Evaluation of odorant-binding protein-1 as a molecular marker for identifying biological forms and delimitating sibling species of *Anopheles stephensi*

Om P. Singh[1]*, Shobhna Mishra[1], Ankita Sindhania[1], Taranjeet Kaur[1], U. Sreehari[2], Manoj K. Das[3], Gunjan Sharma[1]

[1]National Institute of Malaria Research, Sector 8, Dwarka, New Delhi, India
[2]National Institute of Malaria Research, Field Unit, Bengaluru, India
[3]National Institute of Malaria Research, Field Unit, Ranchi, India

*Correspondence: dropsingh@gmail.com, singh@mrcindia.org

Abstract:

Background: *Anopheles stephensi*, an invasive malaria vector, has been reported to have three biological forms identifiable based on the number of ridges present on the egg's floats and the dimension of eggs. Recently, these forms have been designated as sibling species based on the fixed differences in the DNA sequence of the first intron of the odorant-binding protein-1 (*AsteObp1*). In this study, we evaluated the utility of this neutral marker in designating sibling species or identifying biological forms.

Methods: Field collected and laboratory-reared *An. stephensi* were characterized for biological forms based on the number of floats on egg-ridge. DNA sequencing of the partial *AsteObp1* gene of *An. stephensi* individuals were performed by Sanger's method, either directly or after cloning with a plasmid vector.

Results: *AsteObp1* intron-1 in Indian *An. stephensi* populations are highly polymorphic with the presence of more than 12 haplotypes exhibiting nucleotide- as well as length-polymorphism (90-to-121 bp). A majority of the field samples were heterozygous (up to 89% in the field populations). The phasing of haplotypes in heterozygotes through Sanger's sequencing was challenging due to indels (1-to-24 bp) at multiple loci. No specific haplotype or monophyletic clade of intron-1 was found associated with a specific biological form. The inbreeding coefficient for this marker was close to zero in field and laboratory populations which refute the existence of sibling species based on the *AsteObp1* marker.

Conclusions: *AsteObp1* cannot serve as a marker for the identification of biological forms of *An. stephensi*. The probable existence of sibling species in *An. stephensi* based on the *AsteObp1* intron-1 is refuted.

Keywords: *Anopheles stephensi*, *var. mysorensis*, odorant-binding protein, sibling species, malaria

1. Introduction

Anopheles stephensi is an efficient malaria vector mainly responsible for urban malaria. Earlier, this species was reported distributed mainly in countries of the Middle East and South Asia (Dash et al., 2007). During the last few decades, the distribution of this species has expanded to new geographical localities, such as Lakshadweep island of India (Sharma & Hamzakoya, 2001), Sri Lanka (Gayan Dharmasiri et al., 2017), countries of the Horns of Africa (Faulde et al., 2014; Carter et al., 2018; WHO, 2019; Seyfarth et al., 2019; Balkew et al., 2020) and the Republic of Sudan (WHO, 2019; Ahmed et al., 2021), and established as a malaria vector at least in Djibouti, Horn of Africa (Faulde et al., 2014; de Santi et al., 2019; Seyfarth et al., 2019). A high probability of their presence within many urban cities across Africa has been predicted by Sinka et al. (2020), which poses a major threat to the elimination of urban malaria in African continent. In consequence of the recent invasions of this species in several countries, World Health Organization issued an alert to the affected and surrounding countries to take immediate remedial actions (WHO, 2019).

An. stephensi in India is predominantly found in urban areas due to the presence of favourable breeding habitats in urban settings and is considered as an urban malaria vector. Perhaps an adaptation to the urban environment makes this species invasive. The presence of two races, i.e., 'type form' and '*var. mysorensis*' in *An. stephensi* was described by Sweet and Rao (1937) and Rao et al. (1938) based on the eggs' dimension, length of eggs-floats and numbers of ridges on the float. The 'type form' was reported mainly from the urban area and '*var. mysorensis*' from the rural area (Sweet & Rao, 1937; Subbarao et al., 1987). Initial reciprocal crossing experiments showed a certain degree of sterility between these two races (Rao et al., 1938) but subsequently, it was demonstrated that there is no reproductive barrier between these two races (hereafter, described as 'biological form') and the proposed status of subspecies was refuted by Rutledge et al. (1970). Later, an additional form "intermediate" was reported by Subbarao et al. (1987) based on the genetic evidence derived from crossing experiments between the 'type form' and '*var. mysorensis'*.

The 'type form' is reported to be an efficient vector while '*var. mysorensis'* is considered as a poor malaria vector (Subbarao et al., 1987), though the latter is reported to be a vector in Iran (Manouchehri et al., 1976; Oshaghi et al., 2006a). Laboratory feeding experiments suggested differences in the success of sporogonic development of rodent malaria parasites in the biological forms (Shinzawa et al., 2013; Basseri et al., 2013). Due to the suggested differential epidemiological implications of biological forms, it is desirable to determine their biological forms in all epidemiological studies. The classical method for the identification of biological forms of *An. stephensi* involves morphometrics of eggs which is a cumbersome process; for which, live female mosquitoes are to be transported to the laboratory and their eggs have to be obtained for the counting of ridges on the egg's float. Nagpal et al. (2003) reported that the 'spiracular index' (ratio of the length of anterior spiracle with the length of the thorax) of *An. stephensi* can be used for the discrimination of biological forms in the adult mosquitoes, which was higher in 'type form' as compared to '*var. mysorensis'* in Rajasthan, India (an arid zone). They also reported that the two forms of *An. stephensi*, i.e., 'type form' and '*var. mysorensis'*, occupies different ecological niches where '*var. mysorensis*' prefer outdoor habitat and 'type form' indoor. The 'spiracular index', which measures the degree of spiracular opening, may serve as an adaption to water-loss regulation, and therefore need to be evaluated in humid areas also. Recently, a molecular marker was introduced by

Gholizadeh et al. (2015) for the identification of all the three biological forms. They sequenced odorant-binding protein 1 (*AsteObp1*) of *An. stephensi* from four laboratory strains, of which two strains were *'var. mysorensis'*, one was 'type form' and one was 'intermediate'. Based on sequences of a total of 20 samples from all four colonies, they demonstrated fixed differences (no heterozygotes were reported) in the first intron of the gene among three biological forms. This prompted them to propose *AsteObp1* intron-1 as a molecular marker for the identification of biological forms. However, this marker was not tested on the field populations. Subsequently, the same research group (Firooziyan et al., 2018), designated the three biological forms as distinct sibling species, i.e., A, B and C, that corresponds to type form, 'intermediate' and *'var. mysorensis'*, respectively. However, they didn't establish an association of the *AsteObp1* intron-1 haplotypes with eggs' morphology. In absence of any study showing the association of *AsteObp1* haplotypes with biological form in field populations, we investigated the extent of polymorphism in *AsteObp1* intron-1 in Indian *An. stephensi* (field populations as well as laboratory colonies) and validate their association with biological forms or proposed sibling species.

2. Material & methods

2.1.Mosquito sampling and processing

The fully blood-engorged and semi-gravid female mosquitoes were collected from urban localities of Bengaluru (13°00' N, 77°63' E) from Karnataka state, villages Karamcnand Pur (28°04' N, 77°13' E) and Chilawali (28°17' N, 76°95' E) of Mewat district (Haryana), and village Ghummanhera (28°53' N, 76°92' E), of South-West Delhi district, in the morning between 6:00 and 8:00 AM with the help of an aspirator and flash-torch. The mosquitoes were transferred to a thermocol (polystyrene) box designed to transport live mosquitoes. The live mosquitoes were transported to the insectary maintained at 27 °C and 70-75% RH and transferred into insect cage. Mosquitoes were provided access to soaked raisins and water-pads during transportation and maintenance in the insectary. Mosquitoes were allowed to attain the gravid stage. Once gravid, individual mosquitoes were identified provisionally using a magnifying glass and *An. stephensi* were transferred inside a paper cup containing a small amount of water and having the inner side lined with water-soaked filter paper. The mosquitoes were kept in the cup overnight to allow them to lay their eggs. The next morning, female mosquitoes which laid their eggs were removed, anesthetized with diethyl ether, and confirmatory identification of species was performed under a binocular microscope. The mosquitoes were preserved in individual microfuge tubes containing a few drops of isopropanol for molecular studies. The eggs of corresponding mosquitoes laid on filter paper were removed onto a glass micro-slide and checked for the number of ridges present on one side of eggs' float under a stereo-binocular microscope using a 10X objective lens. Finally, the mode number of ridges was determined for each egg batch. Mosquitoes from laboratory colonies of *An. stephensi*, one originating from New Delhi (urban area) and being maintained since the year 2011, was also characterized for the number of ridges on eggs' float. DNA was isolated from individual mosquitoes using the method by Livak (1984).

