Applied Biological Research 17(2): 132-138; (2015)
DOI: 10.5958/0974-4517.2015.00020.8

PHYLOGENY AND *in-silico* MINING OF SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) IN CYTOCHROME OXIDASE I GENE OF INDIAN *Helicoverpa armigera* (NOCTUIDAE: LEPIDOPTERA) POPULATIONS

A.R.N.S. Subbanna, B. Kalyana Babu*and M.S. Khan

Department of Entomology, College of Agriculture, GB Pant University of Agriculture and Technology, Pantnagar, Uttarakhand - 263 601 (India)
**ICAR-Vivekananda Parvatiya Krishi Anusandhan Sansthan (VPKAS), Almora, Uttarakhand (India)*
*e-mail: *subbanna.ento@gmail.com*

(Received 20 February, 2015; accepted 20 April, 2015)

ABSTRACT

Single nucleotide polymorphisms (SNPs) represent stable, bi-allelic nucleotide variations that are distributed throughout the genome and are the most abundant genetic variation. The present study targeted SNP related molecular differentiation and phylogenetic relationship of Indian *Helicoverpa armigera* populations by using expressed sequence tags (ESTs) of partial cytochrome oxidase subunit I (COI). The phylogenetic evaluation clearly separated the northern and southern Indian populations of *H. armigera* with little admixtures of central Indian populations. The SNP mining also revealed two major clusters as that of phylogeny with a total of 11 potential SNPs, of which 10 were in reliable haplotypes. The potential SNPs were found to contain eight transitions, three transversions and no indels (insertions and deletions). Out of 38 sequences, 13 were haplotypes in which 4 were single haplotypes. In cluster 1, the reliable bi-allelic SNPs were observed at 97, 166 and 265 nucleotide positions; while they were at 411, 423, 522, 655 and 656 positions in cluster 2. The present study highlighted the genetic polymorphism and relationship of Indian *H. armigera,* besides confirming the long distance migration and gene flow between the geographic populations.

Key words: Cytochrome oxidase I, *Helicoverpa armigera*, Indian populations, phylogeny, SNP mining

INTRODUCTION

During past few decades numerous studies have been carried out to explore genomes of various vital organisms by detecting the most efficient genetic markers. Recently, single nucleotide polymorphisms (SNPs) have been found to be the most efficient genetic markers for gene identification (Singhal *et al*., 2011). They are co-dominant, bi-allelic, highly polymorphic and have good reproducibility with low mutation rates. As biological markers the SNPs have been found useful in gene mapping, gene identification and drug development, etc. (Botstein and Risch, 2003). Their simplicity, ease of modeling and sheer abundance (Robb *et al*., 2003) make them the marker of choice for many applications in population ecology, evolution and conservation genetics. SNP mapping and correlative analysis are being conducted on an increasing number of organisms, but insufficient attention has been paid to insects, apart from *Drosophila* (Berger *et al*., 2001). Now a days, a large amount of expressed sequence tags (EST) and nucleotide data is being stored and updated for various organisms in databases like NCBI (National Center for Biotechnology Information). With the availability of various bioinformatics tools like Polybayes, SNPhunter (Xiang *et al*., 2009), and Haplo-SNPer (Tang *et al*., 2006), it has become easier to detect SNPs in a given genome. These *in-silico* tools detect SNPs using various publically available EST databases. The detected SNPs have been very useful in polymorphism studies, functional genomics, pharmacogenetics studies and agronomic studies (Singhal *et al*., 2011).

