Materials and Methods

*Anopheles stephensi* mosquitoes were collected from different regions of Madhya Pradesh (Central India) from March, 2011 to September, 2013, covering different geographic regions along latitude and longitudinal axis.

3.1. STUDY AREA

Madhya Pradesh is situated in the central part of India with an area of 308 thousand km² of which forest cover 76,429 km² (about 25% of the total land area). Madhya Pradesh (population 72.6 Million) along with other states like Orissa (population 42 Million), Jharkhand (population 33 Million), and Chhattisgarh (population 25.5 Million) contributes for more than 60% of reported (confirmed) malaria cases in India. According to National Vector Borne Diseases Control Program (NVBDCP) epidemiological data for 2010, these tribal states with a total population of 173.1 million (out of a total country population i.e., 1.21 billion) showed a persistent malaria transmission with high API (annual parasite incidence), slide positivity rate (SPR) and very high Pf% (Sharma, 2012). Madhya Pradesh alone accounts for 6% of the total population of the country but contributes about 8.6% to the total malaria cases. Malaria is complex in Madhya Pradesh because of vast tracts of forest with tribal settlement (20% of state population) (Singh *et al.*, 2004, Anon, 2007). The magnitude of the problem can be accessed from an estimate made in 1987 which revealed that 54 million individuals of various ethnic origins residing in forested area of India and accounting for 8% of the total population contributed 30% of total malaria cases, 60% of total falciparum cases and 50% of malaria deaths in the country (Sharma, 1996).

3.2. COLLECTION OF MOSQUITOES

*Anopheles stephensi* mosquitoes were collected from nine different geographic regions from Madhya Pradesh (Central India) to assess the exact population genetic structure of *An. stephensi* in Central India (Figure 3.1). Mosquitoes were collected from Gwalior (GWL), Sagar (SGR), Umaria (UMR), Anuppur (ANP), Indore (IND), Bhopal (BPL), Khandwa (KNW) and Jabalpur (JBL) and Ujjain (UJN). The latitudinal and longitudinal coordinates ranged from 21.48 N to 26.12 N and 75.47 E to 80.50 E (Table
Figure 3.1: Map showing collection sites of *An. stephensi* mosquitoes from Madhya Pradesh (Central India)

Table 3.1: Latitudinal and longitudinal coordinates of *An. stephensi* collection sites in Madhya Pradesh (Central India).

<table>
<thead>
<tr>
<th>Code</th>
<th>Collection site</th>
<th>Latitude/ longitude coordinates</th>
<th>Sample size</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWL</td>
<td>Gwalior</td>
<td>26°12'46.54&quot;N, 78°11'0.66&quot;E</td>
<td>30</td>
<td>Aug, 2011</td>
</tr>
<tr>
<td>BPL</td>
<td>Bhopal</td>
<td>23°14'43.40&quot;N, 77°24'47.75&quot;E</td>
<td>14</td>
<td>Sep, 2011</td>
</tr>
<tr>
<td>ANP</td>
<td>Anuppur</td>
<td>23°65.95&quot;N, 81°41'52.49&quot;E</td>
<td>24</td>
<td>April, 2012</td>
</tr>
<tr>
<td>KNW</td>
<td>Khandwa</td>
<td>21°48'58.10&quot;N, 76°21'4.05&quot;E</td>
<td>30</td>
<td>Aug, 2011</td>
</tr>
<tr>
<td>UJN</td>
<td>Ujjain</td>
<td>23°10'44.39&quot;N, 75°47'5.66&quot;E</td>
<td>30</td>
<td>March 2012</td>
</tr>
<tr>
<td>SGR</td>
<td>Sagar</td>
<td>23°49'51.66&quot;N, 78°44'41.16&quot;E</td>
<td>30</td>
<td>Aug, 2011</td>
</tr>
<tr>
<td>JBL</td>
<td>Jabalpur</td>
<td>23°9'25.42&quot;N, 79°56'59.53&quot;E</td>
<td>30</td>
<td>Sep, 2013</td>
</tr>
<tr>
<td>UMR</td>
<td>Umaria</td>
<td>23°31'31.13&quot;N, 80°50'28.18&quot;E</td>
<td>30</td>
<td>March, 2013</td>
</tr>
<tr>
<td>IND</td>
<td>Indore</td>
<td>22°42'48.53&quot;N, 75°1'16.34&quot;E</td>
<td>30</td>
<td>April, 2012</td>
</tr>
</tbody>
</table>
3.1). The geographic distance between different pairs of population is given in Table 3.2.

From each region, samples were pooled from two or three different sites with at least 5-6 km gap between them to avoid collection of siblings. Mosquitoes were collected using a hand held suction tube and torchlight from indoor resting sites, mostly from human dwellings and cattle sheds.

At some localities where no adult mosquitoes were found, larvae and pupae were collected from all potential breeding sites by using a standard mosquito dipper (350 ml). Preference was given to III and IV instars larvae and pupae, so that they can turn to adults within two-three days. Anopheles stephensi breeds in a wide range of both urban and rural habitats throughout its distribution region. In urban areas, this species breeds in all sources of water bodies, such as wells, cisterns, fountains, ornamental ponds and in water pools used for building constructions. Larvae can be collected from ponds, pools, stream margins, catch basins and seepage canals. It is found even in water with high salinity, sometimes reaching or even exceeding that of seawater. In rural areas, the breeding places are pools, streambeds, palm irrigation canals, at the margin of streams and rivers, seepages, and marshy areas with a gentle water flow (Hanafi-Bojd et al., 2012). The different collection sites in Madhya Pradesh (Central India) where adult mosquitoes and larvae were collected are shown in Figure 3.2. Intact mosquitoes were morphologically identified (Nagpal et al., 2003) and were preserved in 95% ethanol and stored at -20°C.
(a) Water logged agricultural fields and barren lands

(b) Small artificial and natural ponds/ lakes

(c) Fresh water wells
Figure 3.2: Potential breeding sites (a-g) of *An. stephensi* in Madhya Pradesh (Central India).
Table 3.2: Geographic distance between different population pairs of *An. stephensi* mosquitoes.