Besides, preserved DNA of *An. stephensi* individuals from a laboratory colony (F5) originating from Chennai (urban area), wild-caught mosquitoes from urban localities of Gurgaon (National

Capital Region) were also used in this study. Genomic DNA of *An. stephensi* strain STE2 (Indian origin) was obtained through BEI Resources, NIAID, NIH, USA.

2.2.PCR amplification, cloning and sequencing

Initially, the amplification of partial *AsteObp1* was performed using primers OBP1F1 (5'-CGT AGG TGG AAT ATA GGT GG-3') and OBP1R2 (5'-TCG GCG TAA CCA TAT TTG C-3' (Gholizadeh et al., 2015) which covers full intron-1 and partial intron-2. PCR was performed with Hot Start Taq 2X Master Mix (New England Biolabs Inc) in a 20 µL reaction mixture containing 1.5 mM of $MgCl_2$, 0.5 unit of Taq polymerase and 0.25 µM of each primer. The PCR cycling conditions were: a denaturation step at 95 °C for 3 min, 35 cycles each with a denaturation step at 95 °C for 30 sec, annealing at 55 °C for 30 sec and extension at 72 °C for 45 sec followed by a final extension at 72 °C for 5 min. PCR products were cleaned using ExoSAP-IT™ (ThermoFisher Scientific) and sequenced on both strands of DNA using ABI BigDye Terminator v3.2 (ThermoFisher Scientific) following the manufacturer's protocol.

During initial sequencing of the PCR product amplified with primers OBP1F1 and OBP1R2, we failed to identify *AsteObp1* haplotypes because a majority of samples were heterozygote and their sequence chromatograms were inconspicuous with the presence of peaks of mixed bases due to the presence of indels at multiple loci. Therefore, to phase out haplotypes, the PCR products of nine heterozygote individuals were cloned for sequencing. For cloning, PCR products were amplified using a high-fidelity Taq DNA polymerase in order to avoid or minimize PCR error. Briefly, PCR reactions were performed in a total volume of 25 µL containing 0.5 units of Phusion Taq DNA polymerase, Phusion HF reaction buffer containing 1.5 mM $MgCl_2$ (New England Biolabs Inc), 200 mM of each dNTP, 0.5 µM of each primer and 0.5 µL of DNA. PCR thermal cycling conditions were: initial denaturation at 98 °C for 30 sec, 35 cycles of denaturation at 98 °C for 10 sec, annealing at 65 °C for 30 sec and extension at 72 °C for 45 sec, followed by a final extension at 72 °C for 2 min. Five µL of PCR product was loaded on 2% agarose gel and visualized under UV illumination. The remaining PCR products were purified using QIAquick PCR purification kit (Qiagen Inc) following the vendor's protocol. The purified PCR product was 'A'-tailed by dATP in the presence of Taq polymerase. The reaction mixture (50 µL) contained 0.25 µL of Taq polymerase (KAPA Biosystem, USA), 5µL of 10X buffer containing 1.5 mM $MgCl_2$, 5µL of purified PCR product, 1µL of 10mM dATP and incubated at 70 °C for 20 min. Three µL of purified A-tailed PCR product was then ligated into pGemT®-T vector (Promega Corporation) following guidelines recommended by the manufacturer. Then five µL of the ligated product was transformed into DH5-alpha competent *E. coli* (New England Biolab) and 90 µL of the transformed product was spread onto Ampicillin-IPTG/X-Gal LB agar plates for blue/white colony screening, which was incubated overnight at 37 °C. Plasmid DNA of five clones from each mosquito was PCR amplified and sequenced from both directions of the DNA strand as described above.

For subsequent sequencing reactions for the characterization of full intron-1 (short-PCR), we designed a different primer set (OBP_INTF: 5'-CGC CGT GAT GCC GAA TA-3' and OBP_INTR: 5'-ATT GTC GTC CAC CAC CTT G-3') from flanking exons. The majority of samples were amplified and sequenced using this primer set. The PCR conditions were similar to what we used for the PCR amplification for direct sequencing using primers OBP1F1 and OBP1R2, except the duration of extension time in PCR-cycling, which was reduced to 30 sec.

2.3.Haplotype phasing in heterozygotes

Due to the presence of indels at multiple loci in the intron-1, it was challenging to assign the intron-haplotype of individual mosquitoes from their DNA chromatogram. The presence of indel leads to overlapping peaks (**Figure 1**). For correct basecalling of mixed bases (as per IUB nomenclature) in such ambiguous sequence chromatogram, we analysed the 'ab1' files (sequence chromatogram) using ABI Sequence Analysis Software with modified basecaller algorithm set to call IUB mixed bases if the secondary peak is ≥ 25% of the highest peak. Finally, the sequence chromatograms were manually checked and edited for correct basecalling of mixed bases (**Figure 1**). Manual editing was done because the software failed to recognize mixed bases at some positions in some sequence chromatogram. To phase out the mixed haplotypes in individual sequences, we identified the theoretical pattern of sequence of mixed bases for each haplotype pair in the forward direction based on all haplotypes identified through sequencing of cloned samples, from homozygotes and heterozygotes with polymorphism at single-nucleotide only. The theoretical pattern of sequences of various mixed haplotypes combination following NC-IUB/IUPAC nomenclature was created using the Microsoft-Excel's logic function. All the possible 66 sequence combinations (number of combinations: (n2-n)/2, where n= number of haplotypes) formed by 12 haplotypes in the forward sequence reads have been shown in **Figure 2**. These sequence patterns were used to phase out haplotypes in a heterozygote mosquito using a forward sequence read chromatogram.

2.4.Genetic distance and phylogenetic analysis

The genetic distances among haplotypes and molecular phylogenetic analysis were performed through software MEGA7 (Kumar et al., 2016) using the Maximum Likelihood method based on the Kimura 2-parameter model (K2P). All sites were included in the analysis. The bootstrap consensus tree inferred from 500 replicates was taken to represent the evolutionary history.

3. Results:

3.1.Characterization of *AsteObp1* intron-1 haplotypes

Initial direct sequencing chromatogram of partial *AsteObp1*gene, covering full intron-1 and partial intron-2 (amplicon size ~850 bp), was not readable in the majority of samples sequenced (heterozygotes) due to the ambiguity in sequence reads. Assuming that the ambiguity is due to the presence of indel/s in haplotypes of heterozygous individuals, identification of haplotypes was done through sequencing of cloned PCR products. A total of 10 haplotypes (H1 to H10) were identified through sequencing of cloned PCR products from nine heterozygote individuals (**Figure 3**). To avoid the recognition of false haplotype due to PCR error in clonal samples, the haplotype which was identical in at least a minimum of two clones was considered as a definite haplotype. Direct sequencing of the short-PCR product revealed the presence of additional four haplotypes (based on intron-1 only; designated as H11 to H14) identified based on the homozygous sequence or heterozygous sequence with polymorphism at a single nucleotide base. Considering only intron-1 (the region proposed as a marker for the discrimination of biological forms or sibling species by Gholizadeh et al., 2015 and Firooziyan et al., 2018), there were 12 haplotypes because haplotype H1 was identical to H7, and H4 was identical to H5 in the intron-1. Thus there were 12 haplotypes

considering only the intron-1. Indels were present in both introns but the intron-1 exhibited a higher degree of length polymorphism which varied between 90 and 121 bp (90, 114, 116, 120 and 121 bp).

3.2.Genetic distances and phylogenetic relationship

The K2P genetic distances between haplotypes are shown in **Table-1**. The maximum-likelihood phylogenetic tree with the highest log likelihood inferred from haplotypes sequences (intron-1 only) is shown in **Figure 4**. The haplotypes formed two distinct clades, Clade A and B, with high genetic distances between haplotypes ranging from 16 to 25%. The designated haplotypes following Gholizadeh et al. (2015) for 'type form' (H4) and 'intermediate' form (H2) fall under clade A and the designated haplotype for the '*var. mysorensis*' form (H8) fall under clade B. The two clades can be further divided into monophyletic subclades A1 and A2 within clade A and B1, B2 within clade B (**Figure 4**). The monophyletic clades A1 includes the designated haplotype for 'type form' (H4) as well as 'intermediate' form (H2) and B1 includes designated haplotype for '*var. mysorensis*' form (H8). We classified monophyletic clades 'type form' comprising H4, '*var. mysorensis'* comprising H8 and H14, and 'intermediate' comprising H2, H3, H10 and H12 (shown highlighted in **Figure 4**). Five haplotypes, H1, H6, H11, H9 and H13 could not be assigned to any of these monophyletic clades and remain orphan..

