A novel method for single whitefly (*Bemisia tabaci*) transcriptomes reveals an eleven amino acid deletion in the *Nus*G protein in the bacterial endosymbiont *Portiera aleyrodidarum*

*Sseruwagi, P.[1], *Wainaina, J.M.[2,8], Ndunguru, J.[1], Tumuhimbise, R.[3], Tairo, F.[1], Guo, J.[4,5], Vrielink, A.[2], Blythe, A.[2], Kinene, T.[2,8], De Marchi, B.[2,6,8], Kehoe, M.A.[7], Tanz, S.K.[8], and Boykin, L.M.[2,8,9]

* P. Sseruwagi and J. M. Wainaina are joint first authors on this manuscript.

Addresses

[1]Mikocheni Agriculture Research Institute (MARI), P.O. Box 6226, Dar es Salaam, Tanzania

[2]School of Molecular Sciences, University of Western Australia, Crawley, Perth, WA 6009, Australia

[3]National Agricultural Research Laboratories, Kawanda, P.O. Box 7065, Kampala, Uganda

[4]Ministry of Agriculture Key Laboratory of Agricultural Entomology, Institute of Insect Sciences, Zhejiang University, Hangzhou 310058, China

[5]State Key Laboratory for the Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100193, China

[6] UNESP – Faculdade de Ciências Agronômicas, Botucatu, Brazil

[7]Crop Protection Branch, Departments of Agriculture and Food Western Australia, South Perth, WA 6151, Australia

[8]ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Perth, 6009, Western Australia, Australia

[9] Author for correspondence: laura.boykin@uwa.edu.au

Abstract

Background: *Bemisia tabaci* species (whiteflies) are the world's most devastating insect pests within crops in the tropics. They cause billions of dollars (US) of damage each year and are leaving farmers in the developing world food insecure. Understanding the genetic and transcriptomic composition of these insect pests, the viruses they transmit and the microbiota is crucial to sustainable insect and virus management solutions for farmers. Currently, publically available transcriptome data for *B. tabaci* has been generated from pooled samples (mainly inbred lab colonies) consisting of several individuals because whiteflies are small (approximately 0.2 mm wide and 0.1 mm in height). Pooling individuals can lead to high heterozygosity and skewed representation of the genetic diversity. The ability to extract enough RNA from a single whitefly has remained elusive due to their small size and technology limitations. Therefore, the understanding of whitefly-microbiota-viral species composition of an individual field-collected whitefly has also remained unknown. In this study, we developed a single whitefly RNA extraction procedure and subsequently successfully sequenced the transcriptome of four individual adult Sub-Saharan Africa (SSA1) *B. tabaci.*

Results: Transcriptome sequencing on individual whiteflies resulted in between 39-42 million raw reads. *De novo* assembly of trimmed reads yielded between 65,000-162,000 transcripts across all four *B. tabaci* transcriptomes. In addition, Bayesian phylogenetic analysis of mitochondrion cytochrome I oxidase (mtCOI) grouped the four whiteflies within the SSA1 clade. BLAST searches on assembled transcripts within the four individual transcriptomes identified five endosymbionts; the primary endosymbiont *Portiera aleyrodidarum* and four secondary endosymbionts: *Arsenophonus, Wolbachia, Rickettsia,* and *Cardinium spp.* These five endosymbionts were predominant across all four SSA1 *B.*

tabaci study samples with prevalence levels of between 54.1-75%. Nucleotide and amino acid sequence alignments of the *Nus*G gene of *P. aleyrodidarum* for the SSA1 *B. tabaci* transcriptomes of samples WF2 and WF2b revealed an eleven amino acid residue deletion that was absent in samples WF1 and WF2a. Comparison of the protein structure of the *Nus*G protein from *P. aleyrodidarum* in SSA1 with known *Nus*G structures showed the deletion resulted in a shorter D loop. Although *Nus*G is key in regulating of transcription elongation, it is believed that the shortening of the loop region in the N-terminal domain is unlikely to affect transcription termination. Therefore, the effect of variability in this region across species is unknown.

Conclusion: In this study, we optimised a single whitefly high quality RNA extraction procedure and successfully carried out individual whitefly transcriptome sequencing on adult *B. tabaci* whiteflies. This enabled the detection of unique genetic differences in the *Nus*G genes of the primary endosymbiont *P. aleyrodidarum* in four field-collected SSA1 whiteflies that may not have been detected using lab-pooled *B. tabaci* isolines. The use of field-collected specimens means that both time and money will be saved in future studies using single whitefly transcriptomes in monitoring vector and viral interactions. In addition, the methods we have developed here are applicable to any small organism where RNA quantity has limited transcriptome studies.

Keywords: Single whitefly, Transcriptomes, Amino acid, *Nus*G protein, Bacterial endosymbiont, *Portiera aleyrodidarum*

Background

Members of the whitefly *Bemisia tabaci* (Hemiptera: Aleyrodidae) species complex are classified as the world's most devastating insect pests. There are 34 species globally [1] and the various species in the complex are morphologically identical. They transmit over 100 plant viruses [2, 3], become insecticide resistant [4], and ultimately cause billions of dollars in damage annually for farmers. The adult whiteflies are promiscuous feeders and will move between viral infected crops and native weeds that act as viral inoculum 'sources' and deposit viruses to alternative crops that act as viral 'sinks' while feeding.

The crop of importance for this study was cassava (*Manihot esculenta*). Cassava supports approximately 800 million people in over 105 countries as a source of food and nutritional security, especially within rural smallholder farming communities [5]. Cassava production in Sub Saharan Africa (SSA), especially the East Africa region is hampered by both DNA and RNA transmitted viruses.

Whitefly-transmitted viruses cause cassava mosaic disease (CMD) leading to 28-40% crop losses with estimated economic losses of up to $ 2.7 billion dollars per year in SSA [6]. The CMD pandemics in East Africa, and across other cassava producing areas in SSA were correlated with *B. tabaci* outbreaks [7]. African cassava mosaic viruses (ACMVs) occur mostly towards West Africa where a distinct group of *B. tabaci* SSA1 is predominant. On the other hand, East African cassava mosaic viruses (EACMVs) occur mainly in coastal areas of East Africa with highest diversity inland in Kenya, Tanzania and Uganda, yet again with a different group of *B. tabaci* SSA1. The two distinct groups of SSA1 are yet to be named. While some studies have been carried out to determine the relative transmission of CMDs

by different *B. tabaci* species with indications of no significant differences, it is still not clear why some CMDs such as African cassava mosaic viruses (ACMVs) and East African cassava mosaic – Uganda variant (EACMV-UG), which is a recombinant between EACMV and ACMV in the coat protein (CP), do not occur in coastal East Africa.

Relevant to this study are two RNA *Potyviruses*: the Cassava Brown Streak viruses (CBSV) and the Uganda Cassava Brown Streak Virus (UCBSV) both devastating cassava in East Africa. *Bemisia tabaci* species have been hypothesized to transmit these RNA viruses with limited transmission efficiency [8,9]. Recent studies have shown that there are multiple species of these viruses [10], which further strengthens the need to obtain data from individual whiteflies as pooled samples could contain different species with different virus composition and transmission efficiency. In addition, CBSV has been shown to have a higher rate of evolution than UCBSV [11] increasing the urgency of understanding the role played by the different whitefly species in the system.

