Genetic structure in genus *Philautus* (Gistel, 1848) Family: Rhacophoridae, Order: Anura) along an altitudinal gradient in Eastern Himalayas

Submitted towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune

Supervisor

Ramana Athreya
Department of Biology

CERTIFICATE

This is to certify that this dissertation entitled **Genetic structure along an altitudinal gradient in *Philautus* genus of bush frogs** towards the partial fulfilment of the BS-MS dual degree programme at **Indian Institute of Science Education and Research, Pune** represents original research carried out by **Anurag Mishra** under the supervision of **Dr. Ramana Athreya, Associate Professor, Department of Biology** during the academic year **2014-2015.**

Ramana Athreya

Ramana Athreya

Associate Professor

24th March 2015

DECLARATION

I hereby declare that the matter embodied in the report entitled **Genetic structure along an altitudinal gradient in *Philautus* genus of bush frogs** are the results of the investigations carried out by me at the **Department of Biological Sciences, Indian Institute of Science Education and Research, Pune**, under the supervision of **Dr. Ramana Athreya** and the same has not been submitted elsewhere for any other degree.

Anurag Mishra.

Anurag Mishra

20101051

Integrated BS-MS program

24th March 2015

Acknowledgments

The making of this thesis was a lengthy process which could not have been possible without the unflinching support of many people. In this section I want to convey my sincere thanks to these people.

Dr. Ramana Athreya for mentoring me throughout my time at IISER Pune, for always trying to make me more organized and pushing me towards doing better.

Dr. Neelesh Dahanukar, Dr. Farhat Habib, Dr. Raghav Rajan and Dr. Deepak Barua for help and guidance at various times.

Mansee and Gauri for help with the lab work, Rohan, Aamod and Chintan for the great times in the field and my other lab-mates Naresh, Hrutuja, Ravi, Jay for all the good times!, Sree Vani and Anur for help with fieldwork, my field guides- Dinesh and Arjun for making fieldwork a success.

Last and most importantly, all the wonderful people managing the camps in Eaglenest Wildlife Sanctuary, especially Nima and Raju for the unforgettable experience that was the fieldwork in Arunachal Pradesh.

ABSTRACT

We studied the genetic structure within *Philautus* genus of bush frogs along an altitudinal gradient in Eaglenest Wildlife Sanctuary, Arunachal Pradesh, using a region of 12s-16srRNA gene of the mitochondrial genome. *Philautus* are the most species rich genus among all Indian amphibians with high phenotypic variation within species.

Haplotype analysis reveals three distinct evolutionary lineages with enough genetic divergence to warrant separate species status. Two of these species have little overlap in terms of their altitudinal distribution. Analysis of morphological variations using Discriminant Analysis shows results mostly congruent with the genetic analysis, though the separations between the species is not as clear. Further, all three groups have very little within group genetic diversity, suggesting strong altitude specific mitochondrial selective sweeps that eliminate novel mutations from the mitochondrial genome.

Table of Contents

List of Figures

List of Tables

INTRODUCTION

Genetic variation and evolutionary forces in natural populations

Genetic variation among natural populations is the starting step towards evolution and speciation. The chief factors that shape the amount of genetic variation in populations are drift, mutation, selection and migration.

Mutation generates genetic diversity. Mutations occur slowly but continuously. Mutations could have positive, neutral or negative impact on the fitness of an organism. A majority of mutations have no effect. Mutations with positive impact are the raw material for adaptive evolution.

Drift is the stochastic variation in frequencies of alleles over different generations. Genetic diversity can be gained and lost through the process of drift.

Migration, or gene flow in the context of population genetics, refers to movement of individuals into or out of defined population units (Elmer et al., 2007). If migrating individuals mate with individuals of the native population, they could transfer alleles into the native population gene pool which changes the allele frequencies of the native population. Migration slows down the genetic divergence between populations and thus, slows down the process of speciation as well.

Natural Selection produces adaptation of an organism with its environment. As such, natural selection has the ability to conserve genetic makeup over long periods of time in the face of migration, mutation and drift which tend to alter the genetic makeup (Rice et al., 2011).

We can look at evolution in terms of changes in gene and genotype frequencies within populations. The four factors - mutation, drift, natural selection, and migration acting within and among populations cause micro-evolutionary changes and these account for macro-evolutionary patterns on larger timescales. These macro-evolutionary patterns characterize the higher taxonomic groups (Hickerson et al., 2010).

Phylogeography

Phylogeography is considered the bridge between population genetics and phylogenetics (Avise, 1978). Species with broad geographical distributions very rarely have a homogeneous genetic make-up throughout its range. The interplay of biotic and abiotic factors in different parts of the range could lead to diverse effects of migration on mutation and selection. This leads to different allele frequencies in different parts of the geographic range of the organism, resulting in population structure within the species.

Variable selection pressures in heterogeneous landscapes can lead to local adaptation of populations (Cheviron & Brumfield, 2009). Gene flow among populations experiencing different selection pressures can influence local adaptation in a number of ways. If the strength of selection is weak compared to the level of gene flow, genetic variation will be persistent among populations, slowing the attainment of local adaptive optima and maintaining some amount of genetic diversity. When selection is strong, genetic differentiation can be maintained because selection against maladapted immigrants and mutants removes genetic variability caused by gene flow, and local adaptation may occur even in the face of on-going gene flow (Abbott et al., 2013; Seehausen et al., 2014). The interplay of evolutionary mechanisms leads to different patterns of spatial structuring of genetic diversity within populations, which is the first step towards speciation.

To detect if selection is operational using orthologous DNA sequences, there are several test statistics such as Tajima's D and Fu's Fs (Holsinger, 2010).

Tajima's D is used to distinguish between a population of orthologous DNA sequences evolving neutrally and one evolving under selection. It is computed as the difference between two measures of genetic diversity (θ) - the average number of pair wise differences (π) and the number of segregating (polymorphic) sites in the DNA dataset (S).

When both estimates of θ are roughly equal, the value of D is close to zero, which indicates evolution of sequences under neutral conditions. However, significant departures from zero indicate selection at work. Large negative values of D result when there are very few rare alleles in the population, indicating strong negative selection on maladapted alleles. Large positive values result when two different alleles are maintained in the population, which is an effect of balancing selection.