3.3.Haplotype-identification, heterozygosity and inbreeding coefficient

The haplotyping of heterozygous mosquitoes was done based on a mixed-base sequence pattern deduced theoretically for each haplotype combination as depicted in **Figure 2**. The distribution of haplotypes in laboratory colonies and field populations is shown in **Table 1**. Haplotypes in some heterozygous sequence reads could not be phased out due to the presence of additional haplotypes that was not characterized in this study. The *AsteObp1* intron-1 exhibited extensive polymorphism and heterozygosity in field populations. The laboratory colony (Delhi) which are being maintained in the laboratory since 2011 had only two haplotypes with balanced heterozygosity (H_O=46%; H_E=45%). A high degree of heterozygosity with balanced polymorphism was observed in field populations, i.e., in Karamcand Pur (H_O=87%, H_E=87%) and Bengaluru (H_O=83%, H_E=79%) (**Table 1**). In all cases, the inbreeding coefficient was close to zero (-0.05 to 0.00), suggesting random breeding (**Table 1**). In populations with a sample size of less than 10, we didn't calculate heterozygosity and inbreeding coefficient, but most of the individuals in such populations were heterozygote.

3.4.Association of haplotypes with biological forms of *An. stephensi*

3.4.1. Delhi and agglomerated urban areas

The laboratory colony of *An. stephensi*, originating from New Delhi, was characterized as 'type form' with the mode number of egg ridges ranging from 16 to 20 (average: 18.29). Out of the 28 samples sequenced from this colony, none of them were found to have haplotype H4—designated haplotype for type form. Instead, two haplotypes, H8 (designated for *var. mysorensis*) and H9 (orphan), were found with frequencies 0.34 and 0.66, respectively, both of which fall in clade B. The haplotypes H8 and H9 were far distantly related to designated haplotype for 'type form' (H4) with K2P distance 19 and 24%, respectively (Table 2). Another strain originating from Delhi, STE2 strain (obtained from BEI Resources, NIAID, NIH), was found homozygote for haplotype H8 (*var. mysorensis*). Sequencing of DNA of a small number of *An. stephensi* from Gurgaon (agglomerated

part of National Capital Region) revealed the presence of H8 (*var. mysorensis*), H9 and H12. None of the above urban strains had H4 haplotype (type form)

3.4.2. Chennai (urban)

A total of six samples were sequenced from another laboratory colony (F5) originating from Chennai--a metropolitan city. None of them had H4 haplotype. The majority of haplotypes (9 out of 12) present were H6 and H11, which fall in monophyletic clade 'A2' unrelated to all three monophyletic clades of biological forms (type form, *var. mysorensis* and intermediate).

3.4.3. Bengaluru (urban)

The mode number of egg ridges of *An. stephensi* collected from Bengaluru urban areas ranged between 16 and 18 (average: 16.55) which can be categorized as 'type form'. Sequencing of 18 samples revealed absence of the H4 haplotype (designated for type form). The haplotypes present in this population belonged to 'intermediate' clade (H2, H3 and H10; 47%), mysorensis clade (H8; 28%), and orphan clades (H6 and H11; 25%)

3.4.4. Karamcand Pur village

The average number of ridges in the Karamcand Pur population ranged between 12 and 16 (average: 14.85) which can be categorized as mixture of 'intermediate' and '*var. mysorensis*' forms . The Karamcand Pur population had all the haplotypes identified in this study (**Table 1**) except H12 and H14. The haplotypes present in this population belonged to 'intermediate' clade (H2, H3, H10 and H12; 40%), mysorensis clade (H8; 17%), 'type form' clade (H4; 4%) and orphan clades (H1, H6, H9, H11 and H13; 39%)

3.4.5. Chilawali village

The average number of ridges in the Chilwali population ranged between 11 and 14 (average: 13.18), which can be categorized as '*var. mysorensis'*. Out of a total of 18 haplotype alleles identified, eight belonged to the 'intermediate' clade (H2, H10 and H12), one to 'type form' (H4) and nine to orphan clades (H6, H9 and H11). Expected haplotype H8 (*var. mysorensis*) was absent.

In all populations we failed to establish the association of number of egg ridges with a specific haplotype. Further, in all field populations the haplotypes belonging to 'type form', '*var. mysorensis'* and 'intermediate' were found in heterozygous conditions exhibiting lack of assortative mating.

4. Discussion

The existence of two races in the *An. stephensi* was suggested long back based on the egg morphology (Sweet & Rao, 1938). The initial inference that genetic incompatibility exists between these races (Sweet et al., 1938) was refuted by Rutledge et al. (1970). Subbarao et al. (1987) also found genetic compatibility between races in laboratory settings but suggested that there may be pre-copulatory barriers in the field conditions. As such, there are conflicting reports with regards to gene flow between races. Extensive gene flow among the three forms of *An. stephensi* (type form, '*var mysorensis'* and 'intermediate') was reported in Iran based on mitochondrial genes *cytochrome oxidase 1* and *cytochrome oxidase 2* (*COI* and *COII*) (Chavshin et al. 2014; Oshaghi et al., 2006b). In contrast, a study in India using microsatellite markers indicated high genetic differentiation (F_{ST}) and negligible gene flow between the three variants (Vipin et al., 2010). Irrespective of biological forms, the Saudi Arabian *An. stephensi* have been found to be different from all other populations, viz., India, Pakistan, Djibouti, Ethiopia and derived colony strains based on *COI* sequence (Munawar

et al., 2020; Walter Reed Biosystematics Unit website). However, the ITS2 sequence of *An. stephensi* from Saudi Arabia (Munawar et al., 2020) is no different from Indian, Sri Lankan and African populations (Mishra et al., 2021, Surendran et al., 2018; Tadesse et al., 2019), exhibiting mito-nuclear discordance (Singh et al., 2021). Recent data on field populations of *An. stephensi* collected from Iran and Afghanistan (Firooziyan et al., 2018) suggest the complete absence of heterozygotes in the *AsteObp1* gene among three biological forms in sympatric populations and suggested the three biological forms to be distinct sibling species. However, the conclusion drawn by Firooziyan et al. (2018) is based on small sample size and required validation.

Regardless of the status of gene flow among biological forms, evidence suggests that they are ecological variants and differences in their biological characteristics such as preferred breeding habitat (Sweet & Rao, 1937) malaria vectorial competence (Shinzawa et al., 2013; Basseri et al., 2013), adaptability to breed in captivity (Sweet & Rao, 1938) and preferred ecological niche (Subbarao et al.,1987; Nagpal et al., 2003) have been documented. 'Type form' is mostly found in urban areas and '*var. mysorensis*' in rural areas with a cline of increasing ridge numbers on their eggs from rural to urban habitats (Subbarao et al.,1987). Nagpal et al. (2003) found that the 'type form' is found resting indoor environment and '*var. mysorensis*' outdoor. The 'type form' has been regarded as an efficient malaria vector responsible for urban malaria and '*var. mysorensis*' is regarded as a poor vector (Subbarao et al., 1987). Due to the supposed variability in the role of 'type form' and '*var. mysorensis*' in malaria transmission, it is vital to identify biological forms in a field population.

Previous attempts to identify a molecular marker for the differentiation of 'type form' and '*var. mysorensis*' based on ribosomal DNA (ITS2 and 28S rDNA) failed (Alam et al. 2008; Mishra et al., 2021). However, Gholizadeh et al. (2015) reported that the three biological forms have fixed *AsteObp1* intron-1 haplotype and proposed intron-1 as a molecular marker for their identification. The observation, by Gholizadeh et al. (2015), that fixed differences do exist between different biological forms in intron-1 was based on studies on the four inbred laboratory colonies (two of which have completed over 200 generations) representing three biological forms. However, the proposed diagnostic marker has never been tested on field mosquitoes that have been characterized for biological forms based on the egg morphology. Such exercise was desirable because the fixation of a haplotype in laboratory colonies can be due to the 'founder effect', where a small subset of a population may have been used to establish a colony, which tends to reduce genetic variability and increases homozygosity over generations (Moorad et al., 2005). Similar effect was seen in this study where we found reduced heterozygosity in a laboratory colony (New Delhi) with just two haplotypes in contrast to field populations that were highly polymorphic. Therefore, any conclusion on the association of a haplotype with biological form based on laboratory colonies is flawed. We also raise doubt if *An. stephensi* colonies studied for the *AsteObp1* gene by Gholizadeh et al. (2015) were pure for a single *AsteObp1* haplotype. It appears that in their article, the prefix 'I' of the sample identifiers stands for Iranshahr strain (*var. mysorensis*) 'B' stands for Bandar-Abbas strain (*var. mysorensis*), K stands for Kazerun strain (intermediate) and T stands for Type form. In that case, all strains, except Bandar-Abbas, had two haplotypes (see Figures 4 and 5 in the article by Gholizadeh et al. (2015).