Endosymbionts and their role in B. tabaci

Viral-vector interactions within *B. tabaci* are further influenced by bacterial endosymbionts forming a tripartite interaction. *B. tabaci* has one of the highest numbers of endosymbiont bacterial infections with eight different vertically transmitted bacteria reported [12, 13], [14, 15]. They are classified into two categories; primary (P) and secondary (S) endosymbionts, many of which are in specialised cells called bacteriocytes, while a few are also found scattered throughout the whitefly body. A single obligate *P-symbiont P. aleyrodidarum* is systematically found in all *B. tabaci* individuals. *P. aleyrodidarum* is essential for whitefly survival as it supplies and complements the host metabolic activities in the synthesis of the

117 essential amino acids threonine and tryptophan along with the non-essential amino acid
118 serine [16]. *Portiera* has long co-evolutionary history with all members of the *Aleyrodinae*
119 subfamily [16]. Although it is yet to be confirmed in whiteflies, most P-symbionts have been
120 characteristically shown to have reduced and static genomes [17]. In this study, we further
121 explore genes within the *P. aleyrodidarum* retrieved from individual whitefly
122 transcriptomes, including the transcription termination/antitermination protein *Nus*G. *Nus*G
123 is a highly conserved protein regulator that suppresses RNA polymerase pausing and
124 increasing the elongation rate. However, its importance within gene regulation is species
125 specific; in *Staphylococcus aureus* it is dispensable [18, 19].
126
127 The S-endosymbionts are not systematically associated with hosts and their contribution is
128 not essential to the survival and reproduction. Seven facultative S-endosymbionts,
129 *Wolbachia, Cardinium, Rickettsia, Arsenophonus, Hamiltonella defensa* and *Fritschea*
130 *bemisae* have been detected in various *B. tabaci* populations [20, 21, 12, 22, 23]. The
131 presence of S-endosymbionts can influence key biological parameters of the host.
132 *Hamiltonella* and *Rickettsia* facilitate plant virus transmission with increased acquisition and
133 retention by whiteflies [24]. This is done by protection and safe transit of virions in
134 haemolymph of insects through chaperonins (*GroEL*) and protein complexes that aid in
135 protein folding and repair mechanisms [21].
136
137 ***Application of next generation sequencing in pest management of* B. tabaci**
138 The advent of next generation sequencing (NGS) and specifically transcriptome sequencing
139 has allowed the unmasking of this tripartite relationship of vector-viral-microbiota within
140 insects [25, 26, 27]. Furthermore, NGS provides an opportunity to better understand the co-

141 evolution of *B. tabaci* and its bacterial endosymbionts [28]. The endosymbionts have been
142 implicated in influencing species complex formation in *B. tabaci* through conducting sweeps
143 on the mitochondrial genome [29]. Applying transcriptome sequencing is essential to reveal
144 the endosymbionts and their effects on the mitogenome of *B. tabaci* and predict potential
145 hot spots for changes that are endosymbionts induced.
146
147 Several studies have explored the interaction between whitefly and endosymbionts and
148 have resulted in the identification of candidate genes that maintain the relationship [30,31].
149 This has been explored as a source of potential RNAi pesticide control targets [32, 31, 32,
150 28]. RNAi-based pest control measures also provide opportunities to identify species-
151 specific genes for target gene sequences for knock-down. However, to date all
152 transcriptome sequencing has involved pooled samples obtained through rearing several
153 generations of isolines of a single species to ensure high quantities of RNA for subsequent
154 sequencing. This remains a major bottle neck in particular within arthropoda where
155 collected samples are limited due to small morphological sizes [33, 34]. In addition, the
156 development of isolines is time consuming and often has colonies dying off mainly due to
157 inbreeding depression [35].
158
159 It is against this background that we sought to develop a method for single whitefly
160 transcriptomes to understand the virus diversity within different whitefly species. We did
161 not detect viral reads, probably an indication that the sampled whitefly was not carrying any
162 viruses, but as proof of concept of the method, we validated the utility of the data
163 generated by retrieving the microbiota *P-endosymbionts* and *S-endosymbionts* that have
164 previously been characterised within *B. tabaci* [36, 37] In this study we report the

165 endosymbionts present within field-collected individual African whiteflies and
166 characterisation and evolution of the *Nus*G genes present within the *P-endosymbionts*.
167
168 **Results**
169 ***RNA extraction and NGS optimised for individual B. tabaci samples***
170 In this study, we sampled four individual adult *B. tabaci* from cassava fields in Uganda (WF2)
171 and Tanzania (WF1, WF2a, WF2b). Total RNA from single whitefly yielded high quality RNA
172 with concentrations ranging from 69 ng to 244 ng that were used for library preparation and
173 subsequent sequencing with Illumina Hiseq 2000 on a rapid run mode. The number of raw
174 reads generated from each single whitefly ranged between 39,343,141 and 42,928,131
175 (Table 1). After trimming, the reads were assembled using Trinity resulting into 65,550 to
176 162,487 transcripts across the four SSA1 *B. tabaci* individuals (Table 1).
177
178 ***Comparison of endosymbionts within the SSA1* B. tabaci *samples***
179 Comparison of the diversity of bacterial endosymbionts across individual whitefly transcripts
180 was conducted with BLASTn searches on the non-redundant nucleotide database and by
181 identifying the number of genes from each bacterial endosymbiont (Supplementary Table.
182 1). We identified five main endosymbionts including: *P. aleyrodidarum* the primary
183 endosymbionts and four secondary endosymbionts: *Arsenophonus, Wolbachia, Rickettsia*
184 *sp, and Cardinium* spp (Table 2). *P. aleyrodidarum* predominated all four SSA1 *B. tabaci*
185 study samples with incidences of 74.8%, 71.2%, 54.1% and 58.5% for WF1, WF2, WF2a and
186 WF2b, respectively. This was followed by *Arsenophonus*, *Wolbachia, Rickettsia sp,* and
187 *Cardinium* spp, which occurred at an average of 18.0%, 5.9%, 1.6% and <1%, respectively
188 across all four study samples.

189

190 ***Phylogenetic analysis of single whitefly mitochondrial cytochrome oxidase I (COI)***

191 *B. tabaci* is recognized as a species complex of 34 species based on the mitochondrion
192 cytochrome oxidase I [38, 1, 39]. We therefore used cytochrome oxidase I (COI) transcripts
193 of the four individual whitefly to ascertain *B. tabaci* species status and their phylogenetic
194 relation using reference *B. tabaci* COI GenBank sequences found at www.whiteflybase.org.
195 All four COI sequences clustered within Sub Saharan Africa clade 1 (SSA1) species (data not
196 shown).

197

198 ***Sequence alignment and Bayesian phylogenetic analysis of Nus*G *gene***

199 Nucleotide and amino acid sequence alignments of the *Nus*G in *P. aleyrodidarum* were
200 conducted for several whitefly species including: *B. tabaci* (SSA1, Mediterranean (MED) and
201 Middle East Asia Minor 1 (MEAM1) New World 2 (NW2), *T. vaporariorum* (Greenhouse
202 whitefly) and *Alerodicus dispersus.* The alignment identified 11 missing amino acids in the
203 *Nus*G sequences for the SSA1 *B. tabaci* samples: WF2 and WF2b, *T. vaporariorum*
204 (Greenhouse whitefly) and *Alerodicus disperses.* However, all 11 amino acids were present
205 in samples WF1 and WF2a, MED, MEAM1 and NW2 (Fig. 1). Bayesian phylogenetic
206 relationships of the *Nus*G sequences of *P. aleyrodidarum* for the different whitefly species
207 clustered all four SSA1 *B. tabaci* (WF1, WF2, WF2a and WF2b) within a single clade together
208 with ancestral *B. tabaci* from GenBank (Fig. 2). The SSA1 clade was supported by posterior
209 probabilities of 1 with *T. vaporariorum* and *Alerodicus,* which formed clades at the base of
210 the phylogenetic tree (Fig. 2).

211

212 ***Structure analysis of Portiera Nus*G *genes***

213 Structures of the *Nus*G protein sequence of the primary endosymbiont *P. aleyrodidarum* in
214 the four SSA1 *B. tabaci* samples were predicated using Phyre2 with 100% confidence and
215 compared to known structures of *Nus*G from other bacterial species including (*Escherichia*
216 *coli, Thermus thermophiles*, and *Aquifex aeolicus*; (PDB entries 2KO6, 1NZ8 and 1M1H,
217 respectively) and Spt4/5 from yeast (*Saccharomyces cerevisiae;* PDB entry 2EXU) [18, 40,
218 41]. The 11-residue deletion was found in a loop region that is variable in length and
219 structure across bacterial species, but is absent from archaeal and eukaryotic species (Fig. 3
220 and Fig. 4A). The effect of the deletion appears to shorten the loop in *Nus*G from the African
221 whiteflies (WF2 and WF2b). A model of bacterial RNA polymerase (orange surface
222 representation; PDB entry 2O5I) bound to the N-terminal domain of the T. thermophiles
223 *Nus*G shows that the loop region is not involved in the interaction between *Nus*G and RNA
224 polymerase (Fig. 4B).
225

226 ## Discussion

227 In this study, we developed a single whitefly RNA extraction method for field-collected
228 samples. We subsequently successfully conducted transcriptome sequencing on individual
229 Sub-Saharan Africa 1 (SSA1) *B. tabaci*, revealing unique genetic diversity in the bacterial
230 endosymbionts as proof of concept.
231

232 ### *Nus*G deletion and implications within *P. aleyrodidarum* in SSA *B. tabaci*

