GAVIN PUBLISHERS

Advances in Biochemistry and Biotechnology

Review Article

Damodaran D and Sudarsanam D. Adv Biochem Biotechnol 2: 142.
DOI: 10.29011/2574-7258.000042

Bioinformatic Analysis of Mosquito COX1 Gene Based on the Distance Values of Species Across India and the Globe

Divya Damodaran*, Sudarsanam D

Department of Advanced Zoology and Biotechnology, Loyola college, University of Madras, Chennai, India

***Corresponding author:** Divya Damodaran, Department of Advanced Zoology and Biotechnology, Loyola college, University of Madras, Chennai, India. Email: divya.shaheed@gmail.com

Citation: Damodaran D, Sudarsanam D (2017) Bioinformatic Analysis of Mosquito COX1 Gene Based on the Distance Values of Species Across India and the Globe. Adv Biochem Biotechnol 2: 142. DOI: 10.29011/2574-7258.000042

Received Date: 27 September, 2017; **Accepted Date:** 23 October, 2017; **Published Date:** 30 October, 2017

Abstract

Molecular phylogenetics provides insights into relationships among organisms - through "Species" trees and gene trees provides insights into the evolution and history of genes. It applies a blend of molecular and statistical methods to induce evolutionary connections among living beings or genes. The essential target of molecular phylogenetic studies is to recover the order of transformative events and represent them in developmental trees that graphically illustrate connections among species or genes after some time. Distance analysis compares two aligned sequences at a time, and builds a matrix of all possible sequence pairs. During each comparison, the number of changes (base substitutions and insertion/deletion events) are counted and presented as a proportion of the overall sequence length. Final estimates of the difference between all possible pairs of sequences are known as pairwise distances.

Introduction

Currently, the most commonly used barcode region for animals is a 5′-segment of the mitochondrial gene Cytochrome Oxidase I (COI) called the 'Universal' or 'Folmer' region. This region is the standard marker chosen by the Barcode of Life Database (BOLD), which is an online platform for collating and curating DNA barcoding information from around the world (Ratnasingham and Hebert 2007). Genetic techniques are considered to be relatively free from the subjectivity of identifying morphological features and can reveal the presence of cryptic species complexes that are often overlooked (e.g., Hemmerter et al. 2007). As such, barcoding as a method for identifying mosquitoes is vital to the accuracy of a surveillance program. In this study, we evaluated the use of the COI fragment as a barcode and compared the distance values of mosquito species present globally and in the Indian subcontinent. The study also discusses the relationships between different mosquito species and the composition of mosquito genera. Distance matrix is a phenetic approach preferred by many molecular biologists for DNA and protein work. This method estimates the mean number of changes (per site in sequence) in two taxa that have descended from a common ancestor. There is much information in the gene sequences that must be simplified in order to compare only two species at a time. The relevant measure is the number of differences in these two sequences, a measure that can be interpreted as the distance between the species in terms of relatedness. The overall mean distance values of global and Indian mosquito species were also studied.

Keywords: Bioinformatics; DNA Barcoding; Distance Values; Phylogenetic Analysis

Materials and Methods

Distance Matrix

Distance-matrix methods of phylogenetic analysis explicitly rely on a measure of "Genetic distance" between the sequences being classified, and therefore they require an MSA (multiple sequence alignment) as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches. Distance methods attempt to construct an all-to-all matrix from the sequence query set describing the distance between each sequence pair. From this is constructed a phylogenetic tree that places closely related sequences under the same interior node and whose branch lengths closely reproduce the observed distances between sequences. Distance-matrix methods may produce either rooted or unrooted trees, depending on the algorithm used to calculate them. They are frequently used as

the basis for progressive and iterative types of multiple sequence alignment. The main disadvantage of distance-matrix methods is their inability to efficiently use information about local high-variation regions that appear across multiple subtrees.

Estimating Evolutionary Distances Using Pairwise Distance

MEGA 6 software was used to you can calculate average pair-wise distances between sequences in several ways: Overall, Within Groups, Between Groups and Net Between Groups.

Basic Descriptive Statistics in SAS

PROC MEANS is used in a variety of analytic, business intelligence, reporting and data management situations. PROC MEANS capabilities may be employed in "Data cleansing" or "Exploratory data analysis" tasks to determine if incorrect or "Bad" values of analysis variables are contained in the data set that must be transformed or removed prior to further analysis. PROC MEANS is included the BASE Module of SAS System Software.