In another study by the same group (Firooziyan et al., 2018), the biological forms were designated as sibling species (species A, B and C), but the assignment of biological form was based

on the preoccupied assumption of definite association of haplotype with biological form. Association of a haplotype with biological form identified based on the egg morphology was not established. The limited data suggested the absence of heterozygotes based on sequences of five samples from Afghanistan and 13 samples from Iran. There is no mention of the presence or absence of heterozygotes in the paper. It is intriguing that in both the studies by the Gholizadeh group (Gholizadeh et al., 2015; Firooziyan et al., 2018), large numbers of samples (150 and 100 samples, respectively) were PCR-amplified, but only a few samples were sequenced. It is not clear if heterozygotes were present in the population, and the same were excluded from the study because of difficulty in basecalling. In this study, we observed a high degree of heterozygosity in Indian field populations, as expected assuming that there is unrestricted gene flow. This is in contrast to the findings of Firooziyan et al. (2018), where we find high heterozygosity in field population. The observed high degree of heterozygosity (>80%) in the field population is not unusual considering the number of haplotypes present in the population. The heterozygosity is expected to increase with the increase of the number of haplotype alleles. With the presence of 12 haplotype-alleles, theoretically, we expect maximum heterozygosity up to 0.9 when alleles are equally frequent ($H_{E\ \mathrm{Max}}=1-(1/n)$; see Appendix S1 for the derivation)

A limited number of sequencing in this study revealed the presence of at least 12 haplotypes of intron-1. The number will keep increasing with the increase of sample size and populations because introns, being neutral, are subjected to a high rate of evolution. Keeping in view the frequent presence of indels in intron-1, identification of this marker is not feasible through Sanger's sequencing method without cloning.

In this study, we didn't find an association of any monophyletic clade of intron-1 haplotype with specific biological forms. There was a complete lack of assortative mating, as evidenced by the presence of heterozygotes of haplotypes belonging to all monophyletic clades. We didn't find H4 haplotype (designated for type form) in laboratory colony or field populations characterized as 'type form'. Similarly, the H8 haplotype (designated for '*var. mysorensis*' form) was absent from a population characterized as '*var. mysorensis*'. One strain SDA-500 (origin: Pakistan), widely used in different laboratories, was characterized as an 'intermediate' form by Shinzawa et al. (2013) with an average egg ridge number of 15.14. When we retrieved intron-1 sequences of this strain from publicly available whole-genome sequence database from VectorBase and GenBank (GeneBank accession number: CP032232), it was found to be H8 haplotype. Following designation assigned by Gholizadeh et al. (2015), it can be assigned to '*var. mysorensis'*. Therefore, designation of biological form based on *AsObp1* cannot be done.

It is important to note that there are no differences in the deduced amino acid sequence of different haplotypes (Gholizadeh et al., 2015 and this study). Therefore the adaptability of different biological forms in a different ecological niche is certainly not governed by *the AstObp1* gene and probably some other gene/s govern the adaptability of biological forms to a specific niche. Whole-genome sequencing may reveal the genetic basis of ecological adaptation of different biological forms.

This present study deals with Indian *An. stephensi* populations where we do not find any evidence of reproductive barrier between mosquitoes with different *AsteObp1* intron-1 haplotypes. Whether the reproductive barrier does exist in this species in the Pacific region needs to be

confirmed by haplotyping field population for *AsteObp1* marker and establishing their association with egg morphology.

5. Conclusions

AsteObp1-intron is not a suitable marker for the identification of biological forms at least in Indian populations. No specific haplotype was found associated with biological forms. The presence of a high degree of heterozygosity and low inbreeding coefficient in Indian populations refutes the probable existence of sibling species in *An. stephensi* based on *AsteObp1* marker.

Acknowledgment

Authors are thankful to the Director, NIMR for providing facility to conduct the work. The technical assistance rendered by Mr Uday Prakash, Mr Kanwar Singh and Mr Sadarudddin is acknowledged.

The following reagent was obtained through BEI Resources, NIAID, NIH: Genomic DNA from *Anopheles stephensi*, Strain STE2, MRA-144, contributed by Mark Q. Benedict.

Funding

The study was funded by the DST-SERB Science & Engineering Research Board (Department of Science & Technology)

Authors' contribution

OPS conceived the study, analyzed the data and wrote the first draft of the manuscript; SM and GS performed cloning and sequencing experiments, AS and US performed morphological studies; NK contributed to the manuscript.

Availability of data and materials:

All data generated in the study are included in the article.

References

Ahmed A, Khogali R, Elnour MB, Nakao R, Salim B. Emergence of the invasive malaria vector *Anopheles stephensi* in Khartoum State, Central Sudan. Parasit Vectors. 2021;14(1):511. doi: 10.1186/s13071-021-05026-4.

Alam MT, Bora H, Das MK, Sharma YD. The type and mysorensis forms of the *Anopheles stephensi* (Diptera: Culicidae) in India exhibit identical ribosomal DNA ITS2 and domain-3 sequences. Parasitol Res. 2008;103:75-80. doi: 10.1007/s00436-008-0930-7.

Balkew M, Mumba P, Dengela D, Yohannes G, Getachew D, Yared S, et al. Geographical distribution of *Anopheles stephensi* in eastern Ethiopia. Parasit Vectors. 2020; 13:35. doi: 10.1186/s13071-020-3904-y.

Basseri HR, Mohamadzadeh Hajipirloo H, Mohammadi Bavani M, Whitten MM. Comparative susceptibility of different biological forms of *Anopheles stephensi* to *Plasmodium berghei* ANKA strain. PLoS One. 2013; 8:e75413. doi: 10.1371/journal.pone.0075413.

Carter TE, Yared S, Gebresilassie A, Bonnell V, Damodaran L, Lopez K, et al. First detection of *Anopheles stephensi* Liston, 1901 (Diptera: Culicidae) in Ethiopia using molecular and morphological approaches. Acta Trop. 2018; 188:180-186. doi: 10.1016/j.actatropica.2018.09.001.

Chavshin AR, Oshaghi MA, Vatandoost H, Hanafi-Bojd AA, Raeisi A, Nikpoor F. Molecular characterization, biological forms and sporozoite rate of *Anopheles stephensi* in southern Iran. Asian Pac J Trop Biomed. 2014; 4: 47-51. doi:10.1016/S2221-1691(14)60207-0

Dash AP, Adak T, Raghavendra K, Singh OP. The biology and control of malaria vectors in India. Curr Sci. 2007; 92, 1571-1578.

de Santi VP, Khaireh BA, Chiniard T, Pradines B, Taudon N, Larréché S, Mohamed AB, de Laval F, Berger F, Gala F, Mokrane M, Benoit N, Malan L, Abdi AA, Briolant S. Role of *Anopheles stephensi* Mosquitoes in Malaria Outbreak, Djibouti, 2019. Emerg Infect Dis. 2021; 27:1697-1700. doi: 10.3201/eid2706.204557.

Faulde MK, Rueda LM, Khaireh BA. First record of the Asian malaria vector *Anopheles stephensi* and its possible role in the resurgence of malaria in Djibouti, Horn of Africa. Acta Trop. 2014; 139:39-43. doi: 10.1016/j.actatropica.2014.06.016.

Firooziyan S, Dinparast Djadid N, Gholizadeh S. Speculation on the possibility for introducing *Anopheles stephensi* as a species complex: preliminary evidence based on odorant binding protein 1 intron I sequence. Malar J. 2018; 17:366. doi: 10.1186/s12936-018-2523-y

Gayan Dharmasiri AG, Perera AY, Harishchandra J, Herath H, Aravindan K, Jayasooriya HTR, et al. First record of *Anopheles stephensi* in Sri Lanka: a potential challenge for prevention of malaria reintroduction. Malar J. 2017; 16:326. doi: 10.1186/s12936-017-1977-7.