233 We report the presence of the primary endosymbionts *P. aleyrodidarum* and several
234 secondary endosymbionts within SSA1 transcriptome. Furthermore, *P. aleyrodidarum* in
235 SSA1 *B. tabaci* was observed to have a deletion of 11 amino acids on the *Nus*G gene that is

236 associated with cellular transcriptional processes within another bacteria species. On the
237 other hand, *P. aleyrodidarum* from NW2, MED and SSA1 (WF2a, WF1) *B. tabaci* species did
238 not have this deletion (Fig. 1). The deleted 11 amino acids were identified in a loop region of
239 the N-terminal domain of *Nus*G protein resulting in a shortened loop in the SSA1 WF2b
240 sample. This loop region has high variability in both structure and length across bacterial
241 species and is absent from archaea and eukaryotic species.
242
243 *Nus*G is highly conserved and a major regulator of transcription elongation. It has been
244 shown to directly interact with RNA polymerase to regulate transcriptional pausing and rho-
245 dependent termination [19, 42, 18, 43]. Structural modelling of *Nus*G bound to RNA
246 polymerase indicated that the shortened loop region seen in the WF2b sample is unlikely to
247 affect this interaction. Rho-dependant termination has been attributed to the C-terminal
248 (KOW) domain region of *Nus*G, therefore a shortening of the loop region in the N-terminal
249 domain is also unlikely to affect transcription termination. Yet, there has been no function
250 attributed to this loop region of *Nus*G, and thus the effect of variability in this region across
251 species is unknown. However, the deletion could represent the result of evolutionary
252 species divergence. Further sequencing of the gene is required across the *B. tabaci* species
253 complex to gain further understanding of the diversity.
254
255 ***Why the single whitefly transcriptome approach?***
256 The sequencing of the whitefly transcriptome is crucial in understanding whitefly-
257 microbiota-viral dynamics and thus circumventing the bottlenecks posed in sequencing the
258 whitefly genome. The genome of whitefly is highly heterozygous [44]. Assembling of
259 heterozygous genomes is complex due to the de Bruijn graph structures predominantly used

[45]. To deal with the heterozygosity, previous studies have employed inbred lines obtained from raring a high number of whitefly isolines [46, 44, 33]. However, rearing whitefly isolines is time consuming and often colonies may suffer contaminations, leading to collapse and failure to raise the high numbers required for transcriptome sequencing.

We optimised the ARCTURUS® PicoPure® kit (Arcturus, CA, USA) protocol for individual whitefly RNA extraction with the dual aim of determining if we could obtain sufficient quantities of RNA from a single whitefly for transcriptome analysis and secondly, determine whether the optimised method would reveal whitefly microbiota as proof of concept. Using our method, the quantities of RNA obtained from field-collected single whitefly samples were sufficient for library preparation and subsequent transcriptome sequencing. Across all transcriptomes over 30M reads were obtained. The amount of transcripts were comparable to those reported in other arthropoda studies from field collections [34]. However, we did not observe any difference in assembly qualities as did [34]; probably due to the fact that our field-collected samples had degraded RNA based on RIN, and thus direct comparison with [34] was inappropriate.

Degraded insect specimen have been used successfully in previous studies [47]. This is significant, considering that a majority of insect specimen are usually collected under field conditions and stored in ethanol with different concentrations ranging from 70 to 100% [48, 49] rendering the samples liable to degradation. However, to ensure good keeping of insect specimen to be used for mRNA and total RNA isolation in molecular studies and other downstream applications such as histology and immunocytochemistry, it is advisable to collect the samples in an RNA stabilizing solution such as RNAlater. The solution stabilizes

and protects cellular RNA in intact, unfrozen tissue and cell samples without jeopardizing the quality or quantity of RNA obtained after subsequent RNA isolation. The success of the method provided an opportunity to unmask vector-microbiota-viral dynamics in individual whiteflies in our study, and will be useful for similar studies on other small organisms.

Endosymbionts diversity across individual SSA1 *B. tabaci* transcriptomes

In this study, we identified bacterial endosymbionts (Table 2) that were comparable to those previously reported in SSA1 *B. tabaci* on cassava [50, 23, 37]. Secondary endosymbionts have been implicated with different roles within *B. tabaci*. *Rickettsia* has been adversely reported across putative *B. tabaci* species, including the Eastern African region [51, 23, 51]. This endosymbiont has been associated with influencing thermo tolerance in *B. tabaci* species [52]. *Rickettsia* has also been associated with altering the reproductive system of *B. tabaci,* and within the females. This has been attributed to increasing fecundity, greater survival, host reproduction manipulation and the production of a higher proportion of daughters all of which increase the impact of virus [53]. *Arsenophonus, Wolbachia Arsenophonus* and *Cardinium spp* have been detected within MED and MEAM1 *Bemisia* species [12, 52]. In addition, [51] and [23] reported *Arsenophonus* within SSA1 *B. tabaci* in Eastern Africa that were collected on cassava. These endosymbionts have been associated with several deleterious functions within *B. tabaci* that include manipulating female-male host ratio through feminizing genetic males, coupled with male killing [54, 55].

Within the context of SSA agricultural systems, the role of endosymbionts in influencing *B. tabaci* viral transmission is important. Losses attributed to *B. tabaci* transmitted viruses

within different crops are estimated to be in billions of US dollars [48]. Bacterial endosymbionts have been associated with influencing viral acquisition, transmission and retention, such as in *Tomato leaf curl virus* [56, 24]. Thus, better understanding of the diversity of the endosymbionts provides additional evidence on which members of *B. tabaci* species complex more proficiently transmit viruses and thus the need for concerted efforts towards the whitefly eradication.

Conclusions

Our study provides a proof of concept that single whitefly RNA extraction and transcriptome sequencing is possible and the method is optimised and applicable to a range of small insect transcriptome studies. It is particularly useful in studies that wish to explore vector-microbiota-viral dynamics at individual insect level rather than pooling of insects. It is useful where genetic material is both limited as well of low quality, which is applicable to most agriculture field collections. In addition, the single whitefly transcriptome technique described in this study offers new opportunities to understand the biology and relative economic importance of the several whitefly species occurring in ecosystems within which food is produced in Sub-Saharan Africa, and will enable the efficient development and deployment of sustainable pest and disease management strategies to ensure food security in the developing countries.

Materials and methods

Whitefly sample collection and study design

In this study, we sampled whiteflies in Uganda and Tanzania from cassava (*Manihot esculenta*) fields. In Uganda, fresh adult whiteflies were collected from cassava fields at the National Crops Resources Research Institute (NaCRRI), Namulonge, Wakiso district, which located in central Uganda at 32°36'E and 0°31'N, and 1134 meters above sea level. On the other hand, the whiteflies obtained from Tanzania were collected on cassava in a countrywide survey conducted in 2013. The samples: WF2 (Uganda) and WF1, WF2a, and WF2b (Tanzania) used in this study were collected on CBSD-symptomatic cassava plants. In all the cases, the whitefly samples were kept in 70% ethanol in Eppendorf tubes until laboratory analysis. The whitefly samples were used for a two-fold function; firstly, to optimise a single whitefly RNA extraction protocol and secondly, to unmask RNA viruses and endosymbionts within *B. tabaci* as a proof of concept. In addition, data obtained from Nextera – DNA library prep from a Brazilian sample (156_NW2) was also used in this study. The whitefly was collected from a New World 2 colony in Brazil on *Euphorbia heterophylla* and kept in 70% ethanol in Eppendorf tubes until laboratory analysis.

Extraction of total RNA from single whitefly

RNA extraction was carried out using the ARCTURUS® PicoPure® kit (Arcturus, CA, USA). Briefly, 30 µl of extraction buffer was added to an RNase-free micro centrifuge tube containing a single whitefly and ground using a sterile plastic pestle. To the cell extract an equal volume of 70% ethanol was added. To bind the RNA onto the column, the RNA

purification columns were spun for two minutes at 100 x *g* and immediately followed by centrifugation at 16,000 x g for 30 seconds. The purification columns were then subjected to two washing steps using wash buffer 1 and 2 (ethyl alcohol). The purification column was transferred to a fresh RNase-free 0.5 ml micro centrifuge tube, with 30 ul of RNAse-free water added to elute the RNA. The column was incubated at room temperature for five minutes, and subsequently spun for one minute at 1,000 x g, followed by 16,000 x g for one minute. The eluted RNA was returned into the column and re-extracted to increase the concentration. Extracted RNA was treated with DNase using the TURBO DNA free kit as described by the manufacturer (Ambion, life Technologies, CA USA). Concentration of RNA was done in a vacuum centrifuge (Eppendorf, Germany) at room temperature for 1 hour, the pellet was suspended in 15 ul of RNase-free water and stored at -80 ^{0}C awaiting analysis. RNA was quantified, and the quality and integrity assessed using the 2100 Bioanalyzer (Agilent Technologies). Dilutions of up to x10 were made for each sample prior to analysis in the bioanalyzer.

cDNA and illumina library preparation

Total RNA from each individual whitefly sample was used for cDNA library preparation using the Illumina TruSeq Stranded Total RNA Preparation kit as described by the manufacturer (Illumina, San Diego, CA, USA). Subsequently, sequencing was carried out using the HiSeq2000 on the rapid run mode generating 2 x 50 bp paired-end reads. Base calling, quality assessment and image analysis were conducted using the HiSeq control software v1.4.8 and Real Time Analysis v1.18.61 at the Australian Genome Research Facility (Perth, Australia).

