beginning of the aligned reference protein. The stop codon positions were determined by examination of in-frame TAA, TAG and TGA motifs present within 30 bp of the stop codon of the reference protein. The 5′ or 3′UTR regions were defined based on the CDS prediction. The locations of microsatellites were determined based on the predicted 5′ UTR, 3′UTR and CDS regions.

GO and KEGG pathway analysis. To understand their functions, all microsatellite-containing genes were searched against the GenBank nr protein database using BLASTx with an E-value cut-off of 10^{-5}. Blast2GO software was used to assign the Gene Ontology (GO) terms to these microsatellite-containing genes[51]. Blastall software was employed to perform the pathway analysis by searching all genes against the KEGG database.

Sample collection and DNA extraction. The invasive MEAM1 (mtCOI GeneBank accession number: GQ332577), MED (KF452516), the indigenous Asia II 3 (KF452527), Asia II 6 (KC540758), China 1 (KF452525) and Asia II 7 (EU192043) species of the *B. tabaci* complex were collected from Zhejiang, China. These cryptic species were maintained separately on cotton (cultivar Zhe-Mian 1793) with the following controlled conditions: 27 ± 1°C, a photoperiod of 14 h light: 10 h darkness and relative humidity of 70 ± 10%[7]. Total DNA was extracted from individual female adult whiteflies following the method of Frohlich *et al*[52]. The purity of the populations was identified by PCR amplification of a 0.7 kb fragment of mtCOI gene. The following forward and reverse primers were used to amplify the partial mtCOI sequences (5′-TGRTTYTTTGGTCATCCVGAAGT-3′ and, 5′-TTACTGCACTTTCTGCCACATTAG-3′).

Validation of microsatellite markers using PCR. To test these markers, Primer Premier 5.0[53] was used to design PCR primers from the sequences flanking the microsatellites. *Bemisia tabaci* produces males from unfertilized eggs and females from fertilized eggs. Therefore, only females were used to test polymorphism at each locus. M13-specific primers (5′-CACGACGTTGTAAAACGAC-3′) with a fluorescent dye (FAM or HEX, Applied Biosystems) were added to the 5′-end of each forward primer[38]. PCR was carried out in an S1000 thermal cycler (Bio-Rad). A 15 μL PCR reaction contained 0.25 μL 100 pmol/μl forward primer, 0.25 μL 100 pmol/μL reverse primer, 0.25 μL 100 pmol/μL 5′dye-labelled M13 primer, and 1.6 μL 10 × Ex Taq Buffer, 1.2 μL 2.5 mM dNTP and 0.1 μL Ex Taq polymerase (Takara, Japan). PCR cycling conditions were 94°C for 3 min, followed by 32 cycles of 96°C for 15 s, 51–63°C for 20 s, 72°C for 50 s. The PCR reaction products were diluted and detected on a MegaBACE 1000 DNA analysis system (Amersham Biosciences) at the Center of Analysis and Measurement in Zhejiang University. The ET550-R size standard (GE Healthcare) and Genetic Profiler version 2.2 (GE Healthcare) were used to judge the sizes of amplification.

Polymorphism and microsatellite distribution analysis. Software POPGENE (version 1.31)[54] was used to calculate the total number of polymorphic alleles (N), average number of alleles per locus (N_A), average number of effective alleles per locus (Ne), observed heterozygosity (Ho), expected heterozygosity (H_E), the genetic identity (I), genetic distance (D) and the gene diversity for each loci of different cryptic species. FSTAT (Version 1.2)[55] was used to examine the allelic richness (R). The inbreeding index (F_{IS}) and p (F_{IS}) was estimated by GENEPOP 4.0[56]. Polymorphism information content (PIC) was calculated by PIC-CALC version 0.6[57]. The null alleles and technical artifacts like stuttering and large allele dropout was assessed using MICRO-CHEKER v.2.2.3[58].