Helicoverpa armigera (Hubner), commonly known as cotton bollworm or American bollworm, is a major polyphagous (more than 180 plant hosts from more than 45 families) and cosmopolitan pest of global importance. In India, the cropping patterns provide a range of host crops to this pest, round the year, in any given ecological region. Cotton represents the main host crop on which this pest species completes three out of possible seven to eight generations annually (Behere *et al*., 2013). The highly variable traits of *H. armigera* such as life-history parameters (e.g. number of generations, wide host range, presence of summer/winter diapauses, etc.) and seasonal abundance in association with host plant and geographical location gives a unique challenge for population genetic structure and evolutionary studies of this pest species. Previous studies categorized the Indian *H. armigera* population based on host feeding preferences, inter-mating features (Bhattacherjee, 1972; Reed and Pawar, 1982), metabolic mechanisms mediating pyrethroid resistance (Kranthi *et al.*, 1997), DNA markers such as random amplified polymorphic DNA (RAPD) (Zhou *et al*., 2000), isozymes (Nibouche *et al*., 1998), mtDNA (Behere *et al*., 2007; Tay *et al*., 2013) and microsatellites (Vassal *et al*., 2008; Endersby *et al*., 2007). These studies found little genetic variation between widely separated populations, supporting the idea that extensive long distance migration was occurring in *H. armigera.*

In recent past, the taxonomic, population and evolutionary investigations in animals was dominated by the analysis of mitochondrial genes. Among these, the mitochondrial gene encoding subunit I of cytochrome oxidase (COI) possesses some excellent characteristics which make it particularly suitable as a molecular marker for evolutionary studies (Lunt *et al*., 1996). It was the most studied region of insect mitochondrial genome (Kranthi *et al.*, 2006). However, the evolutionary studies of COI can be enriched with the availability of SNP markers since they provide high amount of polymorphism. This high SNP polymorphism could prove very useful for molecular differentiation of *H. armigera* populations even from the same region and also to disclose the evolutionary relationship. So, the present study was aimed at *in-silico* identification of SNPs from published EST sequences of COI region in Indian origin *H. armigera* populations and their phylogenetic relationship by combining the two molecular markers.

MATERIALS AND METHODS

Retrieval of sequence data

The available partial COI sequences of *H. armigera* were retrieved in FASTA format from FTP site of the National Center for Biotechnology Information (NCBI), during December 2014 to January, 2015. The Indian populations were retrieved by key word search and were confirmed manually. A total of 38 sequences were retrieved from the database and details were presented in Table 1. The sequences were categorized into different groups based on geographical locations of India: South India (SI), North India (NI), Central India (CI) and North East India (NEI).

Data analysis

Alignment of sequences: The multiple alignments of selected sequences were done by using Clustal W software (Tamura *et al.,* 2007). The phylogenetic analyses and minimum evolution tree was prepared by using the software MEGA 4 (Molecular Evolutionary Genetic Analysis version 4) (Kumar *et al*., 2004).

Table 1: Details of the sequences used in the study

Accession No.	Collection site	State or affiliation
AY264944, DQ084770, EF432737	Anonymous	CICR, Nagpur
DQ084765	Mehbubabad	Telangana, SI
DQ084766	Sirsa	Haryana, NI
DQ084767	Nagpur	Maharastra, CI
DQ084768, DQ084773	Amravati	Andhra Pradesh, SI
DQ084769	Fatehabad	Haryana, NI
DQ084771	Dharwad	Karnataka, SI
DQ084772	Mansa	Punjab, NI
DQ084774	Yavatmal	Maharashtra, CI
DQ084781, EF432736	Guntur	Andhra Pradesh, SI
FN908003, FN908013, FN908016	Anonymous	FERA, UK
HM854928	Rajkot	Gujarat, CI
HM854929	Akola	Maharastra, CI
HM854930	Anand	Gujarat, CI
HM854931	Aurangabad	Maharastra, CI
HM854932	Surendra Nagar	Gujarat, CI
JF776377, KC911713, KM226881,KM459450	Anonymous	NBAII, Bangalore
JX532104	Umiam	Meghalaya, NEI
KJ940177, KJ940178, KJ940184, KJ940185	Malerkotla	Punjab, NI
KJ940176, KJ940179, KJ940180, KJ940181, KJ940182	Ludhiana	Punjab, NI
KJ940183	Gurdaspur	Punjab, NI
KM403206	Vellanikkara	Kerala, SI