Analysis of NGS data using the supercomputer

Assembly of RNA transcripts: Raw RNA-Seq reads were trimmed using Trimmomatic. The trimmed reads were used for *de novo* assembly using Trinity [57] with the following parameters: time -p srun --export=all -n 1 -c ${NUM_THREADS} Trinity --seqType fq --max_memory 30G --left 2_1.fastq --right 2_2.fastq --SS_lib_type RF --CPU ${NUM_THREADS} --trimmomatic --cleanup --min_contig_length 1000 -output _trinity min_glue = 1, V = 10, edge-thr = 0.05, min_kmer_cov = 2, path_reinforcement_distance = 150, and group pairs distance = 500.

BLAST analysis of transcripts and annotation: BLAST searches of the transcripts under study were carried out on the NCBI (http://www.ncbi.nlm.nih.gov) non-redundant nucleotide database using the default cut-off on the Magnus Supercomputer at the Pawsey Supercomputer Centre Western Australia. Transcripts identical to known bacterial endosymbionts were identified and the number of genes from each identified endosymbiont bacteria determined.

Phylogenetic analysis of whitefly mitochondrial cytochrome oxidase I (COI): The phylogenetic relationship of mitochondrial cytochrome oxidase I (mtCOI) of the whitefly samples in this study were inferred using a Bayesian phylogenetic method implemented in MrBayes 3.2.2 [58] . The optimal substitution model was selected using Akaike Information Criteria (AIC) implemented in the jmodel test 2 [60].

Sequence alignment and phylogenetic analysis of *Nus*G gene in *P. aleyrodidarum* across *B. tabaci* species: Sequence alignment of the *Nus*G gene from the P-endosymbiont *P.*

400 *aleyrodidarum* from the SSA1 *B. tabaci* in this study was compared with another *B. tabaci*
401 species, *Trialeurodes vaporariorum* and *Alerodicus dispersus* using MAFFT *v7.017* [61]. The
402 Jmodel version 2 [60] was used to search for phylogenetic models with the Akaike
403 information criterion selecting the optimal that was to be implemented in MrBayes 3.2.2.
404 MrBayes run was carried out using the command: "lset nst=6 rates=gamma" for 50 million
405 generations, with trees sampled every 1000 generations. In each of the runs, the first 25%
406 (2,500) trees were discarded as burn in.
407
408 ***Analysis and modelling the structure of the Nus*G *gene***
409 The structures for *Portiera aleyrodidarum BT* and *B. tabaci* SSA1 whitefly were predicted
410 using Phyre2 [62] with 100% confidence and compared to known structures of *Nus*G from
411 other bacterial species. All models were prepared using Pymol (The PyMOL Molecular
412 Graphics System, Version 1.5.0.4).
413
414
415
416
417

418 **Table legends**

419

420 **Table 1** Summary statistics from De novo trinity assemble of Illumina paired end individual
421 whitefly transcriptome

422

423 **Table 2** Distribution of endosymbionts and number of genes present in endosymbionts
424 bacteria

425

426 **Figure legends**

427

428 **Fig. 1** Sequence alignment of nucleotide sequences of *Nus*G gene in *P. aleyrodidarum* across
429 whitefly species sequences using MAFFT v 7.017

430

431 **Fig. 2** Bayesian phylogenetic tree of *Nus*G gene of *P. aleyrodidarum* across whitefly species
432 using MrBayes -3.2.2

433

434 **Fig. 3** Structure of the *Nus*G gene showing the 11 amino acid deletion in a transcription
435 factor of the primary endosymbiont *Portiera aleyrodidarum* of the SSA1 *B. tabaci* species

436

437 **Fig. 4** Structure analysis of *Nus*G from *P. aleyrodidarum* in *B. tabaci* and other
438 endosymbionts **A.** Phyre2 based structure prediction of *Nus*G from *Candidatus Portiera*
439 *aleyrodidarum* in *B. tabaci* SSAI whitefly and comparisons to the structures of *Nus*G from
440 other bacterial species as indicated and of Spt4/5 from yeast. *Nus*G is coloured in grey, the
441 loop region in magenta and the 11-residue deletion is shown in green in the C. *Portiera*

442 *aleyrodidarum* structure. **B**. A model of bacterial RNA polymerase (orange surface
443 representation) bound to the N-terminal domain of the *T. thermophiles Nus*G (grey cartoon
444 representation)

References

[1] P. J. De Barro, S.-S. Liu, L. M. Boykin, and A. B. Dinsdale, "Bemisia tabaci: a statement of species status.," *Annu. Rev. Entomol.*, vol. 56, pp. 1–19, 2011.

[2] J. E. Polston and H. Capobianco, "Transmitting plant viruses using whiteflies.," J. Vis. *Exp.*, no. 81, p. e4332, 2013.

[3] D. R. Jones "Plant viruses transmitted by white ies," *Eur. J. Plant Pathol.*, vol. 109, pp. 195–219, 2003.

[4] V. Vassiliou, M. Emmanouilidou, A. Perrakis, E. Morou, J. Vontas, A. Tsagkarakou, and E. Roditakis, "Insecticide resistance in Bemisia tabaci from Cyprus," *Insect Sci.*, vol. 18, no. 1, pp. 30–39, 2011.

[5] FAO, *Save and Grow: Cassava. A Guide to Sustainable Production Intensification*. 2013.

[6] B. L. Patil and C. M. Fauquet, "Cassava mosaic geminiviruses : actual knowledge and perspectives," vol. 10, pp. 685–701, 2009.

[7] J. P. Legg, B. Owor, P. Sseruwagi, and J. Ndunguru, "Cassava Mosaic Virus Disease in East and Central Africa: Epidemiology and Management of A Regional Pandemic," *Adv. Virus Res.*, vol. 67, no. 6, pp. 355–418, 2006.

[8] M. N. Maruthi, R. J. Hillocks, K. Mtunda, M. D. Raya, M. Muhanna, H. Kiozia, A. R. Rekha, J. Colvin, and J. M. Thresh, "Transmission of Cassava brown streak virus by Bemisia tabaci (Gennadius)," *J. Phytopathol.*, vol. 153, no. 5, pp. 307–312, 2005.

[9] B. Mware, R. Narla, R. Amata, F. Olubayo, J. Songa, S. Kyamanyua, and E. M. Ateka, "Efficiency of cassava brown streak virus transmission by two whitefly species in coastal Kenya," *J. Gen. Mol. Virol.*, vol. 1, no. 4, pp. 40–45, 2009.

[10] J. Ndunguru, P. Sseruwagi, F. Tairo, F. Stomeo, S. Maina, A. Djinkeng, M. Kehoe, L. M.

Boykin, and U. Melcher, “Analyses of twelve new whole genome sequences of cassava brown streak viruses and ugandan cassava brown streak viruses from East Africa: Diversity, supercomputing and evidence for further speciation,” *PLoS One*, vol. 10, no. 10, pp. 1–18, 2015.

[11] T. Alicai, J. Ndunguru, P. Sseruwagi, F. Tairo, G. Okao-Okuja, R. Nanvubya, L. Kiiza, L. Kubatko, M. A. Kehoe, L. M. Boykin, , “Cassava brown streak virus has a rapidly evolving genome: implications for virus speciation, variability, diagnosis and host resistance,” *Sci. Rep.*, vol. 6, no. June, p. 36164, 2016.

[12] G. Gueguen, F. Vavre, O. Gnankine, M. Peterschmitt, D. Charif, E. Chiel, Y. Gottlieb, M. Ghanim, E. Zchori-Fein, and F. Fleury, “Endosymbiont metacommunities, mtDNA diversity and the evolution of the Bemisia tabaci (Hemiptera: Aleyrodidae) species complex,” *Mol. Ecol.*, vol. 19, no. 19, pp. 4365–4378, 2010.