PROC MEANS noprint DATA=Aedes SKEW KURTOSIS MEAN STD T PROBT;

OUTPUTOUT=Aedes_mean_global SKEW=SKEWNESS_global KURTOSIS=KURTOSIS_global MEAN=MEAN_global std=STD_global T=NORMAL_global PROBT=PVALUE_global; var Distance_global; RUN;

DATA = - Specify data set to use

NOPRINT - Do not print output

VAR variable - specifies which numeric variables to use

OUTPUT OUT = datasetname - statistics will be output to a SAS data file

SKEW = - specifies the name of the column to be assigned for the Skew column produced in Proc Means procedure

KURTOSIS = - specifies the name of the column to be assigned for the Kurtosis column produced in Proc Means procedure

MEAN - Arithmetic average

STD - Standard Deviation

Line Overlay Plots

A line plot is a graphical display of data along a number line with symbols connected by a line. The symbols represent data points. The symbols can represent frequency. It is best to use a line plot when comparing fewer than 25 numbers. It is a quick, simple way to organize data.

A line plot will have outliers. An outlier is a number that is much greater or much less than the other numbers in the data set. Outliers are usually represented without any data transformation.

A line plot consists of a horizontal line which is the x-axis with equal intervals. It is important for a line to plot to have a title and a label of the x-axis to provide the reader an overview of what is being displayed. Also, line plots must have legends to explain what is being measured.

1.1.1. Line Overlay Plots in SAS

The GPLOT procedure plots the values of two or more variables on a set of coordinate axes (X and Y). The coordinates of each point on the plot correspond to two variable values in an observation of the input data set. The GPLOT procedure creates a temporary SAS data set that is used to generate an image map in an SVG file when you are sending output to the LISTING destination. (This option is not necessary when you are sending output to the HTML destination.) The drill-down URLs in the image map must be provided by variables in the input data set.

proc gplot data=Aedes; title "Aedes";

SYMBOL1 I=JOIN V=DOT WIDTH=1 HEIGHT=1 CV=GREEN CI=GREEN; /*Global*/

SYMBOL2 I=JOIN V=DOT WIDTH=1 HEIGHT=1 CV=BLUE CI=BLUE; /*Indian*/

SYMBOL3 I=JOIN V=DOT WIDTH=1 HEIGHT=1 CV=BROWN CI=BROWN; /*Kerala*/

PLOT (distance_global distance_indian distance_kerala)* distance / CAXIS=BROWN nolegend CFRAME=LIGHTBLUE overlay; run; quit;

DATA = - specifies the SAS dataset name from plots are to be done

TITLE "" - specifies title for the plot output

SYMBOL1, SYMBOL2, SYMBOL3 statements - specifies different output symbols and colours for different line plots

PLOT statement - specifies the list of numeric variables to be plotted in the overlay plot

OVERLAY option - specifies that many line plots should be shown in the same plot

Results and Discussion

Global genus	Indian genus	Kerala genus
Aedes	*Aedes*	*Aedes*
Culex	*Ficalbia*	*Culex*

Anopheles	*Mimomyia*	
Borichinda	*Orthopodomyia*	
Chagasia	*Uranotaenia*	
Limatus	*Aedeomyia*	
Lutzia	*Anopheles*	
Mansonia	*Malaya*	
Nyctomyia	*Tripteroides*	
Ochlerotatus	*Lutzia*	
Psorophora	*Ochlerotatus*	
Uranotaenia	*Armigeres*	
Wyeomyia	*Hodgesia*	
	Toxorhynchites	
Distance value analysis across genus		

Table 1: Distribution of mosquito genus across the globe, in India and in Kerala.

Estimates of Evolutionary Divergence between Sequences

Tables for evolutionary divergence between the mosquito species of Kerala, India and across the globe were generated using the maximum composite likelihood model. The number of base substitutions per site from between sequences are shown. Analyses were conducted using the Maximum Composite Likelihood model. The analysis involved 28 nucleotide sequences for Kerala species. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There was a total of 611 positions in the final dataset. Evolutionary analyses were conducted in MEGA6.

For Indian species, the number of base substitutions per site from between sequences are shown. Analyses were conducted using the Maximum Composite Likelihood model [1]. The analysis involved 16 nucleotide sequences in Indian species. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There was a total of 490 positions in the final dataset. In the case of global species, analyses were conducted using the Maximum Composite Likelihood model. The analysis involved 15 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There was a total of 512 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 Table 2-6.