<table>
<thead>
<tr>
<th>Sr. no.</th>
<th>Distance between</th>
<th>Distance (km)</th>
<th>Sr. no.</th>
<th>Distance between</th>
<th>Distance (km)</th>
<th>Sr. no.</th>
<th>Distance between</th>
<th>Distance (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GWL-SGR</td>
<td>270</td>
<td>13</td>
<td>SGR-KNW</td>
<td>333</td>
<td>25</td>
<td>ANP-UJN</td>
<td>614</td>
</tr>
<tr>
<td>2</td>
<td>GWL-UMR</td>
<td>395</td>
<td>14</td>
<td>SGR-UJN</td>
<td>312</td>
<td>26</td>
<td>ANP-IND</td>
<td>607</td>
</tr>
<tr>
<td>3</td>
<td>GWL-ANP</td>
<td>494</td>
<td>15</td>
<td>SGR-IND</td>
<td>322</td>
<td>27</td>
<td>JBL-BPL</td>
<td>261</td>
</tr>
<tr>
<td>4</td>
<td>GWL-JBL</td>
<td>380</td>
<td>16</td>
<td>UMR-ANP</td>
<td>102</td>
<td>28</td>
<td>JBL-KNW</td>
<td>404</td>
</tr>
<tr>
<td>5</td>
<td>GWL-BPL</td>
<td>333</td>
<td>17</td>
<td>UMR-JBL</td>
<td>100</td>
<td>29</td>
<td>JBL-UJN</td>
<td>429</td>
</tr>
<tr>
<td>6</td>
<td>GWL-KNW</td>
<td>522</td>
<td>18</td>
<td>UMR-BPL</td>
<td>352</td>
<td>30</td>
<td>JBL-IND</td>
<td>426</td>
</tr>
<tr>
<td>7</td>
<td>GWL-UJN</td>
<td>416</td>
<td>19</td>
<td>UMR-KNW</td>
<td>500</td>
<td>31</td>
<td>BPL-KNW</td>
<td>193</td>
</tr>
<tr>
<td>8</td>
<td>GWL-IND</td>
<td>454</td>
<td>20</td>
<td>UMR-UJN</td>
<td>520</td>
<td>32</td>
<td>BPL-IND</td>
<td>167</td>
</tr>
<tr>
<td>9</td>
<td>SGR-UMR</td>
<td>217</td>
<td>21</td>
<td>UMR-IND</td>
<td>523</td>
<td>33</td>
<td>BPL-IND</td>
<td>171</td>
</tr>
<tr>
<td>10</td>
<td>SGR-ANP</td>
<td>314</td>
<td>22</td>
<td>ANP-JBL</td>
<td>180</td>
<td>34</td>
<td>KNW-UJN</td>
<td>162</td>
</tr>
<tr>
<td>11</td>
<td>SGR-JBL</td>
<td>144</td>
<td>23</td>
<td>ANP-BPL</td>
<td>441</td>
<td>35</td>
<td>KNW-IND</td>
<td>114</td>
</tr>
<tr>
<td>12</td>
<td>SGR-BPL</td>
<td>150</td>
<td>24</td>
<td>ANP-KNW</td>
<td>574</td>
<td>36</td>
<td>UJN-IND</td>
<td>53</td>
</tr>
</tbody>
</table>

3.3. METHODS

3.3.1. Isolation of genomic DNA

Genomic DNA of individual female *An. stephensi* mosquito from each population was isolated from the legs and wings, following the modified protocol from Sambrook *et al.* (1989) as described below -

i. Individual ethanol preserved mosquitoes were allowed to dry overnight on filter paper inside of a covered Petri dish.

ii. Legs and wings of individual mosquito were plucked with the help of sterilized fine forceps.

iii. Legs and wings of individual mosquito were placed in a 1.5 ml grinding tube and were ground with polypropylene pellet pestle until the legs and wings were powdered.

iv. 60µl of warm DNA extraction buffer (DEB) and 23µg of RNase A was added. The mixture was ground for 30 second to homogenize the solution and incubated at 37°C for 60 minutes.

v. 3µl of Proteinase K was added and incubated at 50°C for 60 minutes.

vi. 60µl of phenol, chloroform and isoamyle (25:24:1) was added, vortexed and centrifuged at 14,000 rpm.

vii. The supernatant was transferred to a fresh 1.5ml tube and 150µl of ice-cold 95% ethanol was added. The supernatant was allowed to precipitate overnight in the -20°C freezer.
viii. Centrifuged for 10 minutes at 14,000 rpm and the supernatant was discarded. The DNA pellet was washed with 200µl of ice-cold 70% ethanol for 2 minutes.

ix. The excess ethanol was removed with micropipette. The pellet was allowed to dry completely at 37°C for 10 minutes.

x. The pellet was resuspended in 50µl of autoclaved double distilled water and stored at -20°C freezer.

3.3.2. Qualitative and Quantitative estimation of DNA

Qualitative and quantitative estimation of DNA was done by UV spectroscopy and agarose gel electrophoresis.

3.3.2.1. UV-Spectrophotometer method

1µl of stock DNA was dissolved in 1 ml of distilled water. Optical density (O.D.) was determined at 260 nm and 280 nm wavelengths in UV-spectrophotometer (Perkin Elmer)

\[ \text{DNA (µg/ml) = O.D.}_{260} \times 50 \times \text{dilution factor} \]

DNA quality was estimated by the O.D.\text{260}/O.D.\text{280} ratio. Samples of 1.8 to 2.0 were considered of good quality.

3.3.2.2. Agarose Gel Electrophoresis

Agarose gel (0.7%) using 1X TAE buffer was prepared for running the DNA samples, ethidium-bromide dye (1 µg/ml of gel vol.) was added prior to gel polymerization. For checking the quality of DNA, 2µl of DNA from stock solution of each sample was loaded per well of the gel and electrophoresis was carried out at 100 volt for 1 hour. 100ng of standard marker of DNA (100bp ladder, Bangalore Genie) was run alongside to determine the molecular size. For checking the quality of DNA, gel was visualized on a UV transilluminator. The intact DNA samples with high yield and without any smear were considered for making working DNA solutions.

3.3.3. Microsatellite loci

Initially 16 microsatellite loci; 15 developed by Veradi et al. (2002) (F10, H2ii, E7T, E7, E12, B2N, G1, H1, A7, D7, G11, B1, D8T, A10 and B2) and one (MSH1) by
Djadid et al. (2003; unpublished) for *An. stephensi*, were screened for the analysis using allele specific primers (Table 3.3). However, only 13 (H2II, E12, B2N, F10, D8T, E7, H1, A7, A10, G1, B2, MSH1 and B1) microsatellite loci were used for genotyping as three loci (E7T, D7 and G11) could not be amplified in all the studied populations. The microsatellite markers were labeled at 5’ position with 6-FAM dye for further use in the genotyping. All the primers were synthesized from Sigma-Genosys Ltd., U.S.A. After selecting 13 microsatellite loci for further study, different multiplex combinations were tried for PCR amplification, according to their allele size range and annealing temperature (T_a) with random individuals from all the populations. The final combinations made out of the 13 microsatellite markers for multiplex PCR and their annealing temperature (T_a) are given in Table 3.4.