NI= North India SI= South India CI= Central India NEI=North east India

SNP identification: The SNP identification was done using the online tool HaploSNPer **(**Tang *et al*., 2006**)** by using the settings of parameters. The input seed sequence area was filled with one FASTA format sequence of the 38 sequences retrieved from NCBI database. Here, there were two options where we can give the known similar sequences or the publicly available database. Since, we already retrieved the known EST sequences of *H. armigera* COI region; we gave all the sequences as similar sequences. The other parameters were also selected. 1) For alignment, PHRAP and CAP3 were the provided options, however, in the present study PHRAP was used for sequence alignment. The CAP3 uses individual sequence overlap for constructing clusters, while PHRAP tends to extend the consensus sequence by overlap. 2) The pre-processing of sequences was done using the repeat masker option available. 3) For BLAST analysis, an E-value of 1e-60 was used and for CAP3 analysis, a minimum of 95% was taken as criteria. It could prevent most paralogous sequences and keep all available allelic sequences in a cluster. 4) The default settings were used for haplotype reconstruction. The threshold value of similarity per polymorphic site was taken as 75% and that of similarity over all polymorphic sites was taken as 80%. 5) Settings for low quality regions: The data used in the study were EST sequences. Thus, the default settings like removal of 20 and 30 nucleotides from 5’ and 3’ side, respectively, was used. 6) The sequence redundancy is also used to prevent sequencing errors. In the current study, the values for minimum cluster size, minimum allele size and minimum confidence score were 4, 2 and 2, respectively. The higher confidence score was the reliability of the SNP on sequence redundancy.

RESULTS AND DISCUSSION

Phylogeny

The sequences were first aligned to determine the conserved regions using ClustalW software in order to estimate the phylogenetic relationship between partial sequences of *H. armigera* COI regions belonging to different geographic regions of India. The results showed that 630 bp region had similarity across all the selected sequences which expanded between 1475 to 2105 bp regions of whole mitochondrial genome of *H. armigera* (Accession No. NC 014668.1). The phylogentic relationships were determined using MEGA 4 software which grouped all the sequences into two major clusters (cluster 1 and 2) (Fig. 1). The clustering pattern showed that grouping was mainly according to the geographical origin or place of collection. The cluster 1 consisted of 14 sequences dominated by south Indian populations, while cluster 2 comprised of 24 sequences dominated by south Indian populations. Both the clusters also contained central Indian populations. This genetic structure of Indian *H. armigera* populations showed gene flow patterns between host crop, temporal and spatial levels. The soaring host range (Singh and Singh, 1996; Razmjou *et al*., 2014), fecundity, mobility (Behere *et al*., 2013) and development of resistance (Kranthi *et al*., 1997) were the main reasons for this gene flow. Pedgley *et al*. (1987) reported windborne long-distance migration of *H. armigera* in central India at the end of cropping season (December- January), while rains prolonged the growing season in northern and southern India, resulting in adult migration in these regions around March-April. Nested alternative EPIC markers (RpL3, RpL12, RpL29, RpS6 and RpS2) also detected moderate null allele frequencies (4.3 to 9.4%) in Indian *H. armigera* populations but the apparently genome-wide heterozygote deficit suggested in-breeding or a Wahlund effect rather than a null allele effect (Behere *et al*., 2013). Furthermore, the maternally inherited mitochondrial genome transmits any change to the entire progeny ensuring rapid spread of evolutionary changes. Such changes which can be micro-evolutionary in nature can be a function of selection pressure induced by both biotic and abiotic stresses (Kranthi *et al*., 2006).