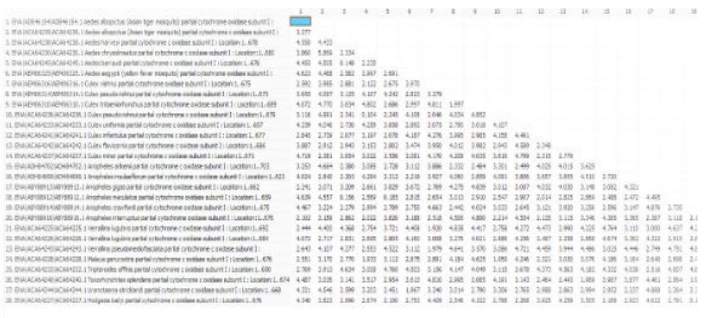

Table 2: Estimates of Evolutionary Divergence between Sequences of Kerala mosquito sp.

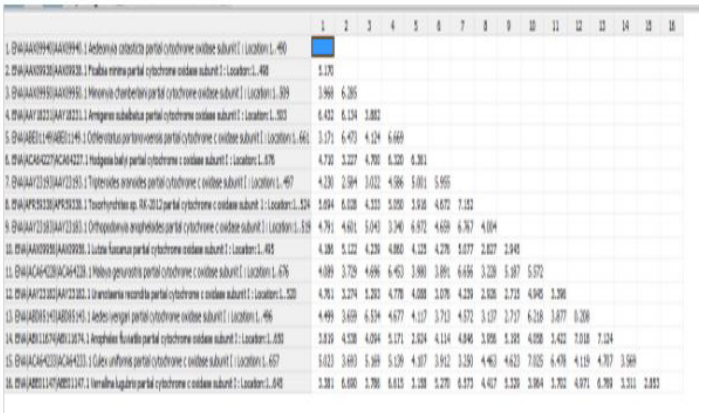

Table 3: Estimates of Evolutionary Divergence between Sequences of Indian mosquito species

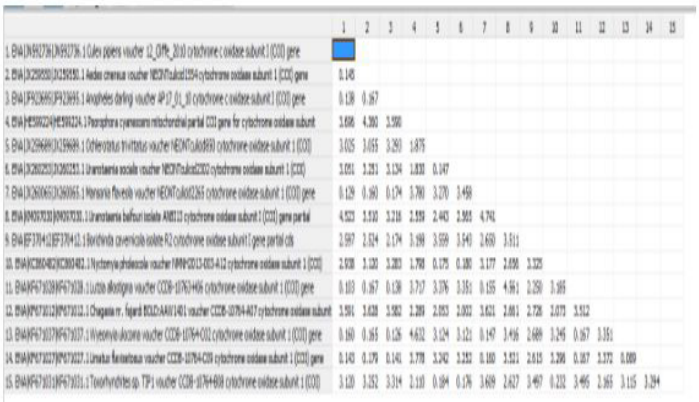

Table 4: Estimates of Evolutionary Divergence between Sequences of global mosquitoe

Genus	No of species	MEAN_Global	STD_Global	MEAN_Indian	STD_Indian	MEAN_Kerala	STD_Kerala
Aedeomyia	14			4.61015968	0.8209613		
Aedes	84	2.14150969	1.6723311	5.91519674	1.70899657	3.34697739	1.02060716
Anopheles	13	2.19742519	1.56109312	3.56893041			
Armigeres	11			5.18577073	0.97010325		
Borichinda	6	2.89695411	0.48829907				
Chagasia	3	2.99673371	0.68455408				
Culex	50	1.92248503	1.63237414			3.55076552	0.83660842
Ficalbia	13			4.56522958	1.32820502		
Hodgesia	9			4.25205447	0.80472496		
Limatus	1	3.30051513					
Lutzia	9	1.84121742	1.92963567	5.56352614	1.14175693		
Malaya	4			4.29329846	1.47338557		
Mansonia	8	2.27360141	1.85616911				
Mimomyia	12			4.59420894	0.87637907		
Nyctomyia	5	2.4187205	1.33145748				
Ochlerotatus	20	2.16678597	1.45729738	4.55916076	1.22457303		
Orthopodomyia	6			3.89686994	1.23080656		
Psorophora	11	3.03573939	1.17872096				
Toxorhynchites	7			3.50589741	0.62954431		
Tripteroides	8			5.31990604	1.39029998		
Uranotaenia	19	2.74427837	1.11870989	3.78157214	3.41777328		
Wyeomyia	2	1.618676	2.16322465				
The mean and standard deviation values for all the three groups are tabulated below.							

Table 5: Basic Descriptive Statistics of Distance Values of all Genus.