[13] G. Gueguen, F. Vavre, O. Gnankine, M. Peterschmitt, D. Charif, E. Chiel, Y. Gottlieb, M. Ghanim, E. Zchori-Fein, and F. Fleury, “Endosymbiont metacommunities, mtDNA diversity and the evolution of the Bemisia tabaci (Hemiptera: Aleyrodidae) species complex,” *Mol. Ecol.*, vol. 19, pp. 4365–4378, 2010.

[14] X. L. Bing, Y. M. Ruan, Q. Rao, X. W. Wang, and S. S. Liu, “Diversity of secondary endosymbionts among different putative species of the whitefly Bemisia tabaci,” *Insect Sci.*, vol. 20, no. 2, pp. 194–206, 2013.

[15] J. M. Marubayashi, V. a. Yuki, K. C. G. Rocha, T. Mituti, F. M. Pelegrinotti, F. Z. Ferreira, M. F. Moura, J. Navas-Castillo, E. Moriones, M. a. Pavan, and R. Krause-Sakate, “At least two indigenous species of the Bemisia tabaci complex are present in Brazil,” *J. Appl. Entomol.*, vol. 137, pp. 113–121, 2013.

[16] M. L. Thao and P. Baumann, “Evolutionary relationships of primary prokaryotic

493 endosymbionts of whiteflies and their hosts," *Appl. Environ. Microbiol.*, vol. 70, no. 6,
494 pp. 3401–3406, 2004.
495 [17] N. Moran, J. P. McCutcheon, and A. Nakabachi, "Genomics and evolution of heritable
496 bacterial symbionts.," *Annu. Rev. Genet.*, vol. 42, pp. 165–190, 2008.
497 [18] R. A. Mooney, K. Schweimer, P. Rösch, M. Gottesman, and R. Landick, "Two
498 Structurally Independent Domains of E. coli NusG Create Regulatory Plasticity via
499 Distinct Interactions with RNA Polymerase and Regulators," *J. Mol. Biol.*, vol. 391, no.
500 2, pp. 341–358, 2009.
501 [19] A. V. Yakhnin, K. S. Murakami, and P. Babitzke, "NusG is a sequence-specific RNA
502 polymerase pause factor that binds to the non-template DNA within the paused
503 transcription bubble," *J. Biol. Chem.*, vol. 291, no. 10, pp. 5299–5308, 2016.
504 [20] E. Zchori-Fein and J. K. Brown, "Diversity of prokaryotes associated with Bemisia
505 tabaci (Gennadius) (Hemiptera : Aleyrodidae)," *Ann. Entomol. Soc. Am.*, vol. 95, no. 6,
506 pp. 711–718, 2002.
507 [21] Y. Gottlieb, M. Ghanim, E. Chiel, D. Gerling, V. Portnoy, S. Steinberg, G. Tzuri, a Rami,
508 E. Belausov, N. Mozes-daube, M. Gershon, S. Gal, N. Katzir, E. Zchori-fein, a R.
509 Horowitz, and S. Kontsedalov, "Identification and Localization of a Rickettsia sp . in
510 Bemisia tabaci (Homoptera : Aleyrodidae) Identification and Localization of a
511 Rickettsia sp . in Bemisia tabaci (Homoptera : Aleyrodidae)," *Appl. Environ.*
512 *Microbiol.*, vol. 72, no. 5, pp. 3646–3652, 2006.
513 [22] J. M. Marubayashi, A. Kliot, V. A. Yuki, J. A. M. Rezende, R. Krause-Sakate, M. A.
514 Pavan, and M. Ghanim, "Diversity and Localization of Bacterial Endosymbionts from
515 Whitefly Species Collected in Brazil," *PLoS One*, vol. 9, no. 9, p. e108363, 2014.
516 [23] S. Ghosh, P. S. Mitra, C. A. Loffredo, T. Trnovec, L. Murinova, E. Sovcikova, S.

517 Ghimbovschi, S. Zang, E. P. Hoffman, and S. K. Dutta, “Transcriptional profiling and
518 biological pathway analysis of human equivalence PCB exposure in vitro: Indicator of
519 disease and disorder development in humans,” *Environ. Res.*, vol. 138, pp. 202–216,
520 2015.
521 [24] A. Kliot, M. Cilia, H. Czosnek, and M. Ghanim, “Implication of the bacterial
522 endosymbiont *Rickettsia* spp. in interactions of the whitefly *Bemisia tabaci* with
523 tomato yellow leaf curl virus.,” *J. Virol.*, vol. 88, no. 10, pp. 5652–5660, 2014.
524 [25] K. Rosario, H. Capobianco, T. F. F. Ng, M. Breitbart, and J. E. Polston, “RNA viral
525 metagenome of whiteflies leads to the discovery and characterization of a whitefly-
526 transmitted carlavirus in North America.,” *PLoS One*, vol. 9, no. 1, p. e86748, 2014.
527 [26] K. Rosario, C. Marr, A. Varsani, S. Kraberger, D. Stainton, E. Moriones, J. E. Polston,
528 and M. Breitbart, “Begomovirus-associated satellite DNA diversity captured through
529 vector-enabled metagenomic (VEM) surveys using whiteflies (Aleyrodidae),” *Viruses*,
530 vol. 8, no. 2, pp. 1–16, 2016.
531 [27] K. Rosario, Y. M. Seah, C. Marr, A. Varsani, S. Kraberger, D. Stainton, E. Moriones, J. E.
532 Polston, S. Duffy, and M. Breitbart, “Vector-enabled metagenomic (VEM) surveys
533 using whiteflies (Aleyrodidae) reveal novel begomovirus species in the new and old
534 worlds,” *Viruses*, vol. 7, no. 10, pp. 5553–5570, 2015.
535 [28] M. F. Poelchau, B. S. Coates, C. P. Childers, A. A. Peréz De León, J. D. Evans, K. Hackett,
536 and D. Shoemaker, “Agricultural applications of insect ecological genomics,” *Curr.*
537 *Opin. Insect Sci.*, vol. 13, no. December 2015, pp. 61–69, 2016.
538 [29] D. E. Kapantaidaki, I. Ovčarenko, N. Fytrou, K. E. Knott, K. Bourtzis, and A.
539 Tsagkarakou, “Low Levels of Mitochondrial DNA and Symbiont Diversity in the
540 Worldwide Agricultural Pest, the Greenhouse Whitefly Trialeurodes vaporariorum

(Hemiptera: Aleyrodidae).," *J. Hered.*, pp. 1–13, 2014.

[30] S. Morin, M. Ghanim, I. Sobol, and H. Czosnek, "The GroEL protein of the whitefly Bemisia tabaci interacts with the coat protein of transmissible and nontransmissible begomoviruses in the yeast two-hybrid system.," *Virology*, vol. 276, pp. 404–416, 2000.

[31] J. Xue, X. Zhou, C.-X. Zhang, L.-L. Yu, H.-W. Fan, Z. Wang, H.-J. Xu, Y. Xi, Z.-R. Zhu, W.-W. Zhou, P.-L. Pan, B.-L. Li, J. K. Colbourne, H. Noda, Y. Suetsugu, T. Kobayashi, Y. Zheng, S. Liu, R. Zhang, Y. Liu, Y.-D. Luo, D.-M. Fang, Y. Chen, D.-L. Zhan, X.-D. Lv, Y. Cai, Z.-B. Wang, H.-J. Huang, R.-L. Cheng, X.-C. Zhang, Y.-H. Lou, B. Yu, J.-C. Zhuo, Y.-X. Ye, W.-Q. Zhang, Z.-C. Shen, H.-M. Yang, J. Wang, J. Wang, Y.-Y. Bao, and J.-A. Cheng, "Genomes of the rice pest brown planthopper and its endosymbionts reveal complex complementary contributions for host adaptation.," *Genome Biol.*, vol. 15, no. 12, p. 521, 2014.

[32] J. K. Kim, Y. J. Won, N. Nikoh, H. Nakayama, S. H. Han, Y. Kikuchi, Y. H. Rhee, H. Y. Park, J. Y. Kwon, K. Kurokawa, N. Dohmae, T. Fukatsu, and B. L. Lee, "Polyester synthesis genes associated with stress resistance are involved in an insect-bacterium symbiosis.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 26, pp. E2381-9, 2013.