**Table 3.3: Sixteen microsatellite loci screened for the analysis.**

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Locus</th>
<th>Primer sequence</th>
<th>Repeat array</th>
<th>Gene Bank Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F10*</td>
<td>5’-GTGGGTGCAGTAGTGGGATT-3’</td>
<td>(TGC)9</td>
<td>AF418589</td>
</tr>
<tr>
<td>2</td>
<td>H2II*</td>
<td>5’-ACATCGGTGGAGTTATCT-3’</td>
<td>(TG)18</td>
<td>AF418595</td>
</tr>
<tr>
<td>3</td>
<td>E7T</td>
<td>5’-GCTGCTACCGACGCCTAAC-3’</td>
<td>(CAA)8</td>
<td>AF418593</td>
</tr>
<tr>
<td>4</td>
<td>E7*</td>
<td>5’-GCTTGTGCTGCTGCTGCT-3’</td>
<td>(TG)17</td>
<td>AF418587</td>
</tr>
<tr>
<td>5</td>
<td>E12*</td>
<td>5’-ACATGCAACACCGAAACCC-3’</td>
<td>(AC)28</td>
<td>AF418586</td>
</tr>
<tr>
<td>6</td>
<td>B2N*</td>
<td>5’-GCTGGGAAGTGCTACAGGG-3’</td>
<td>(CA)30</td>
<td>AF418590</td>
</tr>
<tr>
<td>7</td>
<td>G1*</td>
<td>5’-CAGCCGAAAATTTGGTCA-3’</td>
<td>(CA)15</td>
<td>AF412812</td>
</tr>
<tr>
<td>8</td>
<td>H1*</td>
<td>5’-GATCTGTGATATCTGTAG-3’</td>
<td>(TG)11(TG)5(TGTA)3</td>
<td>AF412813</td>
</tr>
<tr>
<td>9</td>
<td>A7*</td>
<td>5’-CACAATCAACCCCCCT-3’</td>
<td>(CA)25</td>
<td>AF418592</td>
</tr>
<tr>
<td>10</td>
<td>D7</td>
<td>5’-CTGGTGCTCGTGGAAGG-3’</td>
<td>(GT)T</td>
<td>AF418594</td>
</tr>
<tr>
<td>11</td>
<td>G11</td>
<td>5’-GGAAGGAGGAGTGCTTTAACTGG-3’</td>
<td>(CA) (ATG)</td>
<td>AF418584</td>
</tr>
<tr>
<td>12</td>
<td>B1*</td>
<td>5’-ATAGGGCCCGCACTGTAT-3’</td>
<td>(CA)5(CA)4</td>
<td>AF418596</td>
</tr>
<tr>
<td>13</td>
<td>MSH1*</td>
<td>5’-GTTTGCTCCGTTACGGGT-3’</td>
<td>(GT) (AA)</td>
<td>CC894255</td>
</tr>
<tr>
<td>14</td>
<td>D8T*</td>
<td>5’-TCCTTCAGCTCGAGCCAT-3’</td>
<td>(CA)15</td>
<td>AF418585</td>
</tr>
<tr>
<td>15</td>
<td>A10*</td>
<td>5’-GCTGGACACCGAAGTGG-3’</td>
<td>(CA)12</td>
<td>AF418591</td>
</tr>
<tr>
<td>16</td>
<td>B2*</td>
<td>5’-ATACGCACCCCCCTTCCAC-3’</td>
<td>(CA)9</td>
<td>AF418588</td>
</tr>
</tbody>
</table>

*loci used in the study*
Table 3.4: Combination of different microsatellite used for Multiplex PCR

<table>
<thead>
<tr>
<th>S No.</th>
<th>Combination used</th>
<th>Annealing temperature (Tₘ) (° C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H1, A7 and A10</td>
<td>52° C</td>
</tr>
<tr>
<td>2</td>
<td>G1, B2 and MSH1</td>
<td>59° C</td>
</tr>
<tr>
<td>3</td>
<td>B2N and F10</td>
<td>58° C</td>
</tr>
<tr>
<td>4</td>
<td>H2ii and E12</td>
<td>59° C</td>
</tr>
<tr>
<td>5</td>
<td>D8T and E7</td>
<td>59° C</td>
</tr>
<tr>
<td>6</td>
<td>B1</td>
<td>60° C</td>
</tr>
</tbody>
</table>

3.3.4. Polymerase Chain Reaction (PCR)

3.3.4.1. PCR amplification

Sixteen primers were tested for PCR amplification in at least 30 individuals from each population. Genomic DNA (50-100 ng) from each sample was aliquoted into separate reaction tubes. The master mix was prepared for 30 reactions including one extra reaction mixture, for omitting any pipetting error (Table 3.5). The PCR reaction mixture preparation was done at 4° C under sterile conditions. The PCR reaction was equally aliquoted into each reaction tube containing DNA.

Table 3.5: Contents of the PCR reaction mixture.

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Component</th>
<th>Conc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DNA</td>
<td>45 ng</td>
</tr>
<tr>
<td>2</td>
<td>PCR buffer (Mg free)</td>
<td>1 X</td>
</tr>
<tr>
<td>3</td>
<td>MgCl₂</td>
<td>1.5-2.0 mM</td>
</tr>
<tr>
<td>4</td>
<td>Forward primer</td>
<td>1.5 pM</td>
</tr>
<tr>
<td>5</td>
<td>Reverse primer</td>
<td>1.5 pM</td>
</tr>
<tr>
<td>6</td>
<td>dNTPs mix</td>
<td>0.2 mM</td>
</tr>
<tr>
<td>7</td>
<td>Taq polymerase</td>
<td>1 Unit</td>
</tr>
<tr>
<td>8</td>
<td>Sterile distilled water to make total volume 25 µl</td>
<td></td>
</tr>
</tbody>
</table>

The reaction tubes were briefly centrifuged for proper mixing of DNA and other reagents. Tubes were then placed in a thermal cycler (9700 Applied Biosystems). The amplification protocol is depicted in Figure 3.3.
Fig 3.3: Cycling condition for standard multiplex PCR.
*Annealing temperature for each multiplex primer pair is different

3.3.4.2. Detection of PCR products

After completion of polymerase chain reaction, the PCR products were tested for amplification by running on a 2% agarose gel with ethidium-bromide dye, (10 µg/10 ml of gel vol.). 2µl of loading dye (3X) was added to each product prior to loading. 5µl of the above mix was loaded per well. The gel was run at 100 volts for 45 minutes and visualized under UV transilluminator. The PCR amplified product of different multiplex combinations is shown in Figure 3.4. All the populations showed similar banding pattern as alleles cannot be separated on agarose gel.
(a) H1, A7 and A10

(b) G1, B2 and MSH1

(c) B2N and F10
Figure 3.4: PCR product on 2% agarose gel for each multiplex combination (a-f).
The PCR products were stored at -20°C until their further use in genotyping. Genotyping was done by ABI 3730xl genetic analyser and the data was analysed by the Genemapper 4.0 software.

3.4. STATISTICAL ANALYSIS

Various parameters of population genetics such as genetic diversity, gene flow, population structure and genetic distances were analyzed using different population genetic programs available online.