Gholizadeh S, Firooziyan S, Ladonni H, Hajipirloo HM, Djadid ND, Hosseini A, Raz A. The *Anopheles stephensi* odorant binding protein 1 (AsteObp1) gene: a new molecular marker for biological forms diagnosis. Acta Trop. 2015; 146:101-13. doi: 10.1016/j.actatropica.2015.03.012.

Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33:1870-4. doi: 10.1093/molbev/msw054.

Livak KJ. Organization and mapping of a sequence on the *Drosophila melanogaster* X and Y chromosomes that is transcribed during spermatogenesis. Genetics. 1984; 107:611-34.

Manouchehri AV, Javadian E, Eshighy N, Motabar M. Ecology of *Anopheles stephensi* Liston in southern Iran. Trop Geogr Med. 1976;28(3):228-232.

Mishra S, Sharma G, Das MK, Pande V, Singh OP. Intragenomic sequence variations in the second internal transcribed spacer (ITS2) ribosomal DNA of the malaria vector *Anopheles stephensi*. PLoS One. 2021; 16: e0253173. doi: 10.1371/journal.pone.0253173.

Moorad JA, Wade MJ. A genetic interpretation of the variation in inbreeding depression. Genetics. 2005;170: 1373-84. doi: 10.1534/genetics.104.033373.

Munawar K, Saleh A, Afzal M, et al. Molecular characterization and phylogenetic analysis of anopheline (Anophelinae: Culicidae) mosquitoes of the Oriental and Afrotropical Zoogeographic zones in Saudi Arabia. Acta Trop. 2020;207:105494. doi:10.1016/j.actatropica.2020.105494

Nagpal BN, Srivastava A, Kalra NL, Subbarao SK. Spiracular indices in *Anopheles stephensi*: a taxonomic tool to identify ecological variants. J Med Entomol. 2003;40: 747-9. doi: 10.1603/0022-2585-40.6.747. PMID: 14765648.

Oshaghi MA, Yaaghoobi F, Abaie M. Pattern of mitochondrial DNA variation between and within *Anopheles stephensi* (Diptera: Culicidae) biological forms suggests extensive gene flow. Acta Trop. 2006b; 99: 226–233

Oshaghi MA, Yaaghoobi F, Vatandoost H, Abaei MR. *Anopheles stephensi* biological forms: geographical distribution and malaria transmission in malarious region of Iran. Pak J Biol Sci. 2006a; 9: 294–6.

Rao BA, Sweet WC, Subbarao, A.M. Ova measurements of *A. stephensi* type and *A. stephensi var. mysorensis*. Journal of the Malaria Institute of India. 1938; 1,261-266.

Rutledge LC, Ward RA, Bickley WE. Experimental hybridization of geographic strains of *Anopheles stephensi* (Diptera: Culicidae). Ann Entomol Soc Am. 1970;63:1024-30.

Seyfarth M, Khaireh BA, Abdi AA, Bouh SM, Faulde MK. Five years following first detection of *Anopheles stephensi* (Diptera: Culicidae) in Djibouti, Horn of Africa: populations established-malaria emerging. Parasitol Res. 2019; 118:725-732. doi: 10.1007/s00436-019-06213-0.

Sharma SK, Hamzakoya KK. Geographical spread of *Anopheles stephensi* vector of urban Malaria, and *Aedes aegypti*, vector of dengue/ DHF, in the Arabian Sea Islands of Lakshadweep, India. Dengue Bulletin. 2001; 25: 88–91. https://apps.who.int/iris/handle/10665/148798

Shinzawa N, Ishino T, Tachibana M, Tsuboi T, Torii M. Phenotypic dissection of a Plasmodium-refractory strain of malaria vector *Anopheles stephensi*: the reduced susceptibility to *P. berghei* and *P. yoelii*. PLoS One. 2013; 8: e63753. doi: 10.1371/journal.pone.0063753.

Singh OP, Sindhania A, Sharma G, et al. Are members of the *Anopheles fluviatilis* complex conspecific?. Acta Trop. 2021;224:106149. doi:10.1016/j.actatropica.2021.106149

Sinka ME, Pironon S, Massey NC, et al. A new malaria vector in Africa: Predicting the expansion range of *Anopheles stephensi* and identifying the urban populations at risk. Proc Natl Acad Sci U S A. 2020;117(40):24900-24908. doi:10.1073/pnas.2003976117

Subbarao SK, Vasantha K, Adak T, Sharma VP, Curtis CF. Egg-float ridge number in *Anopheles stephensi*: ecological variation and genetic analysis. Med Vet Entomol. 1987; 1: 265-71. doi: 10.1111/j.1365-2915.1987.tb00353.x.

Surendran SN, Sivabalakrishnan K, Gajapathy K, Arthiyan S, Jayadas TTP, Karvannan K, et al. Genotype and biotype of invasive *Anopheles stephensi* in Mannar Island of Sri Lanka. Parasit Vectors. 2018; 11:3. doi: 10.1186/s13071-017-2601-y.

Sweet W, Rao B, Rao A. Cross-breeding of *An. stephensi* type and *An. stephensi var. mysorensis*. J Malar Inst India. 1938;1:149-54.

Sweet WC, Rao BA. Races of *A. stephensi* Liston,1901. Indian Medical Gazette. 1937; 72,665-674.

Tadesse FG, Ashine T, Teka H, et al. *Anopheles stephensi* mosquitoes as vectors of *Plasmodium vivax* and *falciparum*, Horn of Africa, 2019. Emerg Infect Dis. 2021;27(2):603-607. doi:10.3201/eid2702.200019

Tyagi BK. A review of the emergence of *Plasmodium falciparum* dominated malaria in irrigated areas of the Thar Desert, India. Acta Trop. 2004;89:227–39

Vipin, Dube M, Gakhar SK. Genetic differentiation between three ecological variants ('type', 'mysorensis' and 'intermediate') of malaria vector *Anopheles stephensi* (Diptera: Culicidae) Insect Science. 2010; 17: 335–343, doi 10.1111/j.1744-7917.2010.01316.x

WHO. Vector alert: *Anopheles stephensi* invasion and spread. 26 August 2019. https://www.who.int/news/item/26-08-2019-vector-alert-anopheles-stephensi-invasion-and-spread (retrieved: 1 October, 2021)

WRBU. *Anopheles stephensi* Liston, 1901. https://www.wrbu.si.edu/vectorspecies/mosquitoes/stephensi (retrieved 8 November, 2021)

Table 1: Distribution of *AsteObp1* intron-1 haplotypes in different population

Locality	Number of haplotype-alleles (frequency)														H_O	H_E	F_S
	H1	H2	H3	H4	H6	H8	H9	H10	H11	H12	H13	H14	Total	UN*			
Karamchand Pur (rural)	9 (0.10)	17 (0.19)	4 (0.04)	4 (0.04)	13 (0.14)	15 (0.17)	8 (0.09)	13 (0.14)	3 (0.03)	0	4 (0.04)	0	88	2	0.867	0.870	0.0037
Chilawali (rural)	0	2 (0.13)	0	1 (0.06)	4 (0.25)	0	3 (0.19)	3 (0.19)	1 (0.06)	1 (0.06)	1 (0.06)	0	16	6	ND	ND	ND
New Delhi† (urban)	0	0	0	0	0	19 (0.34)	37 (0.66)	0	0	0	0	0	56	0	0.464	0.448	-0.0355
Bengaluru (urban)	0	6 (0.17)	9 (0.25)	0	1 (0.03)	9 (0.25)	0	3 (0.08)	8 (0.22)	0	0	0	36	0	0.833	0.790	-0.0547
Gurgaon (urban)	0	0	0	0	0	3 (0.30)	3 (0.30)	0	0	4 (0.40)	0	0	10	0	ND	ND	ND
Chennai† (urban)	1 (0.08)	1 (0.08)	1 (0.08)	0	7 (0.58)	0	0	0	2 (0.17)	0	0	0	12	0	ND	ND	ND

H_O = observed heterozygosity; H_E = expected heterozygosity; F_S = inbreeding coefficient; *UN= could not be identified due to the presence of new haplotype in heterozygous condition; ND = not done when the sample size was <10; †laboratory colony.