Data accessibility. DNA sequences: Sequences have been submitted to the GenBank with the accession number of: KF916587-KF916610. Detailed information of predicted microsatellites from the transcriptomes of the three species is shown in Supplementary Table S1. PCR primer sequences are presented in Supplementary Table S5.

1. 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 (2011).
2. Boykin, L. M., Armstrong, K. F., Kubatko, L. & De Barro, P. Species delimitation and global biosecurity. *Evol. Bioinform.* **8**, 1–37 (2012).
3. Liu, S. S., Colvin, J. & De Barro, P. J. Species concepts as applied to the whitefly *Bemisia tabaci* systematics: how many species are there? *Journal of Integrative Agriculture* **11**, 176–186 (2012).
4. Firdaus, S. *et al*. The *Bemisia tabaci* species complex: additions from different parts of the world. *Insect Sci.* **20**, 723–733 (2013).
5. Rosell, R. C. *et al*. Analysis of morphological variation in distinct populations of *Bemisia tabaci* (Homoptera: Aleyrodidae). *Ann. Entomol. Soc. Am.* **90**, 575–589 (1997).
6. Dinsdale, A., Cook, L., Riginos, C., Buckley, Y. & Barro, P. D. 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 (2010).
7. Liu, S. S. *et al*. Asymmetric mating interactions drive widespread invasion and displacement in a whitefly. *Science* **318**, 1769–1772 (2007).