3.4.1. Polymorphic Information Content (PIC)

A marker's usefulness for this purpose depends on the number of alleles it has and their corresponding relative frequencies. Qualitatively, a marker is called polymorphic if it has at least two alleles and its most frequent allele in the population has a frequency of at most 99%. Quantitatively, the degree of polymorphism is commonly measured by two distinct quantities. One is known as heterozygosity, and its unbiased estimator and variance formula are well known (Nei and Roychoudhury, 1974). Another measure of polymorphism is the polymorphism information content (PIC) value (Botstein et al., 1980), which was originally defined for a codominant marker used in a linkage study of a rare dominant disease but has more recently been shown to be relevant regardless of the mode of disease inheritance (Guo and Elston, 1999). The polymorphism information content (PIC) value is commonly used in genetics as a measure of polymorphism for a marker locus used in linkage analysis. The PIC value is defined as the probability that a given marker genotype of an offspring of an affected parent will allow deduction of the parental genotype at the marker locus and was shown by Botstein et al. (1980) to be:

\[ PIC = 1 - \sum_i p_i^2 - \sum_{i,j} p_i^2 p_j^2 \]

The PIC values for each microsatellite loci were calculated using Cervus 3.0.7.

3.4.2. Genetic diversity

The various population genetic parameters were studied by using GenAlEx 6.5 (Peakall and Smouse, 2006) computer program which reads information contained in an Excel worksheet that consists of essential parameters and labels, optional labels, and the data itself. Genetic diversity refers to any variation in nucleotides, genes,
chromosomes or whole genomes of organisms (Wang et al., 2009). Genetic diversity can be assessed among different accessions/individuals within same species (intraspecific), among species (interspecific) and between genus and families (Mittal & Dubey, 2009). The following parameters were studied:

3.4.2.1. **Number alleles**

Alleles are set of alternative forms of the same gene occupying the same relative position or locus on homologous chromosomes. Allele number is the total number of alleles for a given locus in a population, which is counted with a non-zero frequency.

3.4.2.2. **Allele frequency/ gene frequency**

The frequency of an allele ‘A’ is the number of ‘A’ alleles in the population divided by the total number of alleles/ genes (Smith, 1998). It gives an indication of the most or least prevalent alleles in the population.

3.4.2.3. **Effective number of alleles**

The effective number of alleles ($N_e$) is the reciprocal of the sum of the square of allele frequencies (Kimura and Crow, 1964). The effective allele number is given by the formula: $N_e = 1 / \sum P_i^2$, where $P_i$ is the frequency of $i^{th}$ allele.

3.4.2.4. **Heterozygosity**

Heterozygosity is the state of possessing different alleles at a given locus with regard to a given character. It is a measure of heterozygotes or genic variation in a population. The expected heterozygosity is given by the formula:

$$H = \sum P_i^2$$

The population heterozygosity at a locus is given by the formula:

$$H = 1 - \sum P_i^2$$

Where $\sum$ stands for summation over all alleles (Nei, 1978) and $P_i$ is the frequency of the $i^{th}$ allele at a locus in a population. The average heterozygosity per locus ($H$) is defined as the mean of $H$ over all structural loci in the genome.
An unbiased estimation of heterozygosity is $He$ (Expected heterozygosity):

$$He = \frac{n(1 - \sum_{i=1}^{k} Pi^2)}{n - 1}$$

where $n$ is the sample size and $Pi$ frequency of $i^{th}$ allele.

### 3.4.3. Analysis of molecular variance (AMOVA)

Analysis of Molecular Variance (AMOVA) is a method for studying molecular variation within a species. AMOVA treats any kind of raw molecular data as a Boolean vector $p_i$, that is, a $1 \times n$ matrix of 1’s and 0’s, 1 indicating the presence of a ‘$i$’ marker and 0 its absence. A marker could be a nucleotide base, a base sequence, a restriction fragment, or a mutational event. Euclidean distances between pairs of vectors are then calculated by subtracting the Boolean vector of one haplotype from another, according to the formula $(p_j - p_k)$. If $p_j$ and $p_k$ are visualized as points in $n$-dimensional space indicated by the intersections of the values in each vector, with $n$ being equal to the length of the vector, then the Euclidean distance is simply a scalar that is equal to the shortest distance between those two points. The squared Euclidean distances are then calculated using the equation:

$$\delta_{jk}^2 = (p_j - p_k)' W (p_j - p_k)$$

Where, $W$ is a weighting matrix; by default, it is an identity matrix and does not change the value of the final product; however, $W$ can be a matrix with a number of values depending upon how one weights molecular change at different locations on a sequence or phylogenetic tree. The AMOVA was performed by using GenAlEx 6.5 program.
3.4.4. Principal coordinate analysis (PCoA)

The PCoA is a multivariate statistical technique for finding the most important axes along which your samples vary. It can also be used to find clusters. PCoA calculates the distance matrix for each pair of samples using the UniFrac metrics. It then turns these distances into points in a space with a number of dimensions one less than the number of samples. The principal components, in descending order, describe how much of the variation each of the axes in this new space explains. The first principal component separates out the data as much as possible; the second principal component provided the next most separation, and so forth. GenAlEx 6.5 was used to perform the PCoA.

3.4.5. Hardy-Weinberg Equilibrium

The Hardy-Weinberg law states that in a large random mating population, the gene and genotype frequencies remain stable in the absence of forces like mutation, migration and selection. The equilibrium is represented by the following equation:

\[ P^2 + 2pq + q^2 = 1 \]
and
\[ P + q = 1 \]

Where, \( P \) and \( q \) are the dominant and recessive allelic frequencies of a given gene (Smith, 1998). The Hardy-Weinberg law could be used to test if random mating produced a population. The goodness of fit of observed genotypic frequencies to the frequencies expected under the Hardy-Weinberg law was determined. "Exact test” of Hardy Weinberg Equilibrium, given by Haldane (1954), Weir (1990) and Guo and Thompson (1992) available in Genepop web version 3.1(Raymond and Rousset, 1995) was used to know the deviations from the Hardy Weinberg expectation in all the \textit{An. stephensi} populations and to decide if the deviation would easily occur by chance alone if real differences in the observed an expected frequencies probably existed (Pirchner, 1969).

3.4.6. Linkage Disequilibrium (LD)

Linkage disequilibrium is the tendency of two alleles to be present on the same chromosome or not to segregate together. As a result, specific alleles at two different loci are found together more or less than expected by chance. Linkage disequilibrium is
the non-independence, at a population level, of the alleles carried at different positions in the genome. In this case, the expected frequency of a two-locus haplotype can be calculated as the probability of the occurrence of two independent (or joint) events simply by multiplying their gene frequencies. The same situation may exist for more than two alleles. Its magnitude is expressed as the delta (Δ) value and corresponds to the difference between the expected and the observed haplotype frequency. If there is no linkage disequilibrium, Δ will be zero (or not significantly different from zero), if there is positive linkage disequilibrium it will be a positive value. It can also be negative if the two alleles tend not to occur together. The statistical significance of linkage disequilibrium, which depends on the sample size, and the magnitude of linkage disequilibrium are separate issues. The statistical significance is determined by usually Fisher test and the magnitude is determined by either Δ value or alternative measures. The magnitude can be normalised (for allele frequencies) to have the same range of values for any frequency.