Table 2. Evolutionary divergence between *AsteObp1* intron-1 haplotypes (number of base substitutions per site from between sequences) using the Kimura 2-parameter model

	H1	H2	H3	H4	H6	H8	H9	H10	H11	H12	H13	H14
H1	-											
H2	0.01	-										
H3	0.01	0.00	-									
H4	0.04	0.05	0.05	-								
H6	0.06	0.08	0.08	0.09	-							
H8	0.16	0.17	0.17	0.19	0.20	-						
H9	0.20	0.22	0.22	0.24	0.25	0.12	-					
H10	0.01	0.00	0.00	0.05	0.08	0.17	0.22	-				
H11	0.06	0.08	0.08	0.09	0.00	0.20	0.25	0.08	-			
H12	0.02	0.01	0.01	0.06	0.09	0.16	0.20	0.01	0.09	-		
H13	0.20	0.22	0.22	0.24	0.25	0.09	0.02	0.22	0.25	0.20	-	
H14	0.16	0.17	0.17	0.19	0.20	0.00	0.12	0.17	0.20	0.16	0.09	-

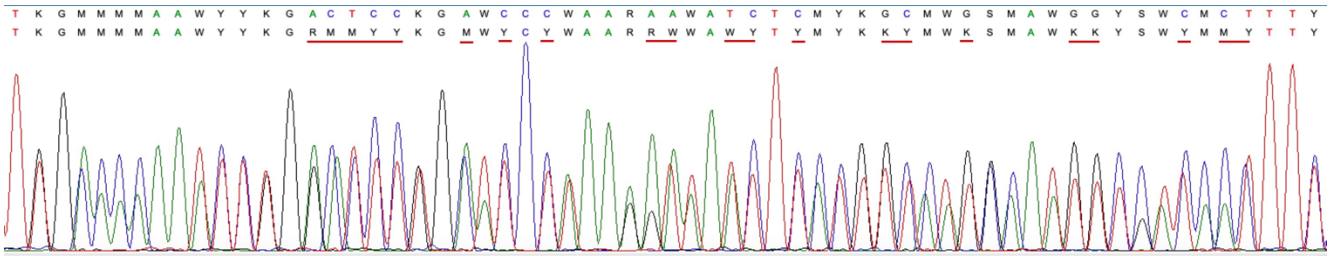

Figure 1 A portion of DNA sequence chromatogram showing mixed bases in a heterozygote for *AsteObp1* intron-1. The first line of basecall was assigned by the ABI Sequence Analysis software. The second line of basecall is manually edited (underlined) where software failed to identify mixed bases at some nucleotide positions.

```
1.    H1+H2     GTGAGYTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTCTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCMTGYMATGYMATSWCWTYMCWKWMSMG
2.    H1+H3     GTGAGYTTGGGTGTCTTCTGGATATTRTTCTAATGTGTTTTCTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCCTGTCATGCAATGTCATCACTTTCCAG
3.    H1+H4     GTGAGCTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTCTGTCTATAAGTTTTWRACACAAATCTGGACTCTGCATCCTAAAGATAATKTCCTGTCATGCAATGTCATCACTTTCCAG
4.    H1+H6     GTGAGCTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTCTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKRMWYCYWAARRWWAWYTYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
5.    H1+H8     GTGAGCTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTYWSTCTAYWWGYTTTTGACASAAATCTGGAYYCTRMAKMYTWWMKAWWRTYWYSYRWYRTSMWMWSTYWYCASWWKCSAK
6.    H1+H9     GTGAGCTTGGGTGTCWTCTGGATATTGWTCTAATGTGTTTTYWSTCYMYWWGTTTTTGACASAAATCTGGAYYCTRMAKMYTWWMKAWWRTYWYSYRWYRTSMWMWSTYWYCASWWKCSAK
7.    H1+H10    GTGAGYTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTCTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCCTGTCATGCAATGTCATCACTTTCCAG
8.    H1+H11    GTGAGCTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTCTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKGMWYCYWAARRWWAWYTYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
9.    H1+H12    GTGAGYTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTYTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCMTGYMATGYMATSWCWTYMCWKWMSMG
10.   H1+H13    GTGAGCTWRKGTGTYTTYWSKMYAYTKKYYTWWKRYRKWWWYYKKSWWYMWRWWKWTKWYMYAAWKYYRKRCWMTGYMWYCWMWWKMYARWWKCSWKYMAWRMRWTSWSMKMMSWWKMSAK
11.   H1+H14    GTGAGCTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTYWSTCTAYWWGYTTTTGACASAAATCTGRAYTCTGMATSCTAAAKATAATYTCCTGTCATGCAATGTCATCACTTTCCAG
12.   H2+H3     GTGAGTTTGGGTGTCTTCTGGATATTRTTCTAATGTGTTTTCTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCMTGYMATGYMATSWCWTYMCWKWMSMG
13.   H2+H4     GTGAGYTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTCTGTCTATAAGTTTTWRACACAAATCTGGACTCTGCATCCTAAAGATAATKTCMTGYMATGYMATSWCWTYMCWKWMSMG
14.   H2+H6     GTGAGYTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTCTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKRMWYCYWAARRWWAWYTYMWKKCAWKSWMWKYMSTYWCYWKMMRSR
15.   H2+H8     GTGAGYTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTYWSTCTAYWWGYTTTTGACASAAATCTGRAYTCTGMATSYTAAAKATAATYTCMTGYMATGYMATSWCWTYMCWKWMSMG
16.   H8+H10    GTGAGYTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTYWSTCTAYWWGYTTTTGACASAAATCTGRAYTCTGMATSYTAAAKATAATYTCCTGTCATGCAATGTCATCACTTTCCAG
17.   H2+H9     GTGAGYTTGGGTGTCWTCTGGATATTGWTCTAATGTGTTTTYWSTCYMYWWGTTTTTGACASAAATCTGGAYYCTRMAKMYTWWMKAWWRTYWYRYRMWRTSWYMWCWYTYYMSARRMGMK
18.   H2+H10    GTGAGTTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTCTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCMTGYMATGYMATSWCWTYMCWKWMSMG
19.   H2+H11    GTGAGYTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTCTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKGMWYCYWAARRWWAWYTYMWKKCAWKSWMWKYMSTYWCYWKMMRSR
20.   H2+H12    GTGAGTTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTYTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCATGCAATGTCATCACTTTCCAGAAGCG
21.   H2+H13    GTGAGYTWRKGTGTYTTYWSKMYAYTKKYYTWWKRYRKWWWYYKKSWWYMWRWWKWTKWYMYAAWKYYRKRCWMTGYMWYCWMWWKMYARWWKCRWKCAAWRWSWTCASYKWCSARRAGMK
22.   H2+H14    GTGAGYTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTYWSTCTAYWWGYTTTTGACASAAATCTGRAYTCTGMATSCTAAAKATAATYTCMTGYMATGYMATSWCWTYMCWKWMSMG
23.   H3+H4     GTGAGYTTGGGTGTCTTCTGGATATTRTTCTAATGTGTTTTCTGTCTATAAGTTTTWRACACAAATCTGGACTCTGCATCCTAAAGATAATKTCCTGTCATGCAATGTCATCACTTTCCAG
24.   H3+H6     GTGAGYTTGGGTGTCTTCTGGATATTRYYCWAATKTGTTTTCTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKRMWYCYWAARRWWAWYTYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
25.   H3+H8     GTGAGYTTGGGTGTCTTCTGGATATTRYTCTAATGTGTTTTYWSTCTAYWWGYTTTTGACASAAATCTGRAYTCTGMATSYTAAAKATAATYTCCTGTCATGCAATGTCATCACTTTCCAG
26.   H3+H9     GTGAGYTTGGGTGTCWTCTGGATATTRWTCTAATGTGTTTTYWSTCYMYWWGTTTTTGACASAAATCTGGAYYCTRMAKMYTWWMKAWWRTYWYSYRWYRTSMWMWSTYWYCASWWKCSAK
27.   H3+H10    GTGAGTTTGGGTGTCTTCTGGATATTRTTCTAATGTGTTTTCTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCCTGTCATGCAATGTCATCACTTTCCAG
28.   H3+H11    GTGAGYTTGGGTGTCTTCTGGATATTRYYCWAATKTGTTTTCTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKGMWYCYWAARRWWAWYTYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
29.   H3+H12    GTGAGTTTGGGTGTCTTCTGGATATTRTTCTAATGTGTTTTYTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCMTGYMATGYMATSWCWTYMCWKWMSMG