Identification of SNPs

SNPs including insertion/deletion (indels) serve as effective genetic markers. Computational strategies for SNP discovery make use of a large number of sequences present in public databases, in most cases as expressed sequence tags (ESTs) and are considered to be faster and more cost-effective than experimental procedures (Tang *et al*., 2006). In present study, SNP mining was donein partial COI sequences of Indian *H. armigera* populations by using HaploSNPer tool. The clustering pattern by MEGA 4 and HaploSNPer software were similar, where both generated two major clusters. However, out of 38 sequences used, HaploSNPer cluster 1 consisted of 13 EST sequences, while cluster 2 had 25 sequences. The SNP mining revealed a total of 11 potential SNPs. Cluster 1 had 5 potential SNPs, while cluster 2 consisted of 6 potential SNPs. The potential SNPs were defined by minimum size of each allele and included bi-allelic, tri-allelic, tetra-allelic, and penta-allelic SNPs. The five potential SNPs of cluster 1 were in reliable haplotypes, of which three SNPs were reliable. Haplotype is a group of sequences within a cluster that represent the same allele of a gene, whereas the reliable haplotype is defined as a haplotype containing at least 2 sequences. The cluster 2 comprised of 5 potential SNPs in reliable haplotypes. There were a total of 7 haplotypes in cluster 1, of which 4 were single haplotypes which contained only single sequence. SNPs in partial COI sequences of *H. armigera* differentiated the populations from Australia, Burkina Faso, Uganda, China, India and Pakistan into 33 mtDNA haplotypes (Behere *et al*., 2007).

The analysis of potential and reliable SNPs revealed that there were no indels (nucleotide insertions and deletions). The potential SNPs were found to contain 8 transitions and 3 transversion. The C/T and A/G transitions were found to be 4. Among the transversion type of SNPs, 2 were A/Transversion and 1 was T/G transversion. However, there were no A/C and C/G transversions. Among the reliable SNPs, 6 were transitions and 2 were transversions. The details of potential SNPs with their nucleotide locations, SNP type, confidence score are given in Table 2. In cluster 1, the identified 5 SNPs were at nucleotide positions 67, 97, 166, 265 and 472. Among them, reliable and bi-alleleic SNPs were at 3 positions (97, 166 and 265). The major allele (A) was found in 8 sequences, while minor allele (G) was in 4 sequences. Three haplotypes were found to contain major allele at nucleotide position 67. However, in the remaining positions, two haplotypes each consisted of major alleles. In case of cluster 2, all the 5 identified SNPs were variable between geno-

Table 2: Details of potential SNPs in partial COI regions of *H. armigera* Indian populations

Parameters	Cluster 1					Cluster 2				
Nucleotide location	67	97	166	265	472	411	423	522	655	656
Major allele	C	A	G	C	G	C	T	G	A	G
Sequences with major allele (No.)	9	8	9	10	10	15	20	12	20	20
Minor allele	T	G	A	T	A	T	C	A	T	T
Sequences with minor allele (No.)	3	4	3	2	2	7	2	10	2	2
Between/within genotypes	1	1	1	1	1	2	2	2	2	2
SNP type	-1	2	2	2	-1	2	2	2	-1	-1
Haplotypes with major allele (No.)	3	2	2	2	2	3	3	3	2	2
Haplotypes with minor allele (No.)	0	1	1	1	0	1	1	1	0	0

Between/within genotypes: 1=Variations within one genotype, 2=Variation between genotypes; SNP type: -1 = Unreliable SNPs, 2 = Reliable bi-allelic SNPs