In the present study ‘exact tests’ of genotypic disequilibrium, between all pairs of loci were performed to evaluate the independence of loci available on Genepop web version 3.1 (Raymond and Rousset, 1995). The null hypothesis is: "Genotypes at one locus are independent from genotypes at the other locus".

3.4.7. F-Statistics

F statistics was estimated to provide a tool for analysis of variance among populations (Weir, 1990; Slatkin and Barton, 1989). The differentiation indices are given by the formulae:

\[ F_{ST} = \frac{\sigma^2a}{\sigma^wa + \sigma^wb + \sigma^ww} \]

\[ F_{IT} = \frac{\sigma^2a + \sigma^2b}{\sigma^wa + \sigma^wb + \sigma^ww} \]

\[ F_{IS} = \frac{\sigma^2b}{\sigma^wb + \sigma^ww} \]

where \( \sigma^2a = \) among sample breed variance component
\[ \sigma^2 b = \text{between individual within sample variance component} \]
\[ \sigma^2 w = \text{within individual variance component} \]

or, the three values are related by the formula:

\[ (1 - F_{IT}) = (1 - F_{IS}) (1 - F_{ST}) \]

where \( F_{IT} \) and \( F_{IS} \) are correlations between two gametes to produce the individuals relative to the subpopulations while \( F_{ST} \) is the correlation between the two samples drawn at random from each subpopulations.

\( F_{IS} \) is concerned with inbreeding in individual (I) relative to the subpopulation (S) to which they belong and a measure of the deviation of genotypic frequencies from panmictic frequencies in terms of heterozygous deficiency or excess. It is known as the inbreeding coefficient (f), which is conventionally defined as the probability that two alleles in an individual are identical by descent (autozygous). The technical description is the correlation of uniting gametes relative to gametes drawn at random from within a subpopulation (Individual within the Subpopulation) averaged over subpopulations. It is calculated in a single population as \( F_{IS} = 1 - (H_{OBS} / H_{EXP}) \) [equal to \( (H_{EXP} - H_{OBS} / H_{EXP}) \)] where \( H_{OBS} \) is the observed heterozygosity and \( H_{EXP} \) is the expected heterozygosity calculated on the assumption of random mating. It shows the degree to which heterozygosity is reduced below the expectation. The value of \( F_{IS} \) ranges between -1 and +1. Negative \( F_{IS} \) values indicate heterozygote excess (outbreeding) and positive values indicate heterozygote deficiency (inbreeding) compared with HWE expectations.

\( F_{ST} \) is concerned with inbreeding in subpopulation (S) relative to the total population (T) of which they are a part and it measures the effect of population subdivision, which is the reduction in heterozygosity in a subpopulation due to genetic drift. \( F_{ST} \) is the most inclusive measure of population substructure and is most useful for examining the overall genetic divergence among subpopulations. It is also called coancestry coefficient (q) (Weir & Cockerham, 1984) or 'Fixation index' and is defined as correlation of gametes within subpopulations relative to gametes drawn at random from the entire population (Subpopulation within the Total population). It is calculated as using the subpopulation (average) heterozygosity and total population expected heterozygosity. \( F_{ST} \) is always positive; it ranges between 0 = panmixis (no subdivision, random mating occurring, no genetic divergence within the population) and 1 = complete isolation (extreme subdivision). \( F_{ST} \) values up to 0.05 indicate negligible
genetic differentiation whereas >0.25 means very great genetic differentiation within the population analyzed. $F_{ST}$ is usually calculated for different genes, then averaged across all loci, and all populations. $F_{ST}$ can also be used to estimate gene flow: $0.25 \left(1 - F_{ST}\right) / F_{ST}$. This highly versatile parameter is even used as a genetic distance measure between two populations instead of a fixation index among many populations.

$F_{ST}$ is related to migration by and is given by formula:

$$F_{ST} = \frac{1}{4Nem + 1}$$

where, $N_e$ is the effective population size and ‘m’ is the effective proportion of immigrants. $N_em$ gives a measure of gene flow across populations.

It is the overall inbreeding coefficient ($F$) of an individual relative to the total population (Individual within the Total population). The F- statistics is estimated from observed frequencies. $F_{IT}$ is rarely used; hence it is not calculated in the present study.

### 3.4.8. R-Statistics

$R_{ST}$ (Slatkin, 1995) is an analogue of $F_{ST}$ (Wright, 1969), adapted to microsatellite loci by assuming a high-rate stepwise mutation model instead of a low-rate $K$- or infinite-allele mutation model. It’s calculated as:

$$R_{ST} = (Sbar - Sw)/Sbar$$

Where, $Sw$ is the sum over all loci of twice the weighted mean of the within-population variances $V(A)$ and $V(B)$, and $Sbar$ is the sum over all loci of twice the variance $V(A+B)$ of the combined population.

$F$-statistic and $R$-statistic were calculated through FSTAT computer programme (version 2.9.3) (Goudet, 1995). Significance of $F_{ST}$ and $R_{ST}$ values were calculated using bootstrapping over loci at 95% confidence level (Goudet, 1995).

### 3.4.9. Gene flow

Gene flow is the transfer of alleles of genes from one population to another. Migration into or out of a population may be responsible for a marked change in allele frequencies. Immigration may also result in the addition of new genetic variants to the established gene pool of a particular species or population.
A number of factors affect the rate of gene flow between different populations. One of the most significant factors is mobility, as greater mobility of an individual tends to give it greater migratory potential and mosquitoes can cover great distances by wind and human activity (passive dispersal) \textit{i.e.} trains, airplanes and trains. Maintained gene flow between two populations can also lead to a combination of the two gene pools, reducing the genetic variation between the two groups. It is for this reason that gene flow strongly acts against speciation, by recombining the gene pools of the groups, and thus, repairing the developing differences in genetic variation that would have led to full speciation and creation of daughter species.

Physical barriers to gene flow are usually, but not always, natural. They may include mountain ranges, oceans or vast deserts. Samples of the same species which grow on either side have been shown to have developed genetic differences, because there is no gene flow to provide recombination of the gene pools. The gene flow was calculated by the formula; $N_{m} = \frac{1}{4} \left[ \frac{1}{F_{ST}} - 1 \right]$ (Slatkin, 1995).

**3.4.10. Null alleles**

Null alleles are non-amplified alleles that, when segregating with another allele, result in an apparent homozygote. For microsatellites, such null alleles can arise when mutations occur in the flanking regions, preventing one or both of the primers from binding (Pemberton \textit{et al.}, 1995; Jones \textit{et al.}, 1998 and Holme \textit{et al.}, 2001).