8. Hu, J. *et al.* An extensive field survey combined with a phylogenetic analysis reveals rapid and widespread invasion of two alien whiteflies in China. *PLoS ONE* **6**, e16061 (2011).
9. De Barro, P. & Driver, F. Use of RAPD PCR to distinguish the B biotype from other biotypes of *Bemisia tabaci* (Gennadius)(Hemiptera: Aleyrodidae). *Australian Journal of Entomology* **36**, 149–152 (1997).
10. Cervera, M. *et al.* Genetic relationships among biotypes of *Bemisia tabaci* (Hemiptera: Aleyrodidae) based on AFLP analysis. *Bull. Entomol. Res.* **90**, 391–396 (2000).
11. Abdullahi, I., Atiri, G. I., Thottappilly, G. & Winter, S. Discrimination of cassava-associated *Bemisia tabaci* in Africa from polyphagous populations, by PCR-RFLP of the internal transcribed spacer regions of ribosomal DNA. *J. Appl. Entomol.* **128**, 81–87 (2004).
12. De Barro, P. J., Driver, F., Trueman, J. W. & Curran, J. Phylogenetic relationships of world populations of *Bemisia tabaci* (Gennadius) using ribosomal ITS1. *Mol. Phylogenet. Evol.* **16**, 29–36 (2000).
13. Chu, D., Gao, C., De Barro, P., Wan, F. & Zhang, Y. Investigation of the genetic diversity of an invasive whitefly (*Bemisia tabaci*) in China using both mitochondrial and nuclear DNA markers. *Bull. Entomol. Res.* **101**, 467–475 (2011).
14. Delatte, H. *et al.* Microsatellites reveal extensive geographical, ecological and genetic contacts between invasive and indigenous whitefly biotypes in an insular environment. *Genet. Res.* **87**, 109–124 (2006).
15. Dickey, A. M., Osborne, L. S., Shatters, R. G., Hall, P. M. & McKenzie, C. L. Population genetics of invasive *Bemisia tabaci* (Hemiptera: Aleyrodidae) cryptic species in the United States based on microsatellite markers. *J. Econ. Entomol.* **106**, 1355–1364 (2013).
16. De Barro, P. Genetic structure of the whitefly *Bemisia tabaci* in the Asia-Pacific region revealed using microsatellite markers. *Molecular Ecology* **14**, 3695–3718 (2005).
17. Tóth, G., Gáspári, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10**, 967–981 (2000).
18. Parchman, T. L., Geist, K. S., Grahnen, J. A., Benkman, C. W. & Buerkle, C. A. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* **11**, 180 (2010).
19. Kirk, H., Dorn, S. & Mazzi, D. Molecular genetics and genomics generate new insights into invertebrate pest invasions. *Evol. Appl.* **6**, 842–856 (2013).
20. Barrett, R. D. & Hoekstra, H. E. Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.* **12**, 767–780 (2011).
21. De Barro, P. J., Scott, K. D., Graham, G. C., Lange, C. L. & Schutze, M. K. Isolation and characterization of microsatellite loci in *Bemisia tabaci*. *Mol. Ecol. Notes* **3**, 40–43 (2003).
22. Tsagkarakou, A. & Roditakis, N. Isolation and characterization of microsatellite loci in *Bemisia tabaci* (Hemiptera: Aleyrodidae). *Mol. Ecol. Notes* **3**, 196–198 (2003).
23. Valle, G., Lourenção, A., Zucchi, M. & Pinheiro, J. Low polymorphism revealed in new microsatellite markers for *Bemisia tabaci* (Hemiptera: Aleyrodidae). *Genet. Mol. Res.* **11**, 3899–3903 (2012).
24. Jing, S. *et al.* Development and use of EST-SSR markers for assessing genetic diversity in the brown planthopper (*Nilaparvata lugens* Stål). *Bull. Entomol. Res.* **102**, 113 (2012).
25. Andersen, J. R. & Lübberstedt, T. Functional markers in plants. *Trends Plant Sci.* **8**, 554–560 (2003).
26. Woodhead, M. *et al.* Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. *Mol. Ecol.* **14**, 1681–1695 (2005).
27. Ellis, J. & Burke, J. EST-SSRs as a resource for population genetic analyses. *Heredity* **99**, 125–132 (2007).
28. Areshchenkova, T. & Ganal, M. Comparative analysis of polymorphism and chromosomal location of tomato microsatellite markers isolated from different sources. *Theor. Appl. Genet.* **104**, 229–235 (2002).
29. Scott, K. D. *et al.* Analysis of SSRs derived from grape ESTs. *Theor. Appl. Genet.* **100**, 723–726 (2000).
30. Wang, X. W. *et al.* Transcriptome analysis and comparison reveal divergence between two invasive whitefly cryptic species. *BMC Genomics* **12**, 458 (2011).
31. Wang, X. W. *et al.* Analysis of a native whitefly transcriptome and its sequence divergence with two invasive whitefly species. *BMC Genomics* **13**, 529 (2012).
32. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, 277–280 (2004).
33. Triwitayakorn, K. *et al.* Transcriptome sequencing of *Hevea brasiliensis* for development of microsatellite markers and construction of a genetic linkage map. *DNA Res.* **18**, 471–482 (2011).
34. Jurka, J. & Pethiyagoda, C. Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.* **40**, 120–126 (1995).