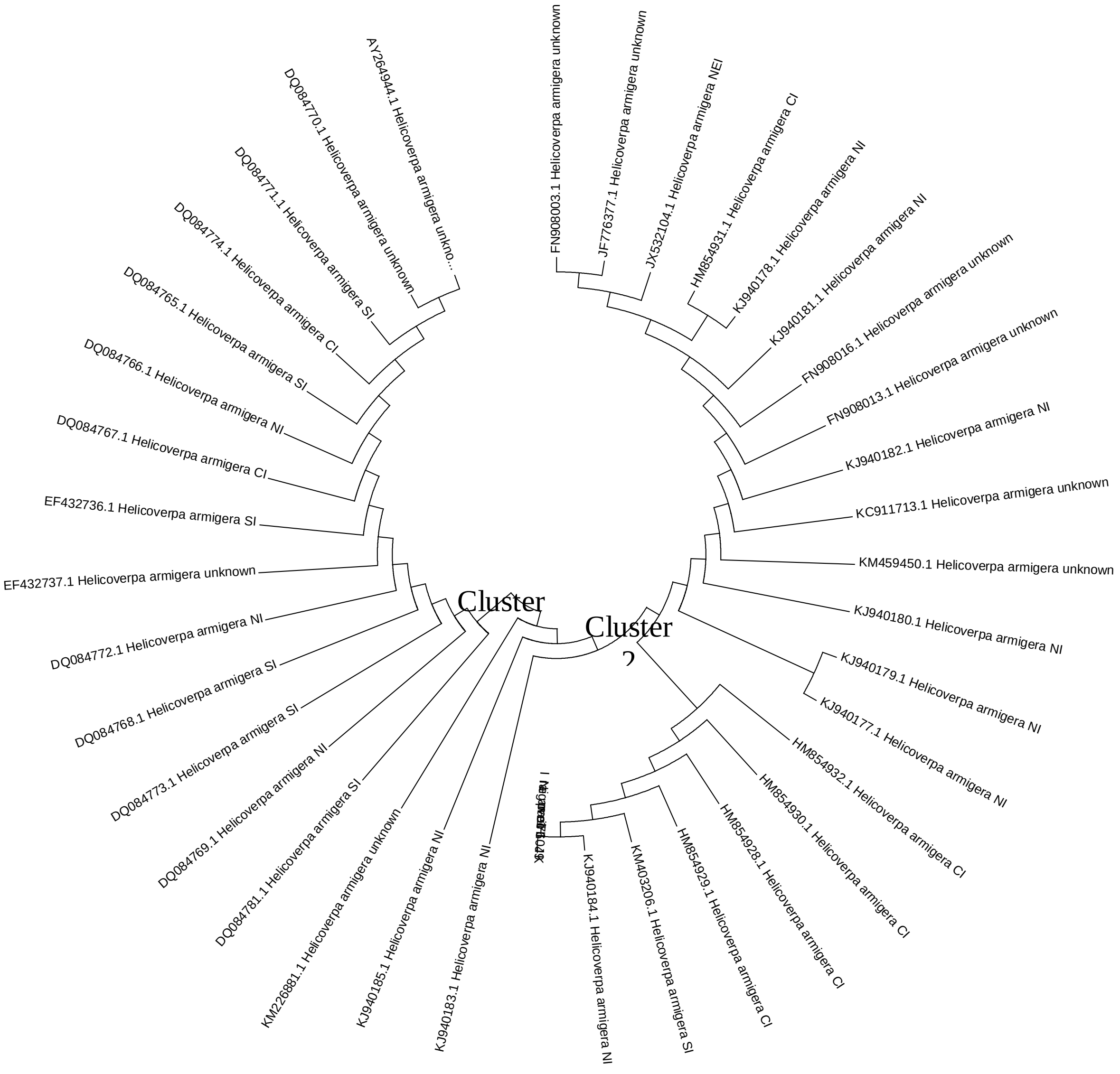
AY264944.1 Helicoverpa armigera unkno...
DQ084770.1 Helicoverpa armigera unknown
DQ084771.1 Helicoverpa armigera SI
DQ084774.1 Helicoverpa armigera CI
DQ084765.1 Helicoverpa armigera SI
DQ084766.1 Helicoverpa armigera NI
DQ084767.1 Helicoverpa armigera CI
EF432736.1 Helicoverpa armigera SI
EF432737.1 Helicoverpa armigera unknown
DQ084772.1 Helicoverpa armigera NI
DQ084768.1 Helicoverpa armigera SI
DQ084773.1 Helicoverpa armigera SI
DQ084769.1 Helicoverpa armigera NI
DQ084781.1 Helicoverpa armigera SI
KM226881.1 Helicoverpa armigera unknown
KJ940185.1 Helicoverpa armigera NI
KJ940183.1 Helicoverpa armigera NI
KJ940184.1 Helicoverpa armigera NI
KM403206.1 Helicoverpa armigera SI
HM854929.1 Helicoverpa armigera CI
HM854928.1 Helicoverpa armigera CI
HM854930.1 Helicoverpa armigera CI
HM854932.1 Helicoverpa armigera CI
KJ940177.1 Helicoverpa armigera NI
KJ940179.1 Helicoverpa armigera NI
KJ940180.1 Helicoverpa armigera NI
KM459450.1 Helicoverpa armigera unknown
KC911713.1 Helicoverpa armigera unknown
KJ940182.1 Helicoverpa armigera NI
FN908013.1 Helicoverpa armigera unknown
FN908016.1 Helicoverpa armigera unknown
KJ940181.1 Helicoverpa armigera NI
KJ940178.1 Helicoverpa armigera NI
HM854931.1 Helicoverpa armigera CI
JX532104.1 Helicoverpa armigera NEI
JF776377.1 Helicoverpa armigera unknown
FN908003.1 Helicoverpa armigera unknown
Cluster
Cluster 2