Deviations from Hardy-Weinberg equilibrium can suggest that the sample does not represent a panmictic population, or alternatively, it might indicate the presence of null alleles and/or aberrations in microsatellite gel scoring. The effects of null alleles appear to be particularly clear when populations differ in null allele frequencies (Oosterhout \textit{et al.}, 2004).

Null allele frequencies were calculated according to the Brookfield. (1996) with the help of Micro-checker computer programme developed by Oosterhout \textit{et al.} (2004). Ignoring all non-amplified samples, the Chakraborty \textit{et al.} (1992) estimation of the null allele is:

$$r = \left[ \frac{(H_e-H_0)}{H_e} \right] / \left[ 2 - \frac{(H_e-H_0)}{H_e} \right]$$
Here, the observed heterozygosity, $H_0$, is measured from the data as $n_2/(n_1 + n_2)$, where $n_1$ is the number of one-banded individuals and $n_2$ is the number if two-banded individuals in the sample. The expected heterozygosity, $H_e$, is calculated as the sum of the product of all observed allele frequencies:

$$\sum P_i P_j, \text{with } i \neq j$$

Here, $P_i$ is the observed allele frequency of allele $i$, and that this frequency (and the frequency of all other visible alleles) is in fact overestimated due to the null allele.

Assuming there are no null allele homozygotes or ignoring all non-amplified samples as degraded DNA, human error, etc., Brookfield (1996) estimate of the null allele frequency is given by:

$$r = \frac{H_e - H_0}{1 + H_e}$$

Here, the observed heterozygosity, $H_0$, is measured from the data as $n_2 / (n_1 + n_2)$, where $n_1$ is the number of one-banded individuals and $n_2$ is the number if two-banded individuals in the sample. The expected heterozygosity, $H_e$, is calculated as the sum of the product of all observed allele frequencies:

$$\sum P_i P_j, \text{with } i \neq j$$

$P_i$ is the observed allele frequency of allele $i$, and that this frequency (and the frequency of all other visible alleles) is in fact overestimated due to the null allele.

Micro-checker distinguishes between the non-amplification caused by mutations in the flanking regions (i.e. null alleles), and other types of scoring errors that might result in non-amplification (such as large allele drop-out). With the help of this software a cut-off frequency was calculated for which allelic classes were analyzed to show a homozygote excess (and were expected to segregate with a null allele). That cut-off frequency equals $p > (2Nq-null)^{-1}$, in which $p$ is the allele frequency, $N$ is the total number of samples analyzed and $q$-null is the estimated null allele frequency. When most allelic classes (with allele frequency $p > (2Nq-null)^{-1}$) show a homozygote excess, the software warns of a potential null allele. The null allele
adjusted dataset was compared to original dataset using FreeNA (Chapuis and Estoup, 2007).

3.4.11. Bottleneck

Populations that have experienced a recent reduction of their effective population size exhibit a correlative reduction of the allele numbers (k) and gene diversity ($H_e$, or Hardy-Weinberg heterozygosity) at polymorphic loci. But the allele numbers is reduced faster than the gene diversity. Thus, in a recently bottlenecked population, the observed gene diversity is higher than the expected equilibrium gene diversity ($H_{eq}$) which is computed from the observed number of alleles (k), under the assumption of a constant-size (equilibrium) population (Luikart et al., 1998a).

This gene diversity excess ($H_e>H_{eq}$) has been demonstrated only for loci evolving under the Infinite Allele Model (IAM) by Maruyama and Fuerst (1985). If the locus evolves under the strict Stepwise Mutation Model (SMM), there can be situations where this gene diversity excess is not observed (Cornuet and Luikart, 1996). However, few loci follow the strict SMM, and as soon as they depart slightly from this mutation model towards the IAM, they will exhibit an gene diversity excess as a consequence of a genetic bottleneck. Because a few microsatellite loci follow the strict (one-step) SMM, it is recommended to use the Two-Phased model of Mutation (TPM) with Bottleneck test. The TPM is intermediate to the SMM and IAM. Most microsatellite data sets better fit the TPM than the SMM or IAM (Di Rienzo et al., 1994). The TPM is recommended for microsatellites consist of mostly one-step mutations, but a small percentage (5-10%) of multi-step changes (Luikart and Cornuet, 1998).

In a population at mutation-drift equilibrium (i.e., the effective size of which has remained constant in the recent past), there is approximately an equal probability that a locus shows gene diversity excess or a gene diversity deficit. To determine whether a population exhibits a significant number of loci with gene diversity excess, three tests have been proposed, namely a "sign test", a "standardized differences test" (Cornuet and Luikart, 1996), and a "Wilcoxon sign-rank test" (Luikart et al., 1998a). The first test suffers from low statistical power. The second test is not very useful since it requires at least 20 polymorphic loci. The Wilcoxon test provides relatively high
power and it can be used with as few as four polymorphic loci and any number of individuals (15-40 individuals and 10-15 polymorphic loci is recommend to achieve high power). A qualitative descriptor of the allele frequency distribution ("mode-shift" indicator) discriminates bottlenecked populations from stable populations (Luikart et al., 1998b). However, this qualitative method is not a proper statistical test as in this type I error rate varies with samples size. It cannot be used (with confidence) with samples of fewer than 30 individuals.

The bottleneck in five populations was tested by the Bottleneck computer programme. The program Bottleneck computes for each population sample and for each locus the distribution of the gene diversity expected from the observed number of alleles (k), given the sample size (n) under the assumption of mutation-drift equilibrium. This distribution is obtained through simulating the coalescent process of n genes under three possible mutation models, the IAM, the TPM and the SMM. This enables the computation of the average (Hexp) which is compared to the observed gene diversity (He, or Hardy-Weinberg heterozygosity; see Nei 1987) to establish whether there is a gene diversity excess or deficit at this locus.

In addition, the standard deviation (SD) of the mutation-drift equilibrium distribution of the gene diversity is used to compute the standardized difference for each locus ((Hobs-Hexp)/SD). The distribution obtained through simulation also enables the computation of a P-value for the observed gene diversity.

The way in which the coalescent process is simulated is unconventional due to the conditioning by the observed number of alleles. The phylogeny of the n genes is simulated as usual (Hudson, 1990).

Under the IAM, a single mutation is allocated at a time and the resulting number of alleles is computed. The process is repeated until the latter reaches the observed number of alleles.

Under the SMM, a Bayesian approach is used as explained by Cornuet and Luikart (1996). Briefly, the likelihood distribution of the parameter theta (= 4Neμ) given the number of alleles (k) and the sample size (n) is evaluated as the proportion of iterations (in the simulation process) producing exactly k alleles for a varying set of
thetas. As a second step, drawing random values of theta according to the likelihood distribution, the coalescent process is simulated as usual. Only gene diversities found in iterations producing exactly k alleles are considered.

Once all loci available in a population sample have been processed, the three statistical tests are performed for each mutation model as explained in Cornuet and Luikart (1996) and Luikart et al. (1998a,b) and the allele frequency distribution is established in order to see whether it is approximately L-shaped (as expected under mutation-drift equilibrium) or not (recent bottlenecks provoke a mode shift).