30.   H3+H13    GTGAGYTWRKGTGTYTTYWSKMYAYTWKYYTWWKRYRKWWWYYKKSWWYMWRWWKWTKWYMYAAWKYYRKRCWMTGYMWYCWMWWKMYARWWKCSWKYMAWRMRWTSWSMKMMSWWKMSAK
31.   H3+H14    GTGAGYTTGGGTGTCTTCTGGATATTRYTCTAATGTGTTTTYWSTCTAYWWGYTTTTGACASAAATCTGRAYTCTGMATSCTAAAKATAATYTCCTGTCATGCAATGTCATCACTTTCCAG
32.   H4+H6     GTGAGCTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTCTGTCTATAAKTTTTRRMMMMAAWYYKGRMYYYKRMWYCYWAARRWWAWYKYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
33.   H4+H8     GTGAGCTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTYWSTCTAYWWGYTTTWRACASAAATCTGRAYTCTGMATSYTAAAKATAATSTCCTGTCATGCAATGTCATCACTTTCCAG
34.   H4+H9     GTGAGCTTGGGTGTCWTCTGGATATTGWTCTAATGTGTTTTYWSTCYMYWWGTTTTWRACASAAATCTGGAYYCTRMAKMYTWWMKAWWRTSWYSYRWYRTSMWMWSTYWYCASWWKCSAK
35.   H4+H10    GTGAGYTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTCTGTCTATAAGTTTTWRACACAAATCTGGACTCTGCATCCTAAAGATAATKTCCTGTCATGCAATGTCATCACTTTCCAG
36.   H4+H11    GTGAGCTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTCTGTCTATAAKTTTTRRMMMMAAWYYKGRMYYYKGMWYCYWAARRWWAWYKYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
37.   H4+H12    GTGAGYTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTYTGTCTATAAGTTTTWRACACAAATCTGGACTCTGCATCCTAAAGATAATKTCMTGYMATGYMATSWCWTYMCWKWMSMG
38.   H4+H13    GTGAGCTWRKGTGTYTTYWSKMYAYTKKYYTWWKRYRKWWWYYKKSWWYMWRWWKWWWWYMYAAWKYYRKRCWMTGYMWYCWMWWKMYARWRKCSWKYMAWRMRWTSWSMKMMSWWKMSAK
39.   H4+H14    GTGAGCTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTYWSTCTAYWWGYTTTWRACASAAATCTGRAYTCTGMATSCTAAAKATAATSTCCTGTCATGCAATGTCATCACTTTCCAG
40.   H6+H8     GTGAGCTTGGGTGTCTTCTGGATATTGCYCWAATKTGTTTTYWSTCTAYWWKYTTTKGMMMRAAWYYKGAMTYYKRAWYSTWAARWWWAWYYYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
41.   H6+H9     GTGAGCTTGGGTGTCWTCTGGATATTGMYCWAATKTGTTTTYWSTCYMYWWKTTTTKGMMMRAAWYYKGRMTCYKAAWSMTWWWSWWAWKYYMWKSWMWKKCAWYRYYKTCMMKWWSCRRW
42.   H6+H10    GTGAGYTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTCTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKRMWYCYWAARRWWAWYTYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
43.   H6+H11    GTGAGCTTGGGTGTCTTCTGGATATTGCCCAAATTTGTTTTCTGTCTATAATTTTTGGCACAAATCTGGACTCTGRATCCTAAAGATAATCTCATGTCATGCAATGTCGTCACTTTCCAGA
44.   H6+H12    GTGAGYTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTYTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKRMWYCYWAARRWWAWYTYMWKKCAWKSWMWKYMSTYWCYWKMMRSR
45.   H6+H13    GTGAGCTWRKGTGTYTTYWSKMYAYTKSCYWWWKWYRKWWWYYKKSWWYMWWWWWKWKKYWCWAATSTSRWSYMWKRWYMYYAMWKWYMAKMWSMKRTCAWRMARTKYMGYSACKWWSMRRW
46.   H6+H14    GTGAGCTTGGGTGTCTTCTGGATATTGCYCWAATKTGTTTTYWSTCTAYWWKYTTTKGMMMRAAWYYKGAMTYYKRAWYSYWAARWWWAWYYYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
47.   H8+H9     GTGAGCTTGGGTGTCWTCTGGATATTGMTCTAATGTGTTTTTACTCYMCTTGYTTTTGACAGAAATCTGRATYCTRAAKRTTWWMTAWWRTCWYSYRWYRTSMWMWSTYWYCASWWKCSAK
48.   H8+H11    GTGAGCTTGGGTGTCTTCTGGATATTGCYCWAATKTGTTTTYWSTCTAYWWKYTTTKGMMMRAAWYYKGAMTYYKGAWYSTWAARWWWAWYYYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
49.   H8+H12    GTGAGYTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTTWSTCTAYWWGYTTTTGACASAAATCTGRAYTCTGMATSYTAAAKATAATYTCMTGYMATGYMATSWCWTYMCWKWMSMG
50.   H8+H13    GTGAGCTWRKGTGTYTTYWSKMYAYTKSYYTWWKRYRKWWWTMYKSWWCYTRMWKWTKWYMKAAWKYYRWRYWMTGWMWKYWMWWTMYARWMKCSWKYMAWRMRWTSWSMKMMSWWKMSAK
51.   H8+H14    GTGAGCTTGGGTGTCTTCTGGATATTGCTCTAATGTGTTTTTACTCTACTTGCTTTTGACAGAAATCTGAATTCTGAATGYTAAATATAATCTCCTGTCATGCAATGTCATCACTTTCCAG
52.   H9+H10    GTGAGYTTGGGTGTCWTCTGGATATTGWTCTAATGTGTTTTYWSTCYMYWWGTTTTTGACASAAATCTGGAYYCTRMAKMYTWWMKAWWRTYWYSYRWYRTSMWMWSTYWYCASWWKCSAK
53.   H9+H11    GTGAGCTTGGGTGTCWTCTGGATATTGMYCWAATKTGTTTTKGMMMRAAWYYKGRMTCYKRAWSMTWWWSWWAWKYYMWKSWMWKKCAWYRYYKTCMMKWWSCRRW
54.   H9+H12    GTGAGYTTGGGTGTCWTCTGGATATTGWTCTAATGTGTTTTTWSTCYMYWWGTTTTTGACASAAATCTGGAYYCTRMAKMYTWWMKAWWRTYWYRYRMWRTSWYMWCWYTYYMSARRMGMK
55.   H9+H13    GTGAGCTWRKGTGTYWTYWSKMYAYTKRYYTWWKRYRKWWWTMYKSMYCYTRWWKWTKWYMKAAWKYYRKRYMMTRWMRWYWYTYTMMWGWMRYGMWMWRWMAKYWCWKYSMMGAAGMGAT
56.   H9+H14    GTGAGCTTGGGTGTCWTCTGGATATTGMTCTAATGTGTTTTTACTCYMCTTGYTTTTGACAGAAATCTGRATYCTRAAKRYTWWMTAWWRTCWYSYRWYRTSMWMWSTYWYCASWWKCSAK
57.   H10+H11   GTGAGYTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTCTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKGMWYCYWAARRWWAWYTYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
58.   H10+H12   GTGAGTTTGGGTGTCTTCTGGATATTGTTCTAATGTGTTTTYTGTCTATAAGTTTTTGACACAAATCTGGACTCTGCATCCTAAAGATAATTTCMTGYMATGYMATSWCWTYMCWKWMSMG
59.   H10+H13   GTGAGYTWRKGTGTYTTYWSKMYAYTKKYYTWWKRYRKWWWYYKKSWWYMWRWWKWTKWYMYAAWKYYRKRCWMTGYMWYCWMWWKMYARWWKCSWKYMAWRMRWTSWSMKMMSWWKMSAK
60.   H10+H14   GTGAGYTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTYWSTCTAYWWGYTTTTGACASAAATCTGRAYTCTGMATSCTAAAKATAATYTCCTGTCATGCAATGTCATCACTTTCCAG
61.   H11+H12   GTGAGYTTGGGTGTCTTCTGGATATTGYYCWAATKTGTTTTYTGTCTATAAKTTTTKGMMMMAAWYYKGRMYYYKGMWYCYWAARRWWAWYTYMWKKCAWKSWMWKYMSTYWCYWKMMRSR
62.   H11+H13   GTGAGCTWRKGTGTYTTYWSKMYAYTKSCYWWWKWYRKWWWYYKKSWWYMWWWWWKWKKYWCWAATSTSRWSYMWKGWYMYYAMWKWYMAKMWSMKRTCAWRMARTKYMGYSACKWWSMRRW
63.   H11+H14   GTGAGCTTGGGTGTCTTCTGGATATTGCYCWAATKTGTTTTYWSTCTAYWWKYTTTKGMMMRAAWYYKGAMTYYKGAWYSYWAARWWWAWYYYMYKKYMWKSMAWKKYSWYMMYTTYCMRR
64.   H12+H13   GTGAGYTWRKGTGTYTTYWSKMYAYTKKYYTWWKRYRKWWWTYKKSWWYMWRWWKWTKWYMYAAWKYYRKRCWMTGYMWYCWMWWKMYARWWKCRWKCAAWRWSWTCASYKWCSARRAGMK
65.   H12+H14   GTGAGYTTGGGTGTCTTCTGGATATTGYTCTAATGTGTTTTTWSTCTAYWWGYTTTTGACASAAATCTGRAYTCTGMATSCTAAAKATAATYTCMTGYMATGYMATSWCWTYMCWKWMSMG
66.   H13+H14   GTGAGCTWRKGTGTYTTYWSKMYAYTKSYYTWWKRYRKWWWTMYKSWWCYTRMWKWTKWYMKAAWKYYRWRYWMTGWMWKCWMWWTMYARWMKCSWKYMAWRMRWTSWSMKMMSWWKMSAK
```