Genus	Number of species	SKEWNESS_ Global	KURTOSIS_ Global	SKEWNESS_ Indian	KURTOSIS_ Indian	SKEWNESS_ Kerala	KURTOSIS_ Kerala
Aedeomyia	14			0.5778686	0.877181		
Aedes	84	-0.328022	-1.869409			-0.5113681	1.4596398
Anopheles	13	-0.637107	-1.734709				
Armigeres	11			0.0026733	0.162524		
Borichinda	6	0.3483025	-1.147263				
Chagasia	3	-1.729392					
Culex	50	-0.20929	-2.116797			-0.32894	-1.0505693
Ficalbia	13			0.2051113	-1.45355		
Hodgesia	9			0.9890915	2.132938		
Limatus	1						
Lutzia	9	8.72E-05	-5.999483	-0.069212	-0.618239		
Malaya	4			1.8710308	3.512792		
Mansonia	8	-0.231863	-1.838699				
Mimomyia	12			0.5489878	1.588763		

Nyctomyia	5	-1.521145	1.71543				
Ochlerotatus	20	-0.682133	-1.50357	1.1189143	0.715222		
Orthopodomyia	6			0.0898244	-3.015679		
Psorophora	11	0.4350519	-1.719519				
Toxorhynchites	7			0.4797761	-1.558486		
Tripteroides	8			0.0635613	-1.328823		
Uranotaenia	19	-1.643276	2.064902	-0.439821			
Wyeomyia	2						
The distance values of global species are more negatively skewed than positive skewness. The distance values of Indian mosquito species are largely positively skewed. The Kerala species which are distributed only in *Aedes* and *Culex* genus are negatively skewed							

Table 6: The Skewness and Kurtosis for the Three Groups are Tabulated Below.

The distance values of global species are more negatively skewed than positive skewness. The distance values of Indian mosquito species are largely positively skewed. The Kerala species which are distributed only in *Aedes* and *Culex* genus are negatively skewed.

Scatter plots - Comparing the distance values of the three groups

Further, scatter plots were obtained for the distance values of each genus. It should be noted here that the top hits of distance values are taken for the analysis though the genus may actually be present in all the 3 groups - Global, Indian and Kerala. In the scatter plots below, the green line represents distance values of global species, the blue line represents distance values of Indian species and the brown line represents distance values of Kerala species.

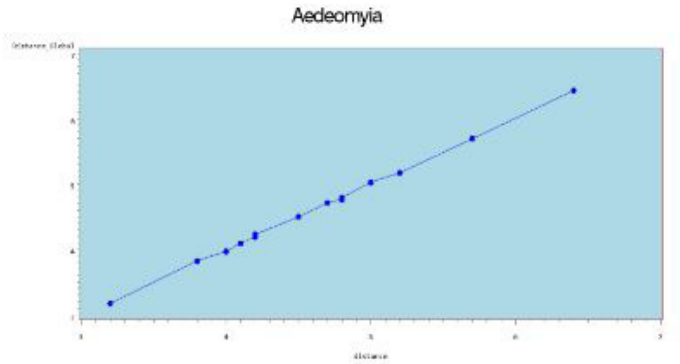

Figure 1: *Aedeomyia (Aedeomyia* was found to be closer to other species among the Indian genus).

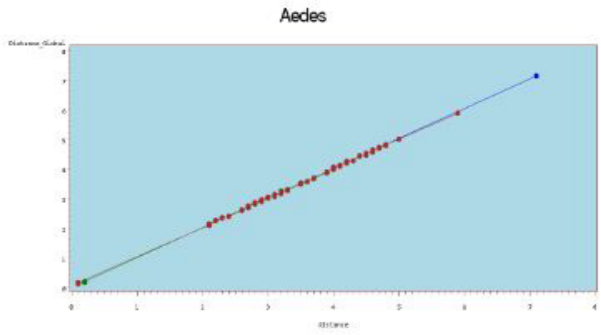

Figure 2: *Aedes (Aedes* species are found across the globe however it was prevalent in Kerala. This was evident from the distance values being more in Kerala).

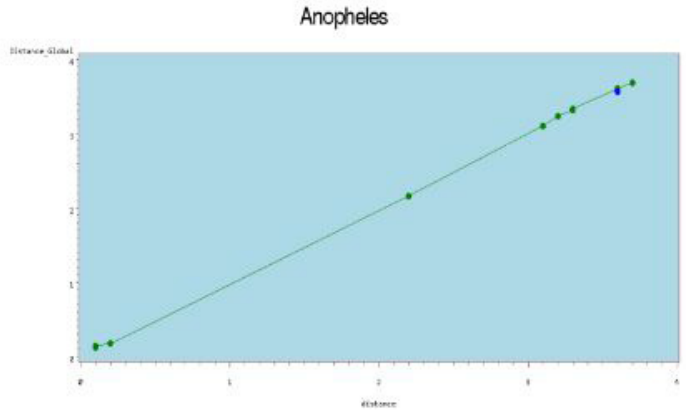

Figure 3: *Anopheles* (The distance values of *Anopheles* were found in global species).