[33] X.-W. Wang, Q.-Y. Zhao, J.-B. Luan, Y.-J. Wang, G.-H. Yan, and S.-S. Liu, "Analysis of a native whitefly transcriptome and its sequence divergence with two invasive whitefly species," *BMC Genomics*, vol. 13, no. 1, p. 529, 2012.

[34] N. Kono, H. Nakamura, and K. Arakawa, "Evaluation of the impact of RNA preservation methods of spiders for de novo transcriptome assembly," pp. 662–672, 2016.

[35] D. Charlesworth and J. H. Willis, "The genetics of inbreeding depression."

[36] X. L. Bing, W. Q. Xia, J. D. Gui, G. H. Yan, X. W. Wang, and S. S. Liu, "Diversity and evolution of the Wolbachia endosymbionts of Bemisia (Hemiptera: Aleyrodidae) whiteflies," *Ecol. Evol.*, vol. 4, pp. 2714–2737, 2014.

[37] Q. Rao, P.-A. Rollat-Farnier, D.-T. Zhu, D. Santos-Garcia, F. J. Silva, A. Moya, A. Latorre, C. C. Klein, F. Vavre, M.-F. Sagot, S.-S. Liu, L. Mouton, and X.-W. Wang, "Genome reduction and potential metabolic complementation of the dual endosymbionts in the whitefly Bemisia tabaci," *BMC Genomics*, vol. 16, no. 1, p. 226, 2015.

[38] L. M. Boykin, R. G. Shatters, R. C. Rosell, C. L. McKenzie, R. A. Bagnall, P. De Barro, and D. R. Frohlich, "Global relationships of Bemisia tabaci (Hemiptera: Aleyrodidae) revealed using Bayesian analysis of mitochondrial COI DNA sequences," *Mol. Phylogenet. Evol.*, vol. 44, pp. 1306–1319, 2007.

[39] C.-H. Hsieh, C.-C. Ko, C.-H. Chung, and H.-Y. Wang, "Multilocus approach to clarify species status and the divergence history of the Bemisia tabaci (Hemiptera: Aleyrodidae) species complex," *Mol. Phylogenet. Evol.*, vol. 76, pp. 172–180, 2014.

[40] P. Reay, K. Yamasaki, T. Terada, S. Kuramitsu, M. Shirouzu, and S. Yokoyama, "Structural and sequence comparisons arising from the solution structure of the transcription elongation factor NusG from Thermus thermophilus," *Proteins Struct. Funct. Genet.*, vol. 56, no. 1, pp. 40–51, 2004.

[41] T. Steiner, J. T. Kaiser, S. Marinkoviç, R. Huber, and M. C. Wahl, "Crystal structures of transcription factor NusG in light of its nucleic acid- and protein-binding activities," *EMBO J.*, vol. 21, no. 17, pp. 4641–4653, 2002.

[42] J. Li, R. Horwitz, S. McCracken, and J. Greenblatt, "NusG, a new Escherichia coli elongation factor involved in transcriptional antitermination by the N protein of phage ??," *J. Biol. Chem.*, vol. 267, no. 9, pp. 6012–6019, 1992.

589 [43] D. G. Vassylyev, M. N. Vassylyeva, A. Perederina, T. H. Tahirov, and I. Artsimovitch,
590 "Structural basis for transcription elongation by bacterial RNA polymerase.," *Nature*,
591 vol. 448, no. 7150, pp. 157–162, 2007.
592 [44] W. Xie, Q. shu Meng, Q. jun Wu, S. li Wang, X. Yang, N. na Yang, R. mei Li, X. guo Jiao,
593 H. peng Pan, B. ming Liu, Q. Su, B. yun Xu, S. nian Hu, X. guo Zhou, and Y. jun Zhang,
594 "Pyrosequencing the Bemisia tabaci transcriptome reveals a highly diverse bacterial
595 community and a robust system for insecticide resistance," *PLoS One*, vol. 7, no. 4,
596 pp. 1–13, 2012.
597 [45] R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M.
598 Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, and T. Itoh,
599 "Efficient de novo assembly of highly heterozygous genomes from whole-genome
600 shotgun short reads," *Genome Res.*, vol. 24, no. 8, pp. 1384–1395, 2014.
601 [46] R. L. S. J. and J. K. B. Dena Leshkowitz, Shirley Gazit, Eli Reuveni, Murad Ghanim,
602 Henryk Czosnek, CindyMcKenzie, "Whitefly (Bemisia tabaci) genome project: analysis
603 of sequenced clones from egg, instar, and adult (viruliferous and non-viruliferous)
604 cDNA libraries.," *BMC Genomics*, vol. 7, p. 79, 2015.
605 [47] I. Gallego Romero, A. A. Pai, J. Tung, and Y. Gilad, "RNA-seq: impact of RNA
606 degradation on transcript quantification.," *BMC Biol.*, vol. 12, no. 1, p. 42, 2014.
607 [48] J. P. Legg, P. Sseruwagi, S. Boniface, G. Okao-Okuja, R. Shirima, S. Bigirimana, G.
608 Gashaka, H. W. Herrmann, S. Jeremiah, H. Obiero, I. Ndyetabula, W. Tata-Hangy, C.
609 Masembe, and J. K. Brown, "Spatio-temporal patterns of genetic change amongst
610 populations of cassava Bemisia tabaci whiteflies driving virus pandemics in East and
611 Central Africa," *Virus Res.*, vol. 186, pp. 61–75, 2014.
612 [49] 7 and L. M. Boykin1 J. M. Wainaina, P. De Barro, L. Kubatko, M. A. Kehoe, J. Harvey, D.

613 Karanja, “Genetic Diversity , Population Structure and Species Delimitation of
614 Trialeurodes vaporariorum (greenhouse whitefly),” 2016.
615 [50] L. S. Tajebe, D. Guastella, V. Cavalieri, S. E. Kelly, M. S. Hunter, O. S. Lund, J. P. Legg,
616 and C. Rapisarda, “Diversity of symbiotic bacteria associated with Bemisia tabaci (
617 Homoptera : Aleyrodidae) in cassava mosaic disease pandemic areas of Tanzania,”
618 2014.
619 [51] L. S. Tajebe, D. Guastella, V. Cavalieri, S. E. Kelly, M. S. Hunter, O. S. Lund, J. P. Legg,
620 and C. Rapisarda, “Diversity of symbiotic bacteria associated with Bemisia tabaci
621 (Homoptera: Aleyrodidae) in cassava mosaic disease pandemic areas of Tanzania,”
622 *Ann. Appl. Biol.*, vol. 166, no. 2, pp. 297–310, 2015.
623 [52] M. Brumin, S. Kontsedalov, and M. Ghanim, “Rickettsia influences thermotolerance in
624 the whitefly Bemisia tabaci B biotype,” *Insect Sci.*, vol. 18, no. 1, pp. 57–66, 2011.
625 [53] A. G. Himler, T. Adachi-Hagimori, J. E. Bergen, A. Kozuch, S. E. Kelly, B. E. Tabashnik, E.
626 Chiel, V. E. Duckworth, T. J. Dennehy, E. Zchori-Fein, and M. S. Hunter, “Rapid spread
627 of a bacterial symbiont in an invasive whitefly is driven by fitness benefits and female
628 bias.,” *Science*, vol. 332, no. 6026, pp. 254–256, 2011.
629 [54] O. Duron, D. Bouchon, S. S. S. Boutin, L. Bellamy, L. Zhou, J. Engelstadter, G. D. Hurst,
630 J. Engelstädter, and G. D. Hurst, “The diversity of reproductive parasites among
631 arthropods: Wolbachia do not walk alone,” *BMC Biol.*, vol. 6, p. 27, 2008.
632 [55] J. Engelstädter and G. D. D. D. Hurst, “The ecology and evolution of microbes that
633 manipulate host reproduction,” *Annu. Rev. Ecol. Evol. Syst.*, vol. 40, no. 1, pp. 127–
634 149, 2009.
635 [56] Q. Su, H. Pan, B. Liu, D. Chu, W. Xie, Q. Wu, S. Wang, B. Xu, and Y. Zhang, “Insect
636 symbiont facilitates vector acquisition, retention, and transmission of plant virus.,”

637 *Sci. Rep.*, vol. 3, p. 1367, 2013.

638 [57] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X.
639 Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A.
640 Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman,
641 and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a
642 reference genome.," *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–52, 2011.

643 [58] J. P. Huelsenbeck, P. Andolfatto, and E. T. Huelsenbeck, "Structurama: Bayesian
644 inference of population structure," *Evol. Bioinforma.*, vol. 2011, no. 7, pp. 55–59,
645 2011.