3.4.12. Isolation by distance (IBD)

The phrase “isolation by distance” was introduced by Sewall Wright in 1943 to describe the accumulation of local genetic differences under geographically restricted dispersal. In the context of population genetics, it is the process by which geographically restricted gene flow generates a genetic structure, because random genetic drift is occurring locally. It is an important phenomenon to consider whenever the genetic structure or the evolutionary trends of natural populations are to be analyzed spatially. Isolation by distance occurs in subdivided populations, when subpopulations exchange genes at a rate dependent upon the distance, or within a continuously distributed population, when dispersal of gametes and/or zygotes is spatially restricted.

Tests of isolation by distance were carried out using Isolation by distance web service (Jensen et al., 2005). IBWS analyzes the significance between the distance similarity matrix and the geographic distance matrix with a Mantel Test. This tells about, if genetic similarity between population pairs associated with the distance between them or not.

The strength of the relationship was determined by regressing all pairwise genetic similarity values against their corresponding geographic distances, using reduced major axis regression. Confidence intervals were calculated using a number of different approaches, including resampling algorithms in IBWS.
3.4.13. Bayesian clustering analysis

The genetic structuring of the *An. stephensi* mosquitoes in Madhya Pradesh was estimated by methods implemented into the STRUCTURE 2.3.4 (Pritchard et al., 2000). There are broadly two types of clustering methods:

1. **Distance-based methods.** These proceed by calculating a pairwise distance matrix, whose entries give the distance (suitably defined), between every pair of individuals. This matrix may then be represented using some convenient graphical representation (such as a tree or a multidimensional scaling plot) and clusters may be identified visually.

2. **Model-based methods.** These proceed by assuming that observations from each cluster are random draws from some parametric model. Inference for the parameters corresponding to each cluster is then done jointly with inference for the cluster membership of each individual, using standard statistical methods (for example, maximum-likelihood or Bayesian methods).

Distance-based methods are usually easy to apply and are often visually appealing. In the genetics literature, it has been common to adapt distance-based phylogenetic algorithms, such as neighbor-joining, to clustering multilocus genotype data (e.g., Bowcock et al., 1994). However, these methods suffer from many disadvantages: the clusters identified may be heavily dependent on both the distance measure and graphical representation chosen; it is difficult to assess how confident we should be that the clusters obtained in this way are meaningful; and it is difficult to incorporate additional information such as the geographic sampling locations of individuals. Distance-based methods are thus more suited to exploratory data analysis than to fine statistical inference, and we have chosen to take a model-based approach here.

The first challenge when applying model-based methods is to specify a suitable model for observations from each cluster. To make our discussion more concrete we introduce very briefly some of our model and notation here; a fuller treatment is given later. Assume that each cluster (population) is modeled by a characteristic set of allele frequencies. Let $X$ denote the genotypes of the sampled individuals, $Z$ denote the
(unknown) populations of origin of the individuals, and P denote the (unknown) allele frequencies in all populations. (Note that X, Z, and P actually represent multidimensional vectors.) Our main modeling assumptions are Hardy-Weinberg equilibrium within populations and complete linkage equilibrium between loci within populations.

Under these assumptions each allele at each locus in each genotype is an independent draw from the appropriate frequency distribution, and this completely specifies the probability distribution $\text{Pr}(XjZ, P)$ (given later in Equation 1). Loosely speaking, the idea here is that the model accounts for the presence of Hardy-Weinberg or linkage disequilibrium by introducing population structure and attempts to find population groupings that (as far as possible) are not in disequilibrium. While inference may depend heavily on these modeling assumptions, we feel that it is easier to assess the validity of explicit modeling assumptions than to compare the relative merits of more abstract quantities such as distance measures and graphical representations. In situations where these assumptions are deemed unreasonable then alternative models should be built.

The detailed description of our modeling assumptions and the algorithms used to perform inference is provided, beginning with the simpler case where each individual is assumed to have originated in a single population (no admixture).

**The model without admixture:** Suppose we genotype N diploid individuals at L loci. In the case without admixture, each individual is assumed to originate in one of K populations, each with its own characteristic set of allele frequencies. Let the vector $X$ denote the observed genotypes, $Z$ the (unknown) populations of origin of the individuals, and $P$ the (unknown) allele frequencies in the populations. These vectors consist of the following elements,

$$(x_{l}^{(i,1)}, x_{l}^{(i,2)}) = \text{Genotype of the } i^{th} \text{ individual at the } l^{th} \text{ locus},$$

where $i = 1, 2, \ldots, N$ and $l = 1, 2, \ldots, L$;

$z^{(i)} = \text{population from which individual } i \text{ originated};$

$P_{klj} = \text{frequency of allele } j \text{ at locus } l \text{ in population } k,$
where \( k = 1, 2, \ldots, K \) and \( j = 1, 2, \ldots, J \),

where \( J_l \) is the number of distinct alleles observed at locus \( l \), and these alleles are labeled \( 1, 2, \ldots, J_l \).

Given the population of origin of each individual, the genotypes are assumed to be generated by drawing alleles independently from the appropriate population frequency distributions,

\[
Pr \left( X_{l}^{(i,a)} = j/Z, P \right) = P_z(i)_j \quad (1)
\]

independently for each \( X_{l}^{(i,a)} \). (Note that \( P_z(i)_j \) is the frequency of allele \( j \) at locus \( l \) in the population of origin of individual \( i \)).

Assume that before observing the genotypes we have no information about the population of origin of each individual and that the probability that individual \( i \) originated in population \( k \) is the same for all \( k \),

\[
Pr(\varepsilon^{(i)} = k) = 1/K, \quad (2)
\]

independently for all individuals. (In cases where some populations may be more heavily represented in the sample than others, this assumption is inappropriate; it would be straightforward to extend our model to deal with such situations).

We follow the suggestion of Balding and Nichols (1995) in using the Dirichlet distribution to model the allele frequencies at each locus within each population. The Dirichlet distribution \( D(\lambda_1, \lambda_2, \ldots, \lambda_J) \) is a distribution on allele frequencies \( p = (p_1, p_2, \ldots, p_J) \) with the property that these frequencies sum to 1. We use this distribution to specify the probability of a particular set of allele frequencies \( p_{kl} \) for population \( k \) at locus \( l \),

\[
p_{kl} \sim D(\lambda_1, \lambda_2, \ldots, \lambda_J), \quad (3)
\]

independently for each \( k, l \). The expected frequency of allele \( j \) is proportional to \( \lambda_j \), and the variance of this frequency decreases as the sum of the \( \lambda_j \) increases. We take \( \lambda_1 = \lambda_2 = \cdots = \lambda_J = 1.0 \), which gives a uniform distribution on the allele frequencies.