Figure 2 Deduced sequence pattern of haplotype combinations in the DNA sequence chromatogram (forward sequence)

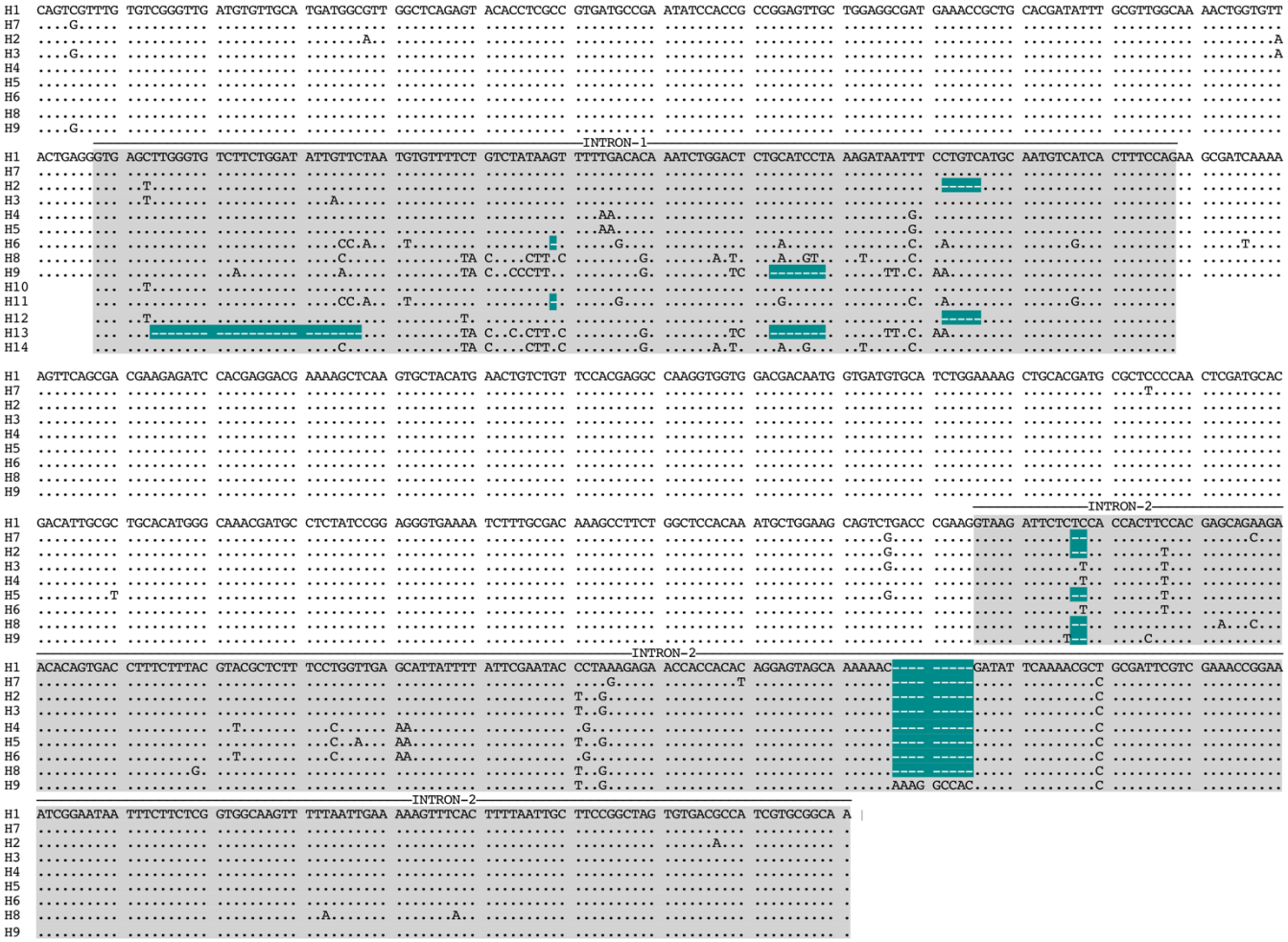

Figure 3 Sequence alignment of *AstObp1* haplotypes identified

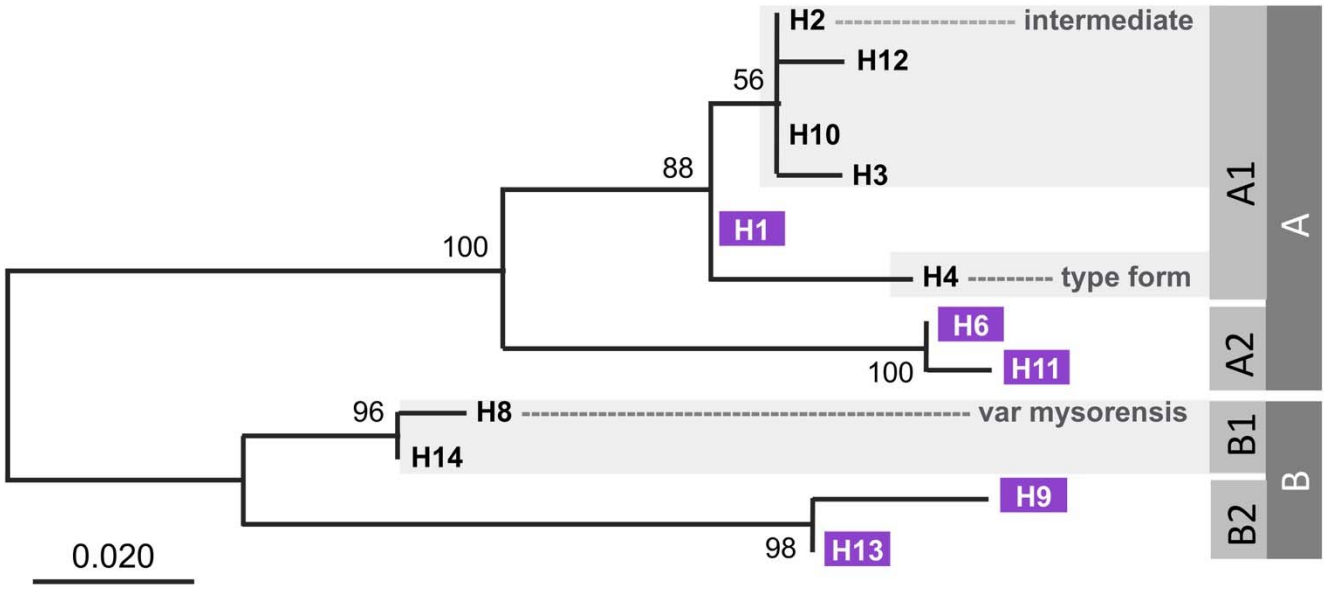

Figure 4. Maximum-Likelihood phylogenetic tree inferred from *AsteObp1* intron-1 sequence. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The specific haplotype that has been designated as a molecular marker for specific biological form by Gholizadeh et al. (2015) and related haplotypes grouped under a monophyletic clade have been highlighted with grey colour. Orphan haplotypes have been labelled with highlighted text (purple colour).

Appendix: S1

Expected heterozygosity (H_E) for *n* alleles

$H_E = 1-\sum_{i=1}^{n}(pi)^2$

(Where *p*i = frequency of *i*th allele of *n* alleles)

Maximum expected heterozygosity ($H_{E\ Max}$) will be when alleles are equally frequent (i.e., $pi = 1/n$) in a population

Then,

$H_{E\ Max} = 1-(n(1/n)^2)$

$= 1-(1/n)$