646 [59] J. P. and R. F. Huelsenbeck, "MrBAYES : Bayesian inference of phylogenetic trees,"
647 *Interface*, vol. 17, no. 8, pp. 754–755, 2001.

648 [60] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada, "jModelTest 2: more models, new
649 heuristics and parallel computing," *Nat. Methods*, vol. 9, no. 8, pp. 772–772, 2012.

650 [61] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid
651 multiple sequence alignment based on fast Fourier transform.," *Nucleic Acids Res.*,
652 vol. 30, no. 14, pp. 3059–3066, 2002.

653 [62] L. A. Kelly, S. Mezulis, C. Yates, M. Wass, and M. Sternberg, "The Phyre2 web portal
654 for protein modelling, prediction, and analysis," *Nat. Protoc.*, vol. 10, no. 6, pp. 845–
655 858, 2015.

656

Declarations

Acknowledgements

J.M.W is supported by an Australian Award scholarship by the Department of Foreign Affairs and Trade (DFAT).

Availabiliy of data

All raw reads for the four whitefly have been deposited in NCBI SRA under the accession SRR5110306, SRR5110307, SRR5109958

Consent for publication

Not applicable

Competing interests

The authors declare that they have no conflict of interest.

Funding

This work was supported by Mikocheni Agricultural Research Institute (MARI), Tanzania through the “Disease Diagnostics for Sustainable Cassava Productivity in Africa” project, Grant no. OPP1052391 that is jointly funded by the Bill and Melinda Gates Foundation and The Department for International Development (DFID). The Pawsey Supercomputing Centre provided computational resources with funding from the Australian Government and the Government of Western Australia supported this work.

681

682 **Authors Contribution**

683 LB, PS, JN, JMW ST designed the research the experiment. JMW, JG performed RNA-seq

684 analysis. RT, FR, BD, TK, MK provided samples and laboratory experiments. AV, AB did the

685 NusG modelling. JMW, PS, LB wrote the manuscript. All authors read and approved the final

686 manuscript.

Table 1 Summary statistics from De novo Trinity assemble of Illumina paired end individual whitefly transcriptome

	WF1	WF2	WF2a	WF2b
Total Number of reads	39,343,141	42,587,057	42,513,188	42,928,131
Number of reads after trimming for quality	34,470,311 (87.61%)	39,898,821 (93.69%)	40,121,377 (94.37%)	40,781,932 (95.00%)
Transcripts	65,550	73,107	162,487	104,539
All transcript Contigs (N50)	505	525	1,084	1,018
Only longest contigs (N50)	468	484	707	746

Table 2 Distribution of endosymbionts and number of genes present within each endosymbiont bacteria present in four SSA1 *B. tabaci* samples from this study

Endosymbionts	**WF2**	**WF2a**	**WF2b**	**WF1**
Candidatus Portiera aleyrodidarum	322	302	408	312
Arsenophonus	3	1	1	1
Arsenophonus endosymbiont of Bemisia tabaci	1	2	1	2
Arsenophonus endosymbiont of Nilaparvata lugens	55	22	73	47
Arsenophonus nasoniae	33	13	34	12
Wolbachia endosymbiont of Cadra cautella	3	6	6	3
Wolbachia endosymbiont of Caudra cautella	NA	NA	NA	3
Wolbachia endosymbiont of Cimex lectularius	NA	5	2	NA
Wolbachia endosymbiont of Culex quinquefasciatus	2	9	3	2
Wolbachia endosymbiont of Diaphorina citri	3	NA	1	4
Wolbachia endosymbiont of Drosophila ananassae	6	14	15	6
Wolbachia endosymbiont of Drosophila simulans	1	6	5	3
Wolbachia endosymbiont of Operophtera brumata	NA	2	NA	2
Wolbachia endosymbiont of Muscidifurax uniraptor	1	NA	NA	NA
Wolbachia endosymbiont wVitA of Nasonia vitripennis phage WOVitA1/Wolbachia endosymbiont wVitB of Nasonia vitripennis phage WOVitB	1	6	3	NA
Wolbachia pipientis	5	9	4	4
Wolbachia pipientis wAlbB	5	4	2	NA
Wolbachia sp. wRi	3	13	8	5
Rickettsia africae	NA	2	1	NA
Rickettsia argasii T170-B	NA	4	7	NA

Rickettsia australis	NA	1	2	NA
Rickettsia buchneri	NA	NA	10	NA
Rickettsia Canadensis	NA	5	NA	NA
Rickettsia Helvetica	3	2	3	5
Rickettsia hoogstraalii	NA	NA	2	NA
Rickettsia japonica	NA	1	NA	NA
Rickettsia massiliae MTU5	NA	1	3	NA
Rickettsia monacensis	NA	NA	5	NA
Rickettsia prowazekii	NA	NA	4	NA
Rickettsia tamurae	NA	NA	1	NA
Rickettsia endosymbiont of Ixodes scapularis	1	20	9	1
Candidatus Rickettsia asemboensis	NA	NA	NA	1
Candidatus Rickettsia gravesii	NA	1	1	NA
Candidatus Rickettsia amblyommii str. Ac/Pa	NA	1	NA	NA
Candidatus Rickettsia amblyommii	NA	1	NA	NA
Candidatus Rickettsia amblyommii str. GAT-30V	NA	1	NA	NA
Rickettsiaceae bacterium Os18	NA	43	34	2
Rickettsiales bacterium Ac37b	1	4	4	NA
Rickettsia peacockii str. Rustic	NA	26	13	NA
Rickettsia bellii	1	10	10	1
Rickettsia felis str. Pedreira	NA	4	4	1
Rickettsia felis str. LSU	NA	5	8	NA
Rickettsia prowazekii str. GvF12	2	2	4	NA
Cardinium endosymbiont of Bemisia tabaci	NA	4	3	NA
Cardinium endosymbiont of Encarsia pergandiella	NA	5	3	NA
Candidatus Hamiltonella defensa	NA	1	NA	NA

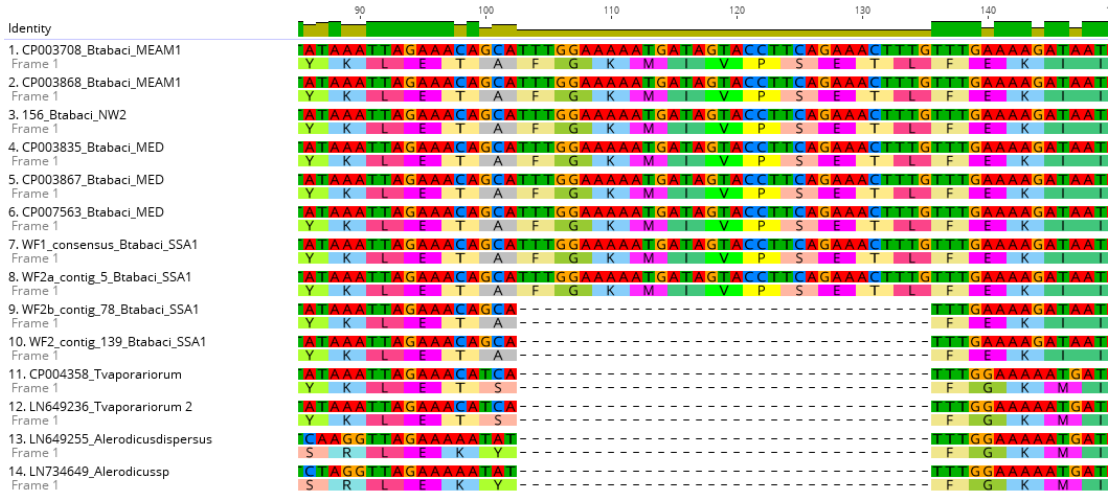

Fig. 1 Sequence alignment of nucleotide sequences of *Nus*G gene in *P. aleyrodidarum* across whitefly species sequences using MAFFT v 7.017