Mitochondrial Markers for Phylogeography

Historically, mitochondrial DNA (mtDNA) has been the most widely used marker for population genetic and demographic studies in animals (Hurst & Jiggins, 2005). The vertebrate mitochondrial genome is extra chromosomal DNA that is 14-23 kb long and consists of 13 protein coding genes which cooperate in electron transport machinery. MtDNA is easy to amplify for sequencing purposes due to high copy number. MtDNA is inherited only through the maternal line and therefore, the molecule can be assumed to have undergone negligible recombination. However, mtDNA only reflects historical processes in females. When the dispersal of individuals of a species is sex dependent, mtDNA analyses will only provide a picture of the maternal line, not of the species as a whole (Hurst & Jiggins, 2005).

Genes that have been characterised to have a high ratio of inter specific to intra specific genetic variation are ideal for characterising species. Cytochrome Oxidase I (COI) gene of the mitochondrial genome has been widely used as a 'barcode' gene to differentiate between species (Hebert et al., 2003). This particular gene exhibits high variation between species and low variation within species, making it ideal for the purpose of delimiting species. The inter-specific to intra-specific variation in the COI gene is roughly of the scale 10:1. For amphibians, the 12srRNA-16srRNA which includes tRNA for Valine in between is used as a barcode gene. It is also extensively used for population level studies in reptiles and amphibians (Chapple et al., 2011; Kotaki et al., 2010; Rodríguez et al., 2010). The priming sections for the 12S-16S locus are highly conserved and therefore, it can be easily amplified in taxa separated by large evolutionary timescales. Within this locus, there is significant variation in the rate of mutation across the entire region which makes it suitable for addressing questions at different timescales.

We wanted to use multiple mitochondrial genes and nuclear genes for studying phylogeography in frogs, but due to reasons beyond our control, we had to use only a single mitochondrial marker for the work presented in this thesis.

Haplotype Networks

Haplotype is a unique sequence of bases over a region of the genome. Individuals that ‘share a haplotype’ have identical sequences in the region of analysis. The longer the sequence length used for comparison between populations, the lower the chance that individuals within or between populations will be identical across the full extent of that sequence. If the sequence is too short, it could be identical within and between populations. Therefore, the length of the sequence used should be decided based on the taxa being studied and the mutation rate of the marker being studied.

Haplotype networks represent the relationships among the different haploid genotypes in the dataset. Identical sequences are pooled into a single terminal and the distances between the terminals are proportional to the genetic distance between them. Haplotype Networks differ from Phylogenetic Trees in enabling identification of the ancestral types of the haplotype under consideration.

Haplotype Diversity is a measure of uniqueness of a haplotype in a given dataset (Nei & Tajima,1981). It is given by $h = (1 - \Sigma x_i{}^{2})\, n / (n - 1)$, where, x_i is the relative frequency of each haplotype and n is the sample size.

Nucleotide Diversity is the average number of nucleotide differences per site between any two DNA sequences chosen randomly from the dataset. It is denoted by π, where $\pi = \Sigma_{ij} x_i x_j \pi_{ij}$, where, x_i and x_j are the ith and jth sequences respectively and π_{ij} is the number of nucleotide differences per nucleotide site between the ith and jth sequences (Nei & Tajima, 1981).

Previous phylogeographic studies of frogs

Frogs are considered an ideal taxon for addressing questions in phylogeography and speciation, due to their limited dispersal capability. As a result of this they exhibit a very high degree of population structure.

A study on the species *Eupsophus calcaratus* (Nuñez et al., 2011) in the Andes using three mitochondrial markers (D-Loop 582bp, Cyt-B 698 bp, 16S 934bp) revealed six distinct haplotype lineages with some haplotypes being present in more than one sampling location.

A study on *Rana dalmatina* populations in the Italian peninsula using 741 bp of Cytochrome Oxidase I (COI) and 653 bp of Cytochrome B revealed four geographically defined divergent lineages, and a higher genetic diversity in the species than studies based on the same species in other parts of Europe (Canestrelli et al., 2014).

Newman et al (2001) analysed the population structure of *Rana sylvatica* in prairie wetlands using microsatellite loci. Their results indicated gene flow on a geographical scale of 50m to 5.5 km, with evidence of differentiation at a scale of 20km.

A study of the *Eleutherodactylus* genus in Central America using the mitochondrialND2 gene of the mitochondrial genome and nuclear c-myc gene revealed high levels of genetic differentiation without much morphological differentiation (García-R et al., 2012). In frogs that require water bodies for their reproductive cycle, long distance migration is rarer, leading to reduced gene flow. The high degree of differentiation in *Eleutherodactylus* was no expected given that these frogs undergo development without a tadpole stage.

Analysis using 7 microsatellite loci in *Rana cascadae* to investigate genetic structure and diversity at both large and small geographical scale to get an estimate of the scale over which gene flow can happen was conducted in Washington and Oregon (Monsen & Blouin, 2004). The results indicate a strong pattern of isolation by distance over the entire species range and a sharp drop in number of migrants exchanged between sites that are farther than 10km away from each other.

There have also been instances where molecular analyses revealed low phylogeographic structure (Vásquez, 2013). In *Rhinella arunco*, analysis of 919 bp from the control region revealed that three of the four haplogroups overlap in geographic distribution. The intra haplogroup variance was almost comparable to the inter haplogroup variance. This pattern is in contrast to closely related species *Rhinella spinulosa* which exhibits a high degree of population structure (Correa et al., 2010). Further, rivers and watersheds which function as barriers to gene flow to certain fish species do not act as barriers to *R. arunco*.

As these studies indicate, different species of frogs seem to have differing patterns of spatial distribution of genetic diversity. Some have pronounced population structure even at short geographical scales, while other do not exhibit clear phylogeographic structure even over large distances. There are also cases where the genetic structure strongly conforms to an Isolation by Distance model (Hoffman & Blouin, 2004).

***Philautus* in Northeast India**

Philautus are a genus of bush frogs belonging to the family Rhacophoridae found in South and Southeast Asia. In India, they occur in the Western Ghats (recently renamed as *Raorchestes* (Biju & Bossuyt, 2009)) and in the Northeast states. *Philautus* are a highly speciose genus that display a very high degree of endemism. They form almost one-sixth of all extant frog species in India (Dinesh et al., 2015). Due to high variability in external appearance, identifying them to species level using traditional morphometric approaches is difficult. Therefore, it is essential to analyse molecular data for better understanding of the dynamics within this genus.

Previously, molecular analyses of *Philautus* using mitochondrial DNA in Western Ghats and Sri Lanka confirmed that this genus consists of a very high number of genetic lineages (Biju & Bossuyt, 2009; Meegaskumbura et al., 2007).

The North-east (NE) region of India can be physiographically categorized into the Eastern Himalayas (Sikkim to Arunachal Pradesh), the Northeast hills (Meghalaya, Manipur, Tripura, Mizoram, Nagaland) and the Brahmaputra Valley plains (Assam). The steep topography and location at the confluence of the Indo-Malayan, Indo-Chinese and Indian biogeographical realms, have resulted in a profusion of habitats hosting a diverse biota with a high level of endemism (Chatterjee, 2006).