35. Pryde, D. C. *et al.* Aldehyde oxidase: an enzyme of emerging importance in drug discovery. *J. Med. Chem.* **53**, 8441–8460 (2010).
36. Puinean, A. M. *et al.* Amplification of a cytochrome P450 gene is associated with resistance to neonicotinoid insecticides in the aphid *Myzus persicae*. *PLoS Gene.t* **6**, e1000999 (2010).
37. Johnson, G. L. & Lapadat, R. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science* **298**, 1911–1912 (2002).
38. Oetting, W. *et al.* Linkage analysis with multiplexed short tandem repeat polymorphisms using infrared fluorescence and M13 tailed primers. *Genomics* **30**, 450–458 (1995).
39. Hemingway, J. & Ranson, H. Insecticide resistance in insect vectors of human disease. *Annu. Rev. Entomol.* **45**, 371–391 (2000).
40. Yamini, K. N. *et al.* Development of EST-SSR markers and their utility in revealing cryptic diversity in safflower (*Carthamus tinctorius L.*). *J. Plant Biochem. Biotechnol.* **22**, 90–102 (2013).
41. Yang, X., Sun, J., Xue, X., Zhu, W. & Hong, X. Development and characterization of 18 novel EST-SSRs from the western flower thrips, *Frankliniella occidentalis* (Pergande). *Int. J. Mol. Sci.* **13**, 2863–2876 (2012).
42. Peakall, R. & Smouse, P. E. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* **6**, 288–295 (2006).
43. Daianede, P. B. *et al.* Identification, characterization and validation of SSR markers from the gerbera EST database. *Plant Omics* **5**, 159–166 (2012).
44. Gao, Y., Lei, Z. & Reitz, S. R. Western flower thrips resistance to insecticides: detection, mechanisms and management strategies. *Pest Manag. Sci.* **68**, 1111–1121 (2012).
45. Eujayl, I., Sorrells, M., Baum, M., Wolters, P. & Powell, W. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theor. Appl. Genet.* **104**, 399–407 (2002).
46. Zeng, S. *et al.* Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* **11**, 94 (2010).
47. Barkley, N., Newman, M., Wang, M., Hotchkiss, M. & Pederson, G. Assessment of the genetic diversity and phylogenetic relationships of a temperate bamboo collection by using transferred EST-SSR markers. *Genome* **48**, 731–737 (2005).
48. Kashi, Y. & King, D. G. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* **22**, 253–259 (2006).
49. De Barro, P., Scott, K., Graham, G., Lange, C. & Schutze, M. Isolation and characterization of microsatellite loci in Bemisia tabaci. *Mol. Ecol. Notes* **3**, 40–43 (2003).
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 (2013).
51. Rotenberg, D. & Whitfield, A. Analysis of expressed sequence tags for *Frankliniella occidentalis*, the western flower thrips. *Insect Mol. Biol.* **19**, 537–551 (2010).
52. Frohlich, D. R., Torres-Jerez, I., Bedford, I. D., Markham, P. G. & Brown, J. K. A phylogeographical analysis of the *Bemisia tabaci* species complex based on mitochondrial DNA markers. *Mol. Ecol.* **8**, 1683–1691 (1999).
53. Pevzner, P. Computational molecular biology: an algorithmic approach. *Biotech Software & Internet Report*. 2, 167–169. (2001).
54. Yeh, F., Yang, R. & Boyle, T. POPGENE version 1.31 Microsoft window-based freeware for Population Genetic Analysis, University of Alberta. *Edmonton, AB, Canada* (1999).
55. Goudet, J. FSTAT A program for IBM PC compatibles to calculate Weir and Cockerham's (1984) estimators of F-statistics (version 1.2). *J. Hered.* **86**, 485–486 (1995).
56. Raymond, M. & Rousset, F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.* **86**, 248–249 (1995).
57. Nagy, S. *et al.* PICcalc: An online program to calculate polymorphic information content for molecular genetic studies. *Biochem. Genet.* **50**, 670–672 (2012).
58. Van Oosterhout, C., Hutchinson, W. F., Wills, D. P. & Shipley, P. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* **4**, 535–538 (2004).
59. Takezaki, N., Nei, M. & Tamura, K. POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. *Mol. Biol. Evol.* **27**, 747–752 (2010).
60. Nei, M. Genetic distance between populations. *Am. Nat.* **106**, 283–292 (1972).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Project number 31272104 and 31321063).

Author contributions

H.L.W., X.W.W. and S.S.L. conceived and designed the experimental plan. H.L.W. and J.Y. preformed the experiments. H.L.W., J.Y., L.M.B., Q.Y.Z., Y.J.W. and X.W.W. analyzed and interpreted the sequence data. H.L.W., J.Y., L.M.B., X.W.W. and S.S.L. drafted the manuscript. All authors read and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Wang, H.L. *et al.* Developing conversed microsatellite markers and their implications in evolutionary analysis of the *Bemisia tabaci* complex. *Sci. Rep.* **4**, 6351; DOI:10.1038/srep06351 (2014).