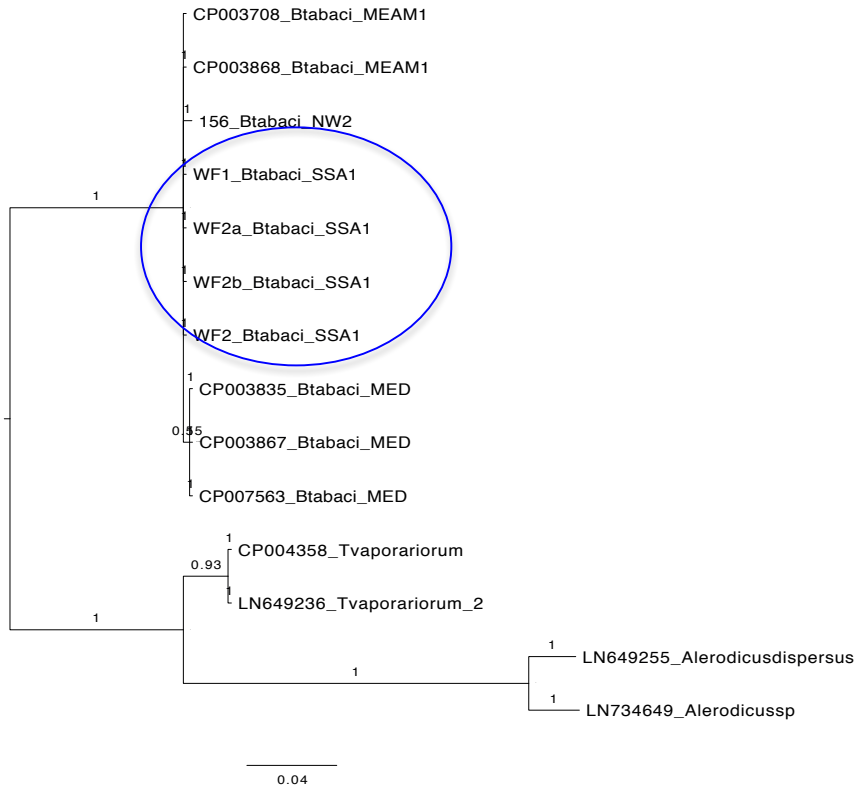

Fig. 2 Bayesian phylogenetic tree of *Nus*G gene of *P. aleyrodidarum* across whitefly species using MrBayes -3.2.2. Circled are *B. tabaci* samples from this study

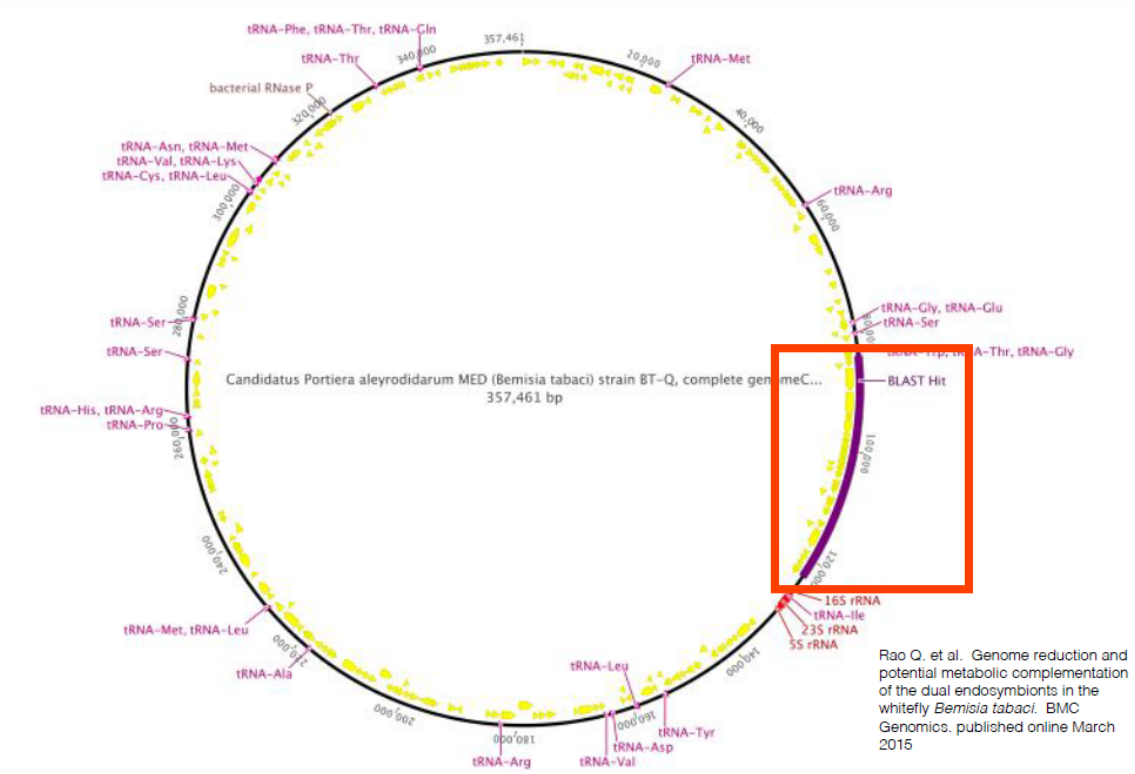

Fig. 3 Structure of the *Nus*G gene showing the 11 amino acid deletion in a transcription factor of the primary endosymbiont *Portiera aleyrodidarum* of the SSA1 *B. tabaci* species

-

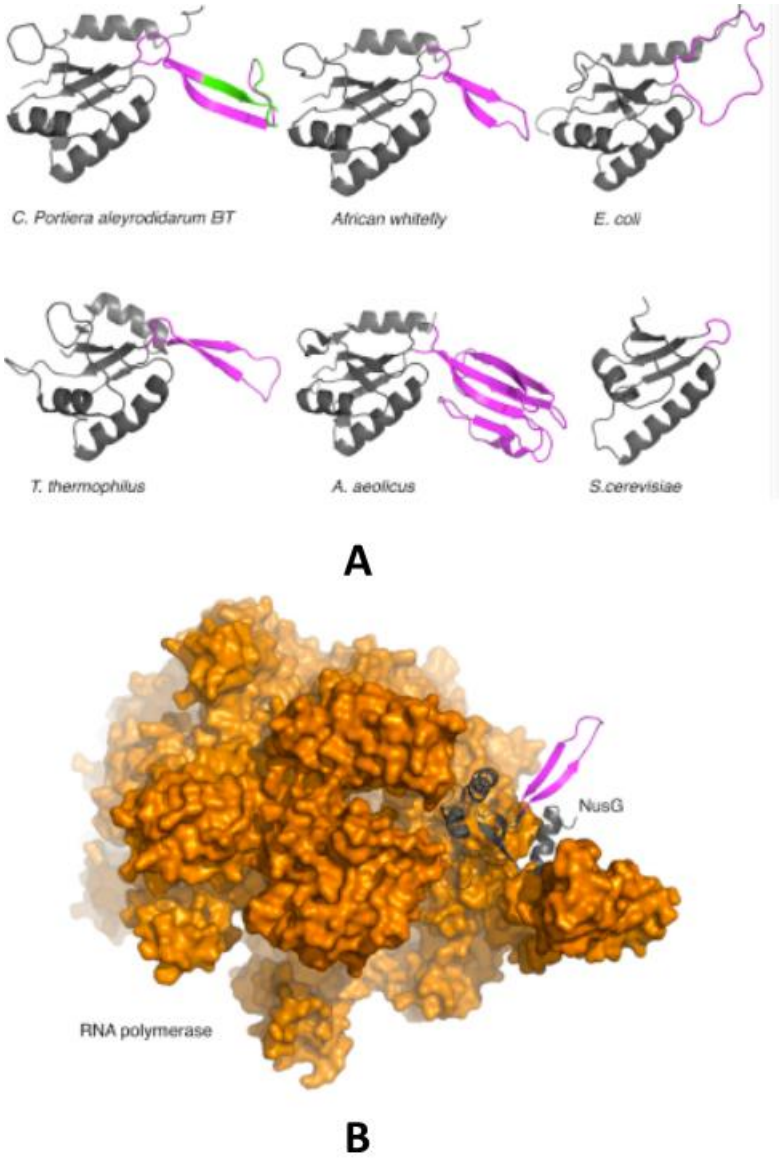

Fig. 4 Structure analysis of *Nus*G from *P. aleyrodidarum* in *B. tabaci* and other endosymbionts **A.** Phyre2 based structure prediction of *Nus*G from *Candidatus Portiera aleyrodidarum* in *B. tabaci* SSAI whitefly and comparisons to the structures of *Nus*G from other bacterial species as indicated and of Spt4/5 from yeast. *Nus*G is coloured in grey, the loop region in magenta and the 11-residue deletion is shown in green in the C. *Portiera aleyrodidarum* structure. **B**. A model of bacterial RNA polymerase (orange surface representation) bound to the N-terminal domain of the *T. thermophiles Nus*G (grey cartoon representation)