The study area, Eaglenest Wildlife Sanctuary (EWS, Athreya, 2006) is spread over 218 km^2 in West Kameng District, Arunachal Pradesh (Figure 1). Rainfall varies from about 1500 mm on the northern slopes to over 3000 mm on the southern slopes (Choudhury, 2003). Set in the Eastern Himalayas, these hills rise from 100 m up to 3250 m (Agarwal, Mistry, & Athreya, 2010). Because of this altitudinal range, there are diverse habitats in EWS which host a rich faunal assemblage.

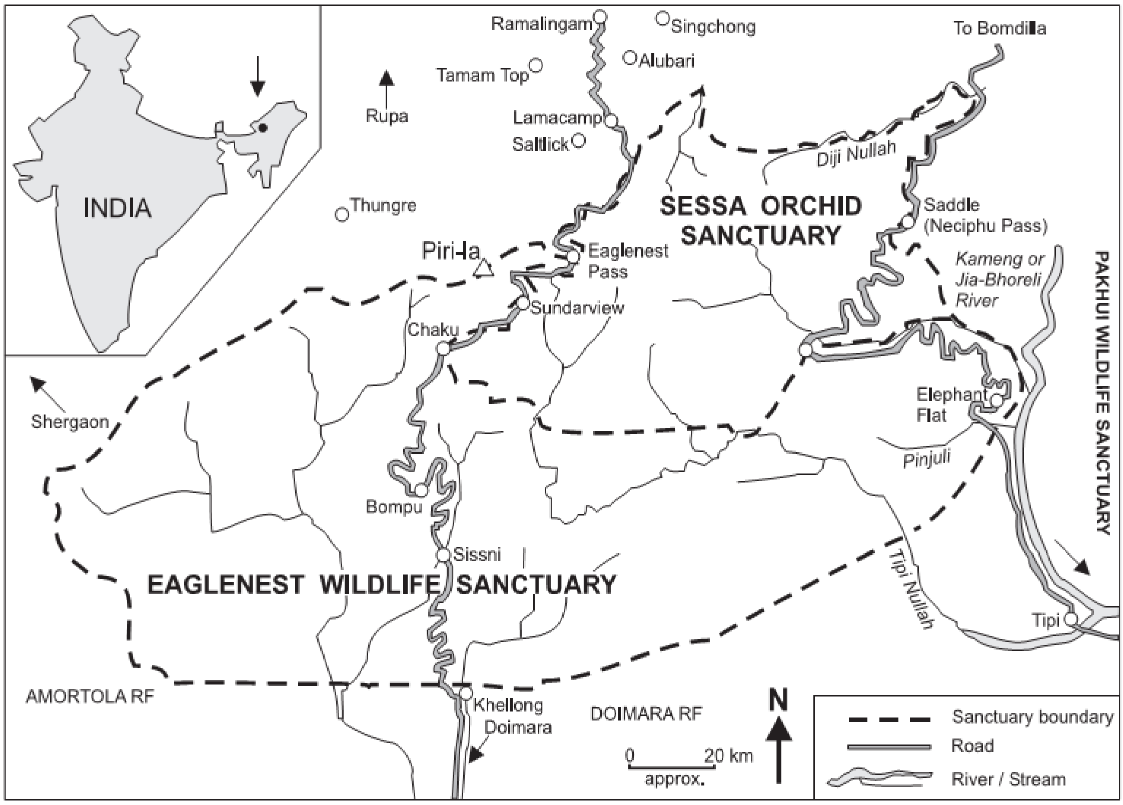

Figure 1: Location of the study site - Eaglenest Wildlife Sanctuary in India Image taken from (Agarwal et al., 2010)

Philautus are found in Eaglenest Wildlife Sanctuary from below 500m till almost 2400m above sea level (asl) in a continuous landscape where their movement is not expected to be hindered in the absence of obvious physical barriers. Elevation gradients are ideal settings to study speciation. Formation of new species in the absence of geographic isolation occurs when natural selection is strong to counteract the effects of gene flow which serves to generate genetic variation (Cadena, 2013). The change in different environmental variables along an altitudinal profile presents a wide variety of niches to organisms where natural selection could potentially act, leading to adaptation to specific niches. Population differentiation in response to varying selective pressures leads to adaptive evolution and speciation along elevation gradients (Lansing, 2005). Speciation events occur gradually over a period of millions of years without immediate reproductive isolation. Populations adapted to different elevations are likely to be in the early stages of speciation and still undergoing substantial gene flow. Estimating the extent and timing of divergence at the species and population levels can help us understand the dynamics of populations diverging into species better (Martin et al., 2013).

Motivation of this study

In this study, we carried out a preliminary analysis of genetic structure within the *Philautus* genus across a 2000m elevation gradient in Eaglenest Wildlife Sanctuary. We collected a large sample of *Philautus* in 2012, which were then assigned to six different morphogroups based on external appearance. Using eighteen different body measurement variables, a discriminant analysis was performed using the six morphotypes as grouping parameters. The morphogroups showed a reasonable degree of divergence and a strong pattern of segregation by elevation as shown in Figure 2.