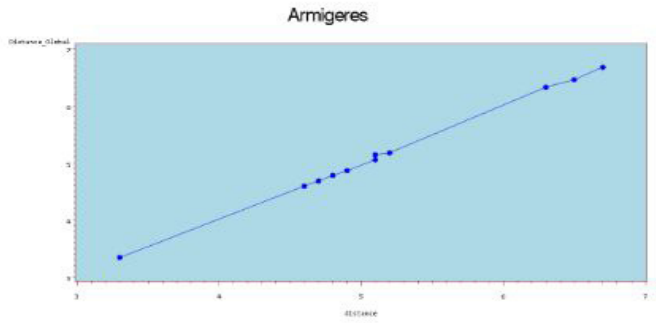

Figure 4: *Armigeres* (The distance values of *Armigeres* species were closer only in Indian species).

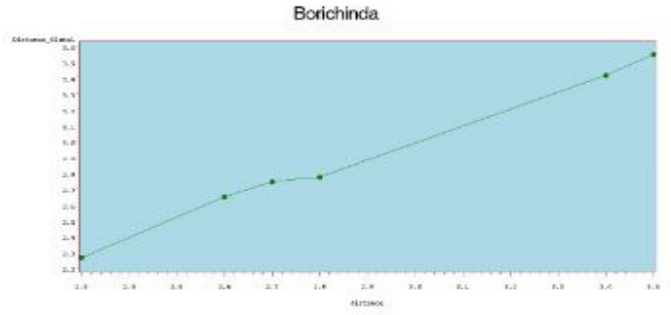

Figure 5: *Borichinda (Borichinda* species are found across the globe but not in India).

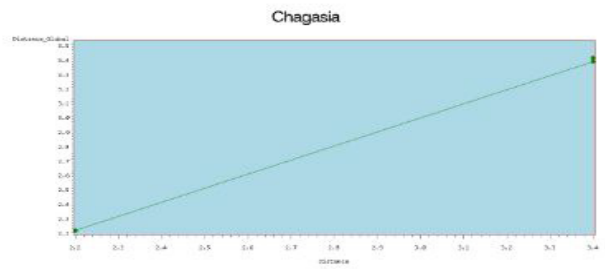

Figure 6: *Chagasia (Chagasia* species are found across the globe but not in India).

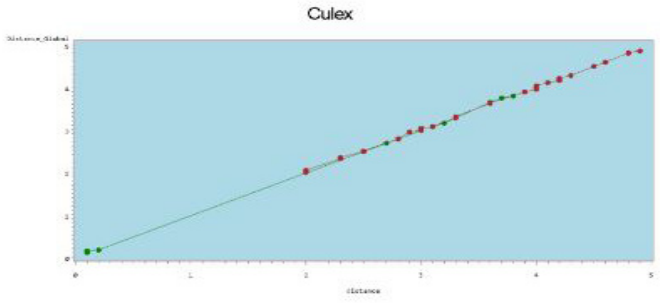

Figure 7: *Culex (Culex* species are found across the globe however it was more prevalent in Kerala. This was evident from the distance values being more in Kerala).

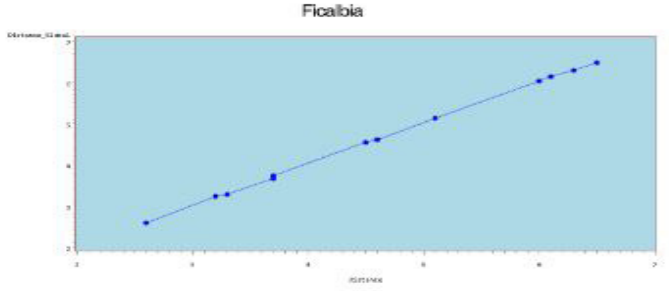

Figure 8: *Ficalbia (Ficalbia* species are closely related to Indian species and hence are present only in the group of distance values of Indian species.

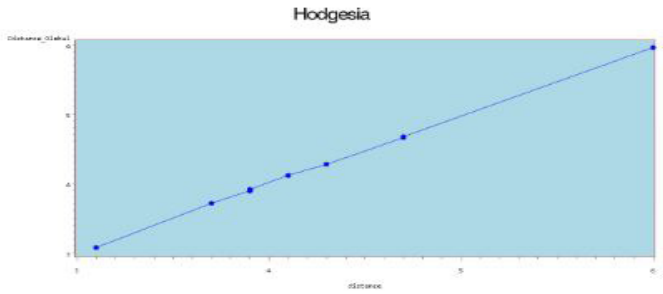

Figure 9: *Hodgesia (Hodgesia* species are found only in India).