**MCMC algorithm (without admixture):** Equations 1, 2, and 3 define the quantities \( Pr \left( XjZ, P \right) \), \( Pr(Z) \) and \( Pr(P) \), respectively. By setting \( \theta = (\theta_1, \theta_2) = (Z, P) \) and
letting $\pi (Z, P) = \Pr(Z, P/X)$ we can use the approach outlined in Algorithm 1 to construct a Markov chain with stationary distribution $\Pr(Z, P/X)$ as follows:

Algorithm 1: Starting with initial values $Z^{(0)}$ for $Z$ (by drawing $Z^{(0)}$ at random using $\Pr(z^{(i)} = k) = 1/K$ for example), iterate the following steps for $m = 1, 2, \ldots$.

Step 1. Sample $P^{(m)}$ from $\Pr(P/X, Z^{(m-1)})$.

Step 2. Sample $Z^{(m)}$ from $\Pr(Z/X, P^{(m)})$.

Informally, step 1 corresponds to estimating the allele frequencies for each population assuming that the population of origin of each individual is known; step 2 corresponds to estimating the population of origin of each individual, assuming that the population allele frequencies are known. For sufficiently large $m$ and $c$, $(Z^{(m)}, P^{(m)})$, $(Z^{(m+c)}, P^{(m+c)})$, $(Z^{(m+2c)}, P^{(m+2c)})$, \ldots will be approximately independent random samples from $\Pr(Z, P/X)$.

The model with admixture: We now expand our model to allow for admixed individuals by introducing a vector $Q$ to denote the admixture proportions for each individual. The elements of $Q$ are

$q_k^{(i)} = \text{proportion of individual i’s genome that originated from population } k.$

It is also necessary to modify the vector $Z$ to replace the assumption that each individual $i$ originated in some unknown population $Z^{(i)}$ with the assumption that each observed allele copy $X_{l}^{(i,a)}$ originated in some unknown population $Z_{l}^{(i,a)}$:

$Z_{l}^{(i,a)} = \text{population of origin of allele copy } x_{l}^{(i,a)}.$

We use the term “allele copy” to refer to an allele carried at a particular locus by a particular individual.

Our primary interest now lies in estimating $Q$. We proceed in a manner similar to the case without admixture, beginning by specifying a probability model for $(X, Z, P, Q)$. Analogues of (1) and (2) are

$\Pr (X_{l}^{(i,a)} = j/Z, P, Q) = P_{l,j}^{(i,a)}$ (4)

and

$\Pr (Z_{l}^{(i,a)} = k/P, Q) = q_k^{(i)}$ (5)
with (3) being used to model $P$ as before. To complete our model we need to specify a distribution for $Q$, which in general will depend on the type and amount of admixture we expect to see. Here we model the admixture proportions $q^{(i)} = (q_1^{(i)},..., q_K^{(i)})$ of individual $i$ using the Dirichlet distribution

$$q^{(i)} \sim D(\alpha, \alpha,..., \alpha)$$

independently for each individual. For large values of $\alpha (>1)$, this models each individual as having allele copies originating from all $K$ populations in equal proportions. For very small values of $\alpha (<1)$, it models each individual as originating mostly from a single population, with each population being equally likely. As $\alpha \to 0$ this model becomes the same as our model without admixture (although the implementation of the MCMC algorithm is somewhat different). We allow $\alpha$ to range from 0.0 to 10.0 and attempt to learn about $\alpha$ from the data (specifically we put a uniform prior on $\alpha \in [0, 10]$ and use a Metropolis-Hastings update step to integrate out our uncertainty in $\alpha$). This model may be considered suitable for situations where little is known about admixture.

**MCMC algorithm (with admixture):** The following algorithm may be used to sample from $Pr(Z, P, Q/X)$.

**Algorithm 2:** Starting with initial values $Z^{(0)}$ for $Z$ (by drawing $Z^{(0)}$ at random using (3) for example), iterate the following steps for $m = 1, 2,....$

1. **Step 1.** Sample $P^{(m)}, Q^{(m)}$ from $Pr(P, Q/X, Z^{(m-1)})$.
2. **Step 2.** Sample $Z^{(m)}$ from $Pr(Z/X, P^{(m)}, Q^{(m)})$.
3. **Step 3.** Update $\alpha$ using a Metropolis-Hastings step.

Informally, step 1 corresponds to estimating the allele frequencies for each population and the admixture proportions of each individual, assuming that the population of origin of each allele copy in each individual is known; step 2 corresponds to estimating the population of origin of each allele copy, assuming that the population allele frequencies and the admixture proportions are known. As before, for sufficiently
large $m$ and $c$, $(Z^{(m)}_m, P^{(m)}_m, Q^{(m)}_m)$, $(Z^{(m-c)}_m, P^{(m-c)}_m, Q^{(m-c)}_m)$, $(Z^{(m-2c)}_m, P^{(m-2c)}_m, Q^{(m-2c)}_m)$, \ldots will be approximately independent random samples from $\Pr(Z, P, Q/X)$.

3.4.14. Effective population size

Effective population size ($N_e$) is defined as "the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration". Generally the effective population size is less than the population under consideration. The effective population size ($N_e$) can be affected by natural forces acting on the population ($N$) resulting in the difference between them. These are as following:

1) One of the most important influences reducing $N_e$ relative to $N$ is fluctuating population size. This is because $N_e$ that accounts for fluctuating population size is calculated as the harmonic mean of the census size. The harmonic mean is the reciprocal of the average of the reciprocals.

$$N_e = \frac{1}{\left( \frac{1}{200} + \frac{1}{250} + \frac{1}{50} + \frac{1}{300} \right)^{-1}} = 123.7$$

The harmonic mean of 123.7 contrasts with the arithmetic mean of 200.

2) A second factor affecting $N_e$ is the breeding sex ratio. The equation is:

$$N_e = \frac{4 \times N_m \times N_\frac{e}{f}}{N_m + N_\frac{e}{f}}$$

3) A third influence on $N_e$ can have an interesting effect that sometimes enters into captive breeding designs. The effective population size ($N_e$) assumes a poisson distribution of family (offspring) numbers. The Poisson is characterized by having the variance equal the mean. If the variance is lower than the mean, then $N_e$ can actually be larger than the census size. In natural populations, if the environment causes the variance to exceed the mean (which may occur fairly frequently) then $N_e$ will again be less than $N$. 

66
4) Overlapping generations can also act to reduce $N_e$.

5) Yet another factor affecting $N_e$ is the spatial dispersion (pattern of spatial distribution) of the population. Its influence on the effective size is given by:

$$N_e = 4 \pi \sigma^2 \delta$$

where $\sigma^2$ is the variance of the dispersal distance and $\delta$ is the density of individuals. This formulation is often called the neighborhood size and assumes a normal (bell-shaped) distribution of dispersal distances (out in a circular shape from the source, hence the $\pi$). So, again, changes in the variance of dispersal size can affect $N_e$ (viscous populations will have smaller $N_e$).