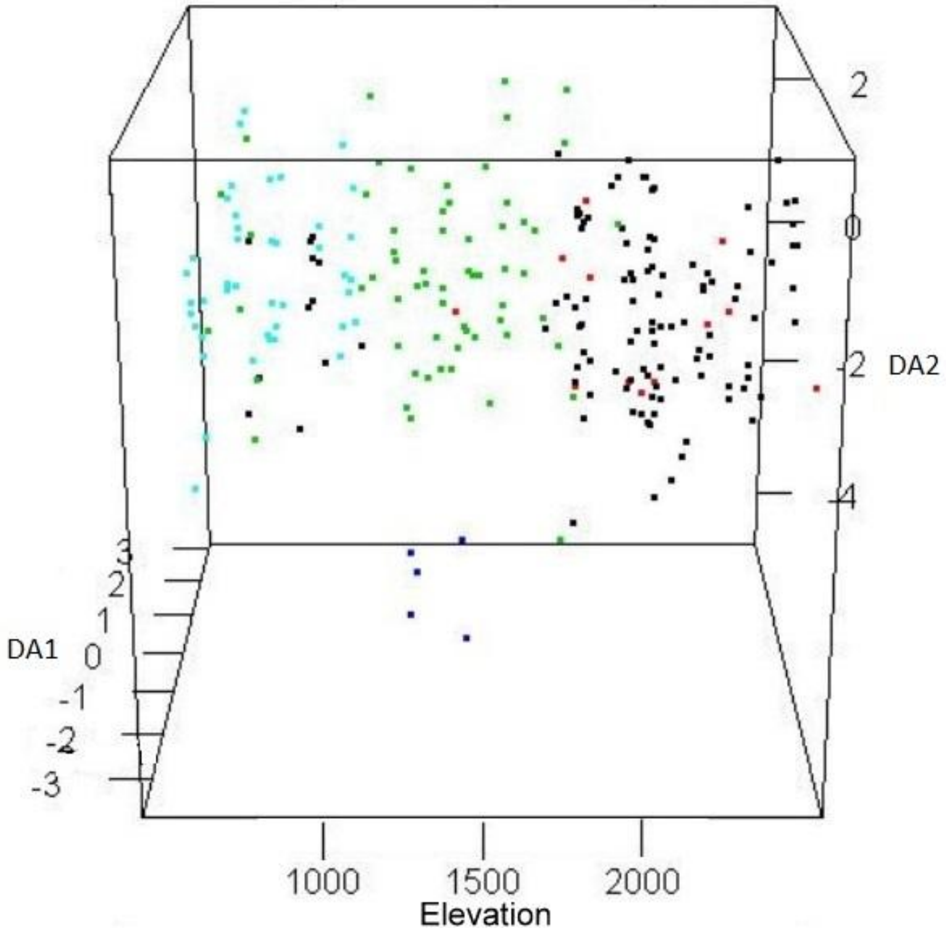

Figure 2: Discriminant Analysis plot for Morphometric separation of morphotypes based on external appearance (Mishra & Athreya, 2012; IISER, Pune BIO 310 poster presentation). Colours represent the different morphotypes.

Therefore, we undertook a genetic analysis of this sample to delineate the species boundaries and understand the correspondence between genetic and morphometric groups. It is hoped that an in depth study of this nature will help to understand the nature and role of evolutionary forces in generating and maintaining diversity in a landscap devoid of physical barriers.

METHODS

Sampling

During the months of May-June 2012, I collected specimens of *Philautus* from the southern slope of Eaglenest Wildlife Sanctuary across the altitudinal range of 500-2400m elevation. We used *Philautus* for our study because it is perhaps the most abundantly found genus in EWS (hence, not a conservation concern) and can be easily located due to distinct calls and habit of sitting up on roadside bushes. We did not obtain any specimens from above 2400m asl in spite of intensive sampling efforts. Logistical issues precluded the collection of specimens below 500m. As this was the first detailed study of *Philautus* across an elevational gradient in North-east India, we did not have an estimate of the species elevational range and width of the transition, especially as this is a cryptic group. Previous reports suggested that the number of species in other genera in the same region such as *Rhacophorus* and *Megophrys* in the same region ranged between three and five species (Athreya, 2006) which suggested an average species elevational range of about 400-600m. Therefore, to study transition regions between species ranges we chose 100m elevational strips as sampling units and collected 10-15 specimens in each strip. We had expected to have between 30 and 60 individuals per species for comparative analysis.

The specimens were captured by following the calls of the males mostly on accessible areas along the trails. The individuals were euthanized using MS-222. Tissue sample for molecular analysis was preserved in 99% ethanol. The specimens were fixed in 99% ethanol and later preserved in 70% ethanol.

Morphometric measurements for eighteen continuous characters (length measurements) were recorded in the field in 2012 using digital callipers (0.01mm least count). These characters were re-measured in the lab in Pune in 2012 for consistency of the measurements. It was found that the accuracy of the measurements was about 0.5mm though the least count of the instrument was 0.01mm. The characters used were Head Length (hl), Head Width (hw), Inter-Narial distance (in), Anterior Inter-Orbital distance (ioa), Posterior Inter-Orbital distance (iop), Eye-Nostril distance (en), Eye-Snout Distance (es), Eye-Diameter (ed), Eye-Tympanum Distance (et), Tympanum Diameter (tym), Humerus Length (h), Radius-

Ulna Length (r), Longest Finger Length (lf), Snout-Vent Length (sv), Axilla-Groin Distance (ag), Femur length (f), Tibia length (t), Longest Toe length (lt). Traditionally, these are the measurements used for morphometric analyses of frogs e. g. (Biju & Bossuyt, 2009). Certain characters such as the length of digital pads are not recorded because the small size of *Philautus* made it difficult and prone to large errors.

Genetic Analysis

At the beginning of this project (July to September 2014) we unsuccessfully tried to standardise a method for library preparation of Next Generation Sequencing (NGS) in order to get sequence data for the entire mitochondrial genome and sections of the nuclear genome (details in Appendix A). Subsequently, we restricted the analysis to a single marker corresponding to the 12S-16S region for which the DNA sequences had been obtained in 2013, even though it was suboptimal for the kind of questions we intended to address.

The DNA isolation, PCR and sequencing were done in 2013 by Ms. Gauri Keskar, a former member of the lab. Genomic DNA isolation was done using phenol chloroform method. A region of the mitochondrial genome consisting of the 12SrRNA, tRNA – Valine and 16SrRNA genes was amplified using the primers - P2: GAAGAGGCAAGTCGTAACATGG and P4: GACCTGGATTACTCCGGTCTGA (from Wilkinson et al., 2002). The expected length of the amplicon was close to 1400bp. Sequencing was done at 1st BASE using the BigDye Terminator v3.1 cycle sequencing kit chemistry.

The sequencing effort provided 173 samples. A 400bp region of good base calls common to almost all sequences was sufficient to identify the species level identity of the specimens. However, for a haplotype analysis, fragments of greater lengths and superior quality are required. Ambiguous signals of nucleotides in the sequences could lead to erroneous inferences such as overestimation of number of haplotypes. An initial analysis based on the base calls from the sequencing facility showed 53 haplotypes in the sequence data. On closer examination, it was found that a number of identified haplotypes were artefacts due to ambiguous base calls.

I re-examined the chromatograms for the sequences in Geneious (Kearse et al., 2012) and manually removed the ambiguous nucleotides. One set of sequences had a 7-8 bp long run of Cytosine bases, followed by low quality of sequences downstream due to enzyme slippage.

To improve the analysis throughput and to use an objective criterion for selecting bases by quality, we developed a script in R (R development team, 2008) to estimate the confidence of the base call at a particular nucleotide site using the ratio of the fluorescence intensity of the strongest base to the next strongest one. The workflow going into preparing the final dataset is detailed in Appendix B and Appendix C.