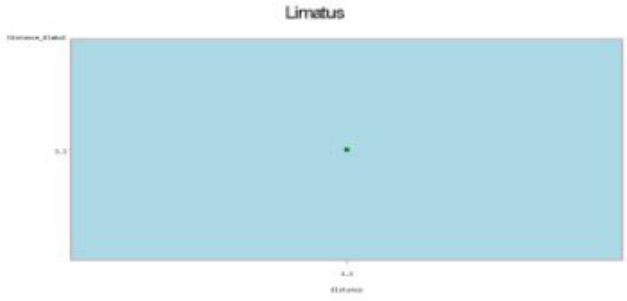

Figure 10: *Limatus* (Only one distance value was obtained for the genus *Limatus* in the group of global mosquito species. *Limatus* was not found to be related to any species in the Indian and Kerala species groups.

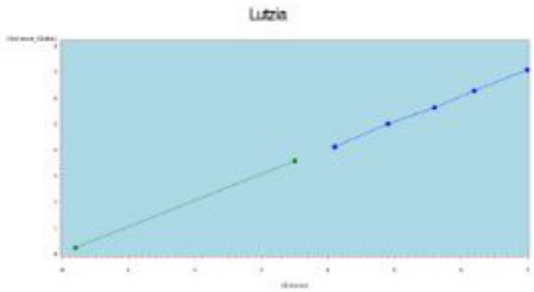

Figure 11: *Lutzia* (The species of *Lutzia* were closely related to global and Indian species as evident from the chart.

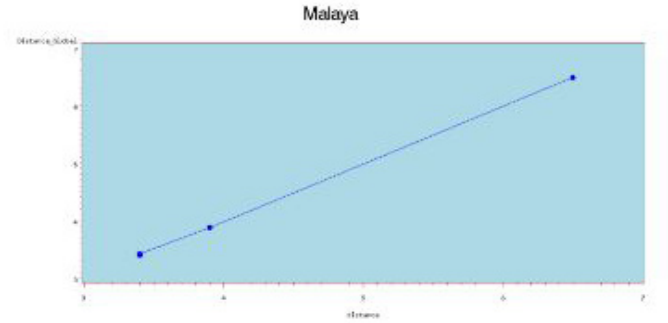

Figure 12: *Malaya* (Species of *Malaya* are present in the global species group with distance value close to few other global species).

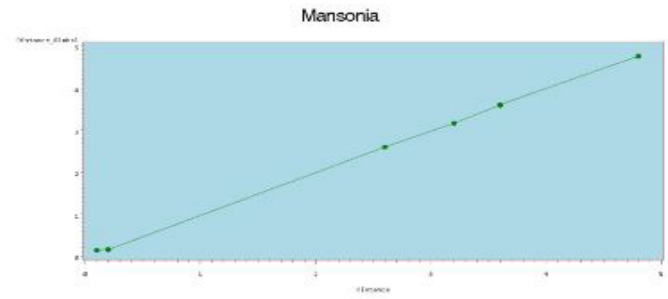

Figure 13: *Mansonia* (*Mansonia* species were found among the global species).

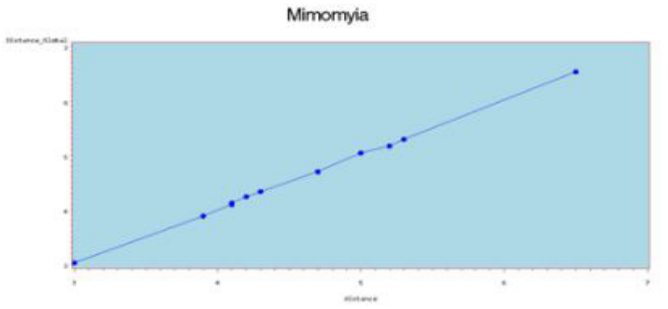

Figure 14: *Mimomyia* (*Mimomyia* species were found only in the Indian species group).

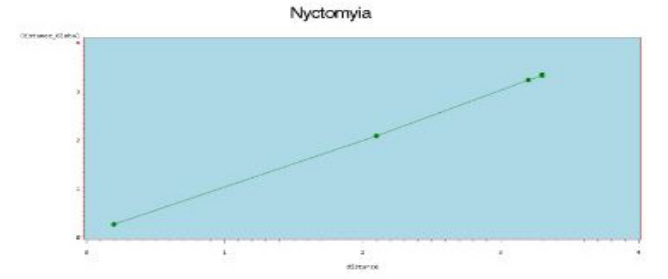

Figure 15: *Nyctomyia* (*Nyctomyia* species occur only in the global species group).