Fig. 1: Minimum Evolution (ME) tree with bootstrap support (1000 replicates) showing clustering of Indian *H. armigera* populations for COI sequences.

types, of which 3 (at nucleotide positions 411, 423 and 522) were reliable and bi-allelic. The number of major allele haplotypes with unreliable SNPs was 2, each at positions 655 and 656.

The present study could prove helpful in detecting SNPs which can be applied not only for making genetic maps (in case of large data analysis) but also for exploring the astonishing features (genes with special features) of a genome sequence. Consistent base pair substitutions between the two Indian species of *Helicoverpa* (*H. armigera* and *H. assulta*) and specific restriction enzymes that can cleave at the point of SNP were used in the specific recognition of two species at their morphologically indistinguishable stages (Kranthi *et al*., 2006). Such features of SNP have recently brought a flurry of SNP discovery and detection. However, the EST sequences available for *H. armigera* COI sequences were less from India and require more EST database for discovery of more SNPs for their genomics applications like polymorphism detections and gene identification and evolutionary relationships. Besides, the high mobility of pest results in shared haplotypes low *F*-statistic values and low nucleotide diversity between countries (Behere *et al*., 2007). SNPs have the potential to place historical demography and speciation studies on a common molecular framework, which could be easily comparable to the decades of mtDNA work already undertaken. Despite intense agricultural interest in *H. armigera*, very little systematic research has been conducted to resolve the question about geographical variation and phylogeny, especially under Indian conditions. It is apparent that this research work can definitely contribute to the functional genomics, agricultural sciences and crop protection studies.

REFERENCES

Behere, G.T., Tay, W.T., Russell, D.A., Heckel, D.G., Appleton, B.R., Kranthi, K.R. and Batterham, P. 2007. Mitochondrial DNA analysis of field populations of *Helicoverpa armigera* (Lepidoptera: Noctuidae) and of its relationship to *H. zea*. *BMC Evolutionary Biology*, **7**: 117-126.

Behere, G.T., Tay, W.T., Russell, D.A., Kranthi, K.R. and Batterham, P. 2013. Population genetic structure of the cotton bollworm *Helicoverpa armigera* (Hubner) (Lepidoptera: Noctuidae) in India as inferred from EPIC-PCR DNA markers. *PLoS One*, **8**: e53448. doi:10.1371/journal.pone.0053448.

Berger, J., Suzuki, T., Senti, K.A., Stubbs, J., Schaffner, G. and Dickson, B.J. 2001. Genetic mapping with SNP markers in *Drosophila*. *Nature Genetics*, **29**: 475-481.

Bhattacherjee, N.S. 1972. *Heliothis armigera* (Hubner) a polytypic species. *Entomology Newsletter*, **2**: 3–4.

Botstein, D. and Risch, N. 2003. Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**: 228-237.

Endersby, N.M., Hoffmann, A.A., McKechnie, S.W. and Weeks, A.R. 2007. Is there genetic structure in populations of *Helicoverpa armigera* from Australia? *Entomologia Experimentalis et Applicata*, **122**: 253-263.

Kranthi, K.R., Armes, N.J., Rao, N.G.V., Sheo, R. and Sundarmurthy, V.T. 1997. Seasonal dynamics of metabolic mechanisms mediating pyrethroid resistance in *Helicoverpa armigera* in central India. *Pesticide Science*, **50**: 91-98.