The distribution of ‘good’ sequence lengths is shown in Figure 3. Based on this, we identified a final dataset of 120 sequences with a length of 694bp.

The number of sequences from each 100m elevation strip going into the final dataset was not uniform because some of the specimens did not yield a sequence while others were shorter than 694 bp. It was seen that many of the sequences exclude were preferentially from a particular genotype. The number of sequences at various stages of the study available to us within each elevation strip is listed in Table 1.

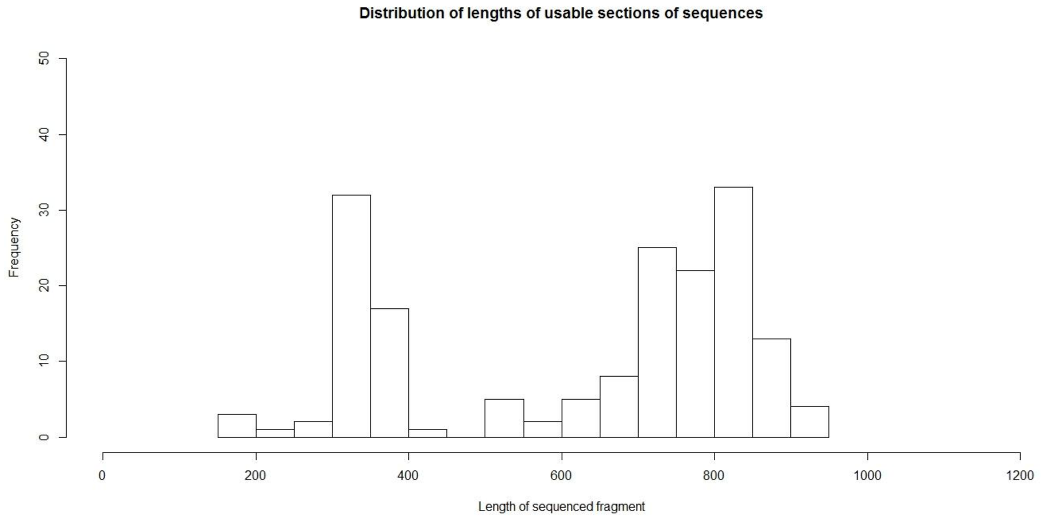

Figure 3: Distribution of lengths of usable sections of sequences after initial quality check

Alignment was done using Clustal W (Thompson et al., 2009) algorithm implemented in MEGA 6.0 version (Tamura et al., 2013). Analyses of genetic diversity were done using DnaSP software (Rozas et al.,1995). Statistical Parsimony Haplotype Networks were constructed using TCS software developed by Templeton, Crandall and Sing. (Clement et al., 2000). Demographic parameters were assessed using DnaSP. Phylogenetic analyses were performed using MEGA 6.0.

Table 1: Number of samples going into final dataset from each elevation strip

Elevation Strip	Number of Samples collected in field	Number of samples sequenced successfully	Number of samples used for analysis
500-600	10	3	1
600-700	11	8	7
700-800	11	10	10
800-900	10	1	1
900-1000	10	6	0
1000-1100	10	7	6
1100-1200	10	10	5
1200-1300	10	10	10
1300-1400	11	10	10
1400-1500	11	11	11
1500-1600	10	10	8
1600-1700	10	10	7
1700-1800	22	20	13
1800-1900	21	4	2
1900-2000	26	23	14
2000-2100	10	9	6
2100-2200	10	8	5
2200-2300	10	7	2
2300-2400	10	6	2
Total	233	173	120

RESULTS AND ANALYSIS

Phylogenetic Analysis

The 120 sequences comprised 12 different haplotypes. A Maximum Likelihood Tree of these 12 haplotypes was constructed using MEGA 6.0. GTR + G + I (General Time Reversible Model with gamma distribution and invariant sites, (Tavaré 1986) was found to be the most optimal model for the data. 1000 bootstrap replicates were performed. Sequences of *Amolops* (*Amolops ricketti* KF956111)*, Nanorana* (*Nanorana pleskei* HQ324232)*, Occidozyga* (*Occidozyga martensii* GU177877) *and Xenophrys* (*Xenophrys shapingensis* JX458090) were used from NCBI database as outgroups.

There are three clear groups among the samples and these correspond to the elevations of Bompu, Sessni and Khellong (3 locations within EWS) at 1950m, 1250m and 780m respectively. The pariwise distance between them is about 50 bases in 694 bp i.e. close to 7% which is well above the separation of 5% seen in frogs. The elevational ranges for the three groups are as follows: Bompu: 1700-2400m; Sessni species: 600-1800m; Khellong species: 500-1100m.

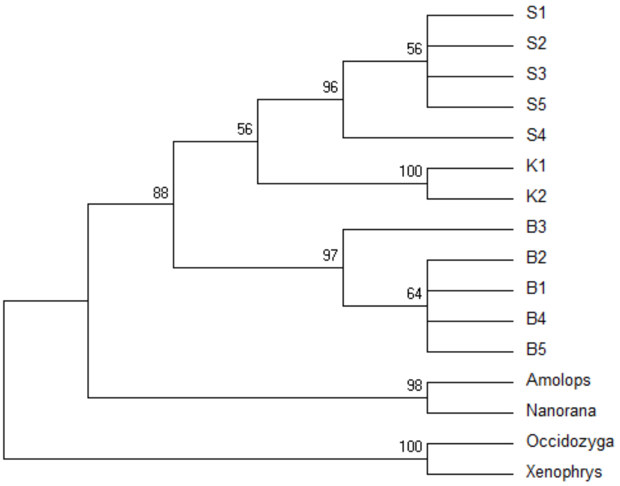

Figure 4: Maximum Likelihood Tree showing relationships for the haplotypes (1000 bootstrap replicates). The numbers on nodes indicate bootstrap support.

The Bompu species split initially and the Khellong and Sessni species diverged later (Figure 4). We also did a divergence time estimate of the species using MEGA 6.0, and a mitochondrial mutation rate of 0.025 per site per million years was used (Evans et al., 2004) and a calibration divergence time between *Nanorana* and *Amolops* lineages as 79 million years (Hedges, Dudley, & Kumar, 2006). The divergence times are 27.5 myr for the Khellong-Sessni split and 43 myr for Bompu-Sessni split.

Haplotype Analysis

Results of the haplotype analysis using DnaSP is shown in Tables 2 and 3.