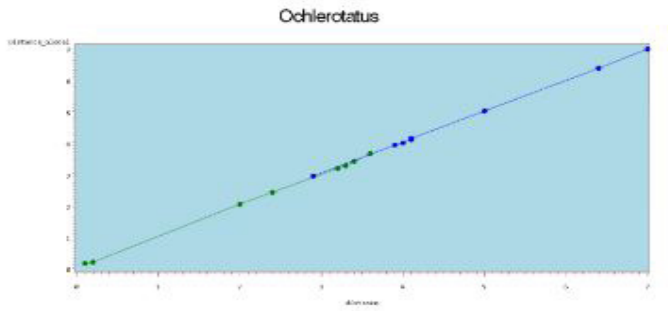

Figure 16: *Ochlerotatus* (Species of *Ochlerotatus* was found in the global and Indian species groups with wide range of distance values.

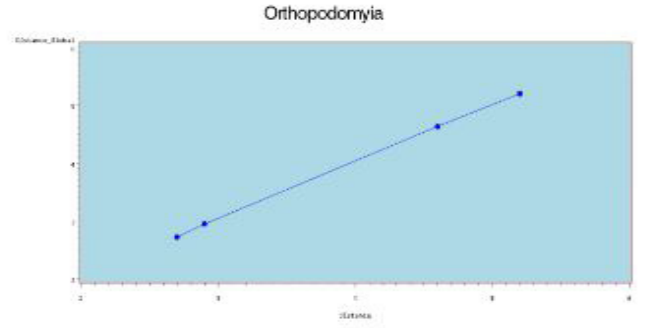

Figure 17: *Orthopodomyia* (Species of *Orthopodomyia* was found only in Indian species with a small range of distance values.

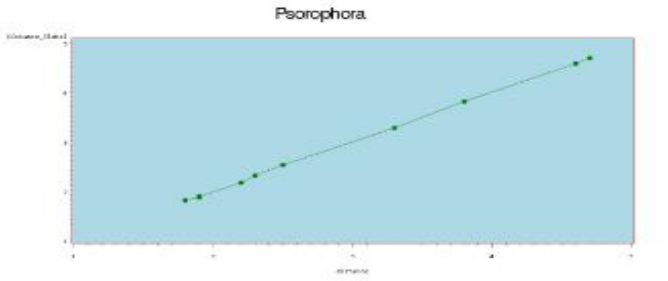

Figure 18: *Psorophora* (*Psorophora* genus occurs only in the global genus group with a small range of distance values).

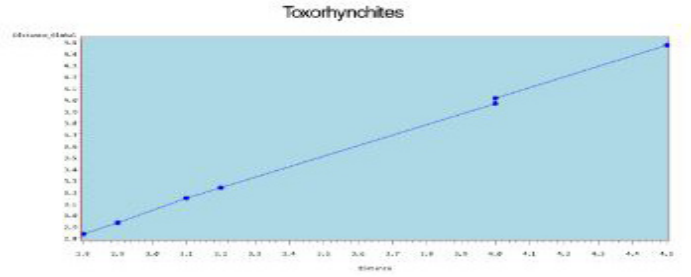

Figure 19: *Toxorhynchites* (*Toxorhynchites* genus occurs only in Indian genus group with a very small range of distance values against few Indian species.

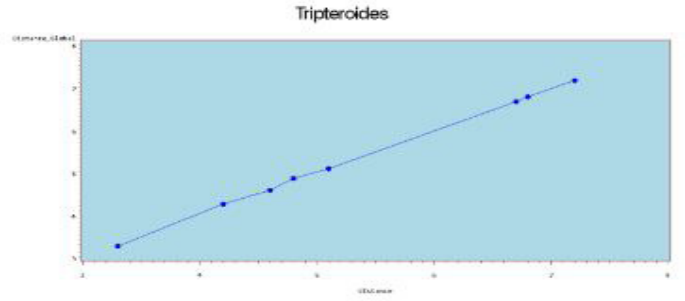

Figure 20: *Tripteroides* (*Tripteroides* genus occurs only in Indian genus group with high distance values against few Indian species.

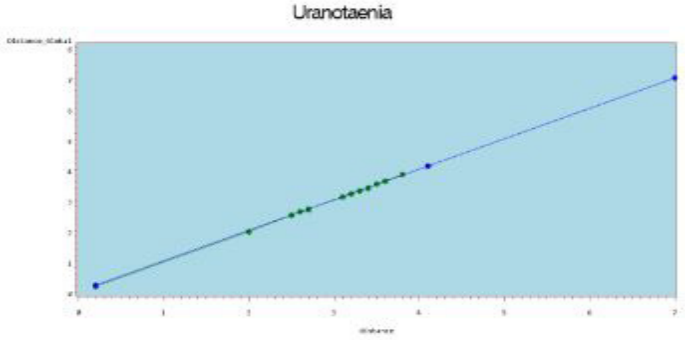

Figure 21: *Uranotaenia* (*Uranotaenia* genus shows a wide range of distance values in the global and Indian species group. Though it shows relatedness only to very few Indian species, the range of distance value is very high.