Kranthi, S., Kranthi, K.R., Bharose, A.A., Syed, S.N., Dhawad, C.S., Wadaskar, R.M., Behere, G.T. and Patil, E.K. 2006. Cytochrome oxidase I sequence of *Helicoverpa* (Noctuidae: Lepidoptera) species in india- Its utility as a molecular tool. *Indian Journal of Biotechnology*, **5**: 195-199.

Kumar, S., Tamura, K. and Nei, M. 2004. MEGA4: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinformatics*, **5**: 150-163.

Lunt, D.H., Zhang, D.X., Szymura, J.M. and Hewitt, G.M. 1996. The insect cytochrome oxidase I gene: evolutionary patterns and conserved primers for phylogenetic studies. *Insect Molecular Biology*, **5**: 153-165.

Nibouche, S., Bues, R., Toubon, J.F. and Poitout, S. 1998. Allozyme polymorphism in the cotton bollworm *Helicoverpa armigera* (Lepidoptera: Noctuidae): Comparison of African and European populations. *Heredity* **80**: 438-445.

Pedgley, D.E., Tucker, M.R. and Pawar, C.S. 1987. Windborne migration of *Heliothis armigera* (Hubner) (Lepidoptera: Noctuidae) in India. *International Journal of Tropical Insect Science*, **8**: 599-604.

Razmjou, J., Naseri, B. and Hemati, S.A. 2014. Comparative performance of the cotton bollworm, *Helicoverpa armigera* (Hubner) (Lepidoptera: Noctuidae) on various host plants. *Journal of Pest Science*, **87**: 29-37.

Reed, W. and Pawar, C.S. 1982. *Heliothis*: A global problem. pp. 9-14. **In**: *Proceedings of the International Workshop on Heliothis Management*. 15-20 November, 1981 (ed. W. Reed), 15-20 November 1981, ICRISAT Center, Patancheru, A.P., India.

Robb, T., Brumfield, Beerli, P., Deborah, A., Nickerson and Edwards, S.V. 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution*, **18**: 249-256.

Singhal, D., Gupta, P., Sharma, P., Kashyap, N., Anand, S. and Sharma, H. 2011. *In-silico* single nucleotide polymorphisms (SNP) mining of *Sorghum bicolor* genome. *African Journal of Biotechnology*, **10**: 580-583.

Singh, O.P. and Singh, K.J. 1996. Outbreak of gram pod borer *Helicoverpa armigera* on winter soybean and its effects on the grain yield in Madhya Pradesh. *Annals of Entomology*, **14**: 23-25.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. 2007. MEGA 4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**: 1596–1599.

Tang, J., Vosman, B., Voorrips, R.E., van der Linden, C.G. and Leunissen, J.A.M. 2006. Quality SNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics*, **7**: p. 438 doi: .1186/1471-2105-7-438.

Tay, W.T., Soria, M.F., Walsh, T., Thomazoni, D., Silvie, P., Behere, G.T., Anderson, C. and Downes, S. 2013. A brave new world for an old world pest: *Helicoverpa armigera* (Lepidoptera: Noctuidae) in Brazil. *PLOS One*, **8**, e80134.

Vassal, J.M., Brevault, T., Achaleke, J. and Menozzi, P. 2008. Genetic structure of the polyphagous pest *Helicoverpa armigera* (Lepidoptera: Noctuidae) across the Sub Saharan cotton belt. *Communications in Agricultural and Applied Biological Sciences*, **73**: 433-437.

Xiang, W., Can, Y., Qiang, Y., Hong, X., Nelson, L. and Weichuan, Y. 2009. MegaSNPHunter: A learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. *BMC Bioinformatics*, **10**: p. 13. doi: 10.1186/1471-2105-10-13.

Zhou, X., Faktor, O., Applebaum, S.W. and Coll, M. 2000. Population structure of the pestiferous moth *Helicoverpa armigera* in the eastern Mediterranean using RAPD analysis. *Heredity*, **85**: 251-256.