Table 2: Haplotypes summary for the 120 sequences

Haplotype	Group	Number of individuals
B1	Bompu	35
B2	Bompu	2
B3	Bompu	3
B4	Bompu	2
B5	Bompu	1
S1	Sessni	11
S2	Sessni	45
S3	Sessni	2
S4	Sessni	1
S5	Sessni	2
K1	Khellong	15
K2	Khellong	1

Table 3: Summary of population genetic parameters for the three species

	Bompu	Sessni	Khellong
Number of individuals	43	61	16
Number of haplotypes	5	5	2
Number of segregating sites	4	4	1
Haplotype Diversity	0.336	0.428	0.125
Nucleotide Diversity	0.00052	0.00067	0.00018
Tajima’s D	-1.422	-0.9989	-1.16221

The Haplotype Network constructed using TCS (Figure 6) reveals three clear ‘haplogroups’ that are separated from each other by a significant number of mutation steps. A haplogroup is a group of haplotypes that are only a few mutational steps away from each other. A histogram for the genetic distances between all the samples was plotted (Figure 7) which reveals three sharp peaks corresponding to these three species which have very little genetic diversity within them as evident from the narrow distributions.

Each of the species has one dominant haplotype and a several other haplotypes that are only a few mutational steps away from the dominant one.

The distributions of genetic distances within these species are shown in Figure 7.

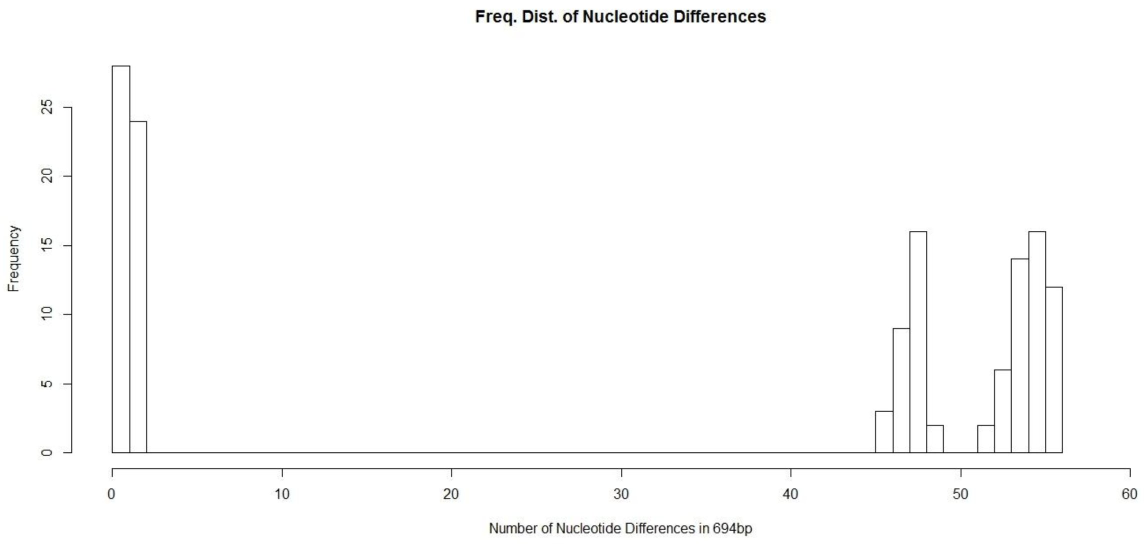

Figure 5: Histogram showing distribution of number of nucleotide differences in the 120 sequences between the 12 haplotypes

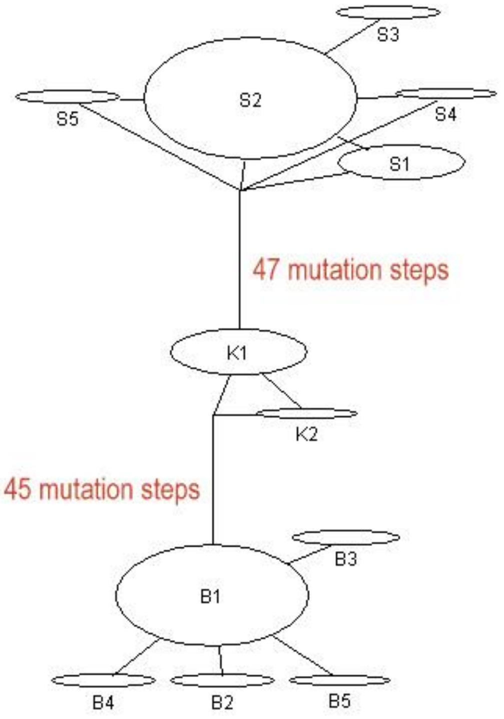

Figure 6: Median Joining Haplotype Network for the samples created using TCS software. The size of the ovals is proportional to the number of individuals the haplotype represents

Each of the three species has a clear majority of one particular haplotype and a few other rare haplotypes. The analysis reveals three distinct lineages within the *Philautus* genus with high genetic divergence between them. The three species have very little intra specific genetic diversity and have similar population genetic parameters.