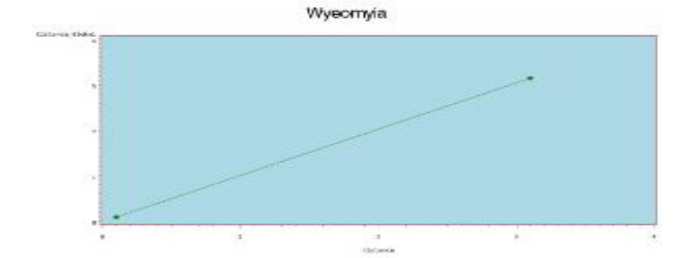

Figure 22: *Wyeomyia* (*Wyeomyia* shows close relatedness to only two species in the global group).

Global species	2.371
Indian species	4.678
Kerala species	3.33

Overall Mean Distance Values in the Three Groups

Discussion

The biostatistical analysis of the mosquito species of the three-geographic region suggested that *Toxorhynchitis*, *Orthopodomyia* and *Mimomyia* occur only in the Indian species. *Limatus* was not found to be related to any species in the Indian and Kerala species groups. *Hodgesia* species was found only in India. *Borichinda*, *Chagasia* and *Nyctomyia* are found only in the global species.

The greater the genetic distance between populations, the less breeding there is between them and the more isolated they are from one another. The lower the genetic distance between populations, the more breeding there is between them and the less isolated they are from one another. Values on the high end indicate some isolation between populations, and most likely mean that the populations are not currently breeding with one another. Values on the low end indicate that the populations are sharing their genetic material through high levels of breeding. Overall mean distance value in the Indian species is recorded as 4.67, suggesting there is more isolation among the Indian species. If two species have a small distance between them (as measured by the number of differences in their character sequences), then they have a recent common ancestor; but if they are far apart, then their common ancestor is in the remote past. We can use the distance between the species as a measure of the distance in time since the species diverged. These two distances, the number of character differences and the time since divergence, will be approximately proportional when they're relatively small.

Conflict of Interests:

The author declares no conflict of interests.

References

1. Berger J. Introduction to Molecular Phylogeny Construction. BIOL 334.
2. Day WHE (1986) Computational complexity of inferring phylogenies from dissimilarity matrices". Bulletin of Mathematical Biology 49: 461-467.
3. Endo T, Ogishima S, Tanaka H (2003) Standardized phylogenetic tree: a reference to discover functional evolution J Mol Evol 57: 174-181.
4. Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.
5. Felsenstein J (2004) Inferring Phylogenies Sinauer Associates: Sunderland, MA.
6. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach, Journal of Molecular Evolution 17: 368-376
7. Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155: 279-284.
8. Hillis DM, Moritz C, Mable BK, eds. (1996) Molecular systematics, 2nd ed. Sinauer Associates, Sunderland, Massachusetts.
9. http://study.com/academy/lesson/what-is-a-line-plot-in-math-definition-examples.html
10. http://support.sas.com/documentation/cdl/en/graphref/67881/HTML/default/viewer.htm#n1ca4rvgoodca6n19cn5npeiwccb.htm
11. http://www2.sas.com/proceedings/sugi29/240-29.pdf

12. http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

13. http://www.mathplanet.com/education/algebra-2/equations-and-inequalities/line-plots-and-stem-and-leaf-plots

14. http://www.socialresearchmethods.net/kb/statdesc.php

15. http://www.stattutorials.com/SAS/TUTORIAL-PROC-MEANS.htm

16. Mount DM (2004) Bioinformatics: Sequence and Genome Analysis 2nd ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor NY.

17. Pagel M (1999) Inferring historical patterns of biological evolution. Nature 401: 877-884.

18. vise JC (2000) Phylogeny: The history and formation of species. Harvard University Press, Cambridge, Massachusetts.

19. Wen-Hsiung Li (1997) Molecular Evolution. Sinauer Associates.

20. Whelan S, Lio P, Goldman N (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past Trends in Genetics 17: 262-272.

21. Zuckerlandl E, Pauling, L (1962) Molecular disease, evolution, and genetic heterogeneity. In Horizons in Biochemistry (Kasha,M. and Pullman, B., eds), 189-225, Academic Press 1921-1930