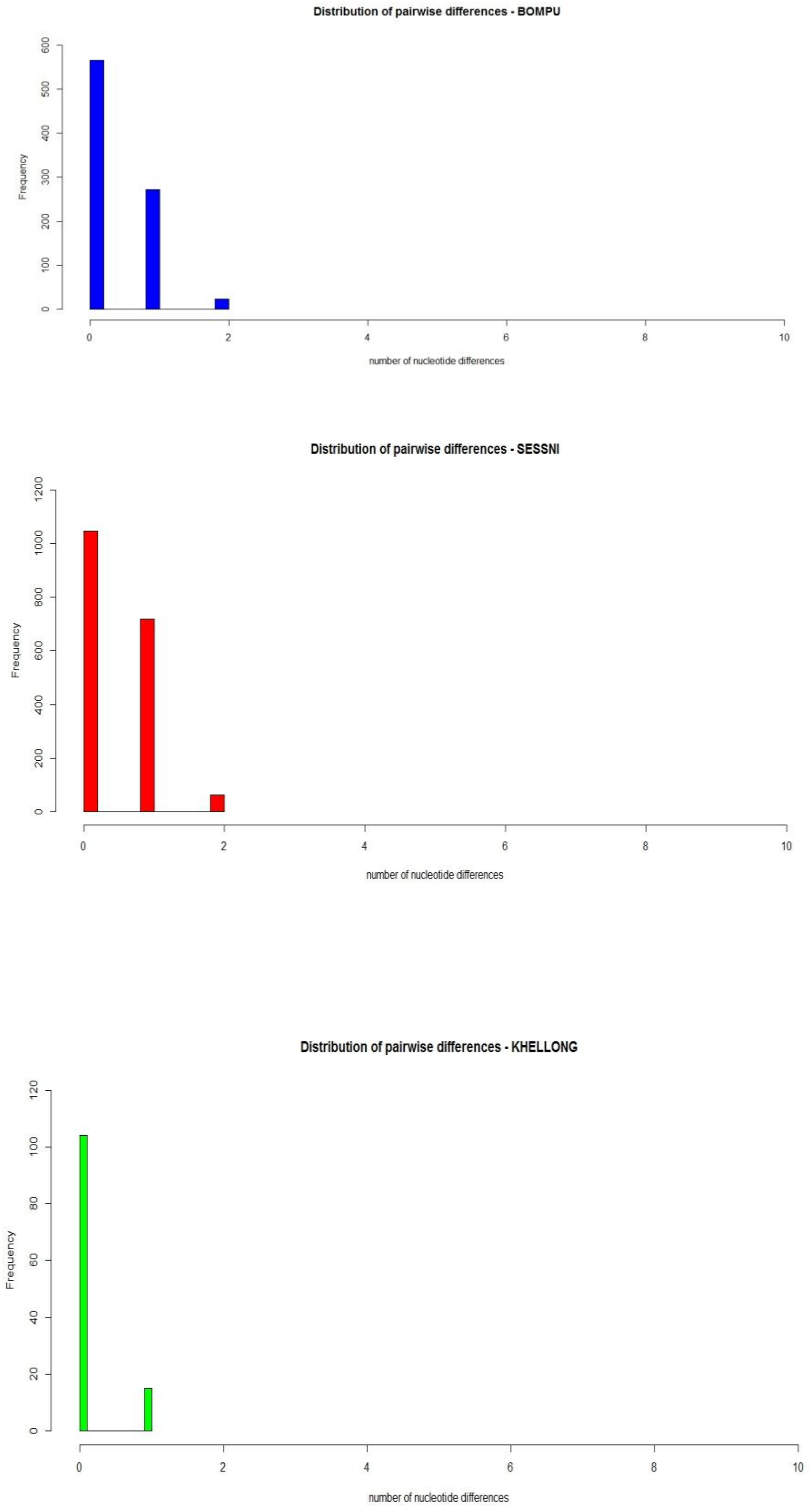

Figure 7: Distributions of genetic distances within the three species

Elevational Distribution of Haplotypes

The elevational distribution of species is shown in Figure 8. The presence of the rarer haplotypes (B2, B3, B4, B5, S2, S3, S4, S5 and K2) is indicated for each elevation strip where they are found.

There is a sharp demarcation of the altitudinal distributions between the Sessni and Bompu species. But, there is considerable overlap between the Sessni and Khellong species.

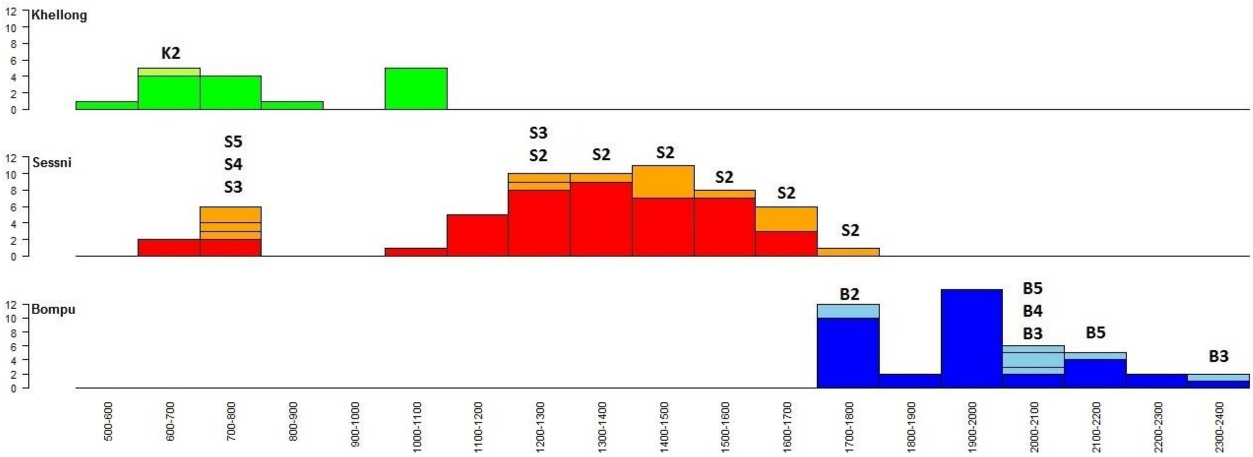

Figure 8: Distribution of the Haplotypes within 100m elevation strips. The rarer haplotypes are plotted in lighter shade and the haplotype labels are indicated above each bar. Data is missing for elevation strip 900m-1000m due to bad quality of sequences

Morphometry-Molecular data Comparisons

Discriminant Analysis was performed using the three species as the grouping parameter to see how the different species separate out in morphometric space. It may be recalled that in a previous work we had done a similar analysis but with colour forms (Figure 2). Assuming allometric growth, the value of all the variables were normalised by dividing it by the Snout-Vent Length to remove size bias due to age differences. The normalised values of the variables were used for further statistical analyses. Table 4 shows the loadings of different variables for the two discriminant axes. Plotting the first two discriminant axes shows differentiation of the Bompu haplotype from the Sessni and Khellong haplotypes, though the Khellong and Sessni haplotypes have significant overlap.

Table 4: Loadings of different morphometric variables for the two discriminant axes for separation of the three broad species

Variable	DA1	DA2
HL	2.8298	-1.9634
HW	0.22702	-10.011
IN	40.167	32.321
IOA	10.418	-1.9632
IOP	-8.2881	-17.678
EN	-33.178	34.619
ES	6.0609	-12.752
ED	6.1485	-13.432
ET	-11.368	-28.627
TYM	-4.8988	6.7255
ARM	-16.512	-3.626
RU	33.658	-10.956
LF	15.344	3.2347
AG	-0.4010	11.008
FEM	-9.2714	-4.4555
TIB	-17.438	13.692
LT	16.057	4.6348

There is differentiation of the Bompu species from Khellong and Sessni species (Figure 9). However, the separation between Sessni and Khellong species is not too clear. Along Discriminant Axis 1, the Bompu species has high values for the specimens, suggesting it has high values for Inter Narial distance and Radius Ulna length (which have high positive loadings on Discriminant Axis 1) whereas the specimens have low values for these variables in the Khellong species. Similarly, the specimens in Bompu species have low values for Eye-Nostril distance whereas specimens of Khellong have larger Eye-Nostril Distance (Eye-Nostril distance having a large negative loading in Discriminant Axis 1). There is a downward shift of the discriminant scores along Discriminant Axis 2 for Sessni species with increase in elevation in the range where the Khellong and Sessni species overlap. Plotting the discriminant scores for the two axes along with elevational in a three-dimensional plot shows a more clear segregation of the different species (Figure 10).

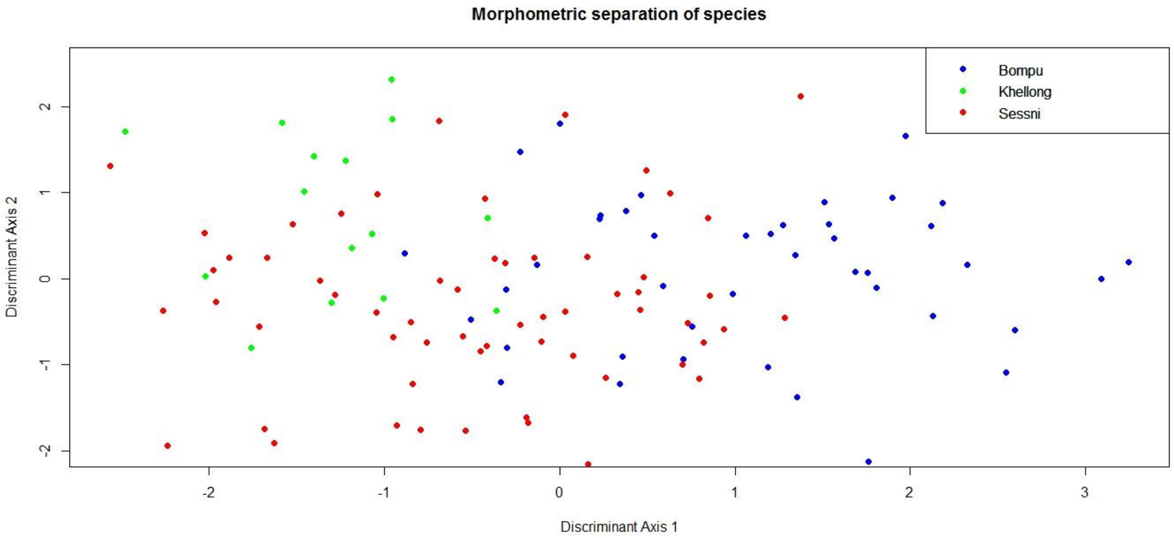

Figure 9: Discriminant Analysis plot for separation of three species in morphometric space

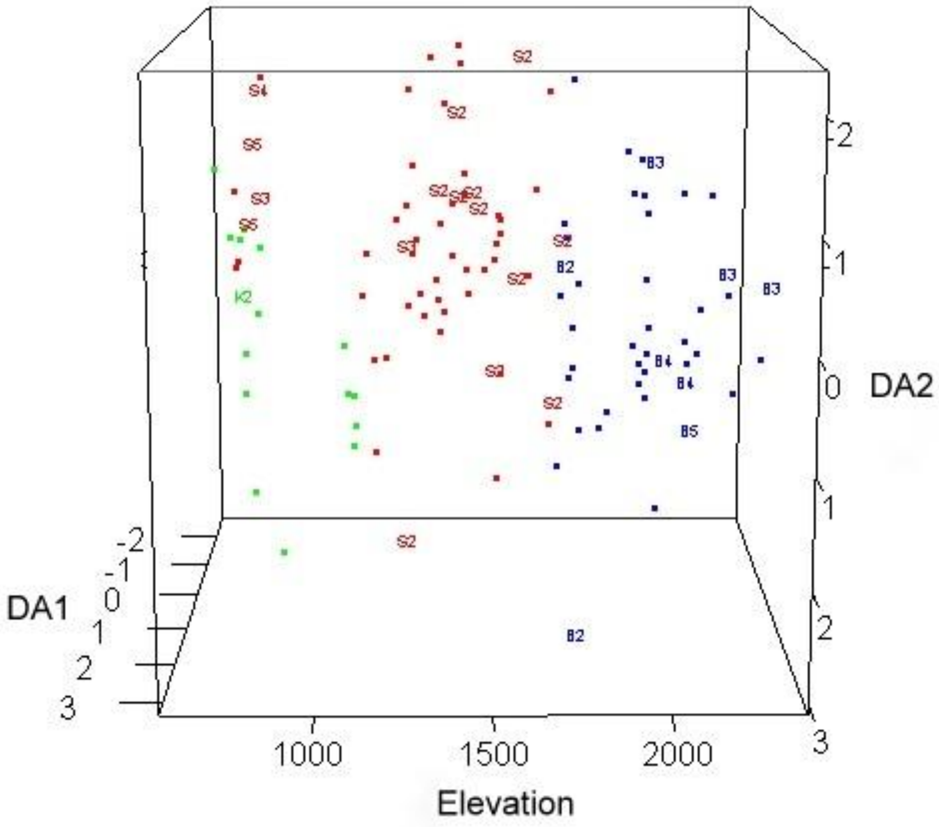

Figure 10: Three dimensional plot with discriminant axes and elevation as a third axis. The dots represent the dominant haplotypes in the three species - B1, S1 and K1. The other haplotypes are denoted in text. The species are coloured as Black – Bompu, Red – Sessni and Green – Khellong.

The samples were further divided into upper elevation group (belonging to the Bompu species) and the lower elevation group (belonging to Sessni and Khellong species) for investigating morphometric patterns after classifying the samples into broad categories based on their external appearances. The Sessni and Khellong species were grouped together because they have a significant overlap in their elevational range distribution.

The upper elevation group, which has all samples belonging to one species, could be classified into the following four groups based on appearance (Figure 11).

- yellowish overall coloration with a hourglass shaped marking on dorsal.
- Buff to yellow in colour with two visible bands running across dorsal from eye to ventral part.
- Bright yellow in dorsal colour with very faint dorsal markings if any.
- Brown to Dark Brown in dorsal colour with thin strips of a darker shade running across dorsal from eye to ventral part.

Some of the samples had confusing external appearance and could not be assigned to any broad group with confidence and hence were left out of this analysis.

Within the lower elevation group, there are samples belonging to two species. These samples were also classified into broad categories based on external appearance (Figure 12). There is very little phenotypic variation within samples belonging to the Khellong species and they can all be categorised into one group based on external appearance.

- Brown dorsal colour with hourglass shaped pattern of a lighter shade across the dorsal.
- Yellow to light brown in dorsal colour with distinct black blotch on ventral side near the hind leg
- Small sized, buff dorsal coloration with no distinct markings on dorsal
- Dark brown dorsal colour with distinct black blotch on ventral side near the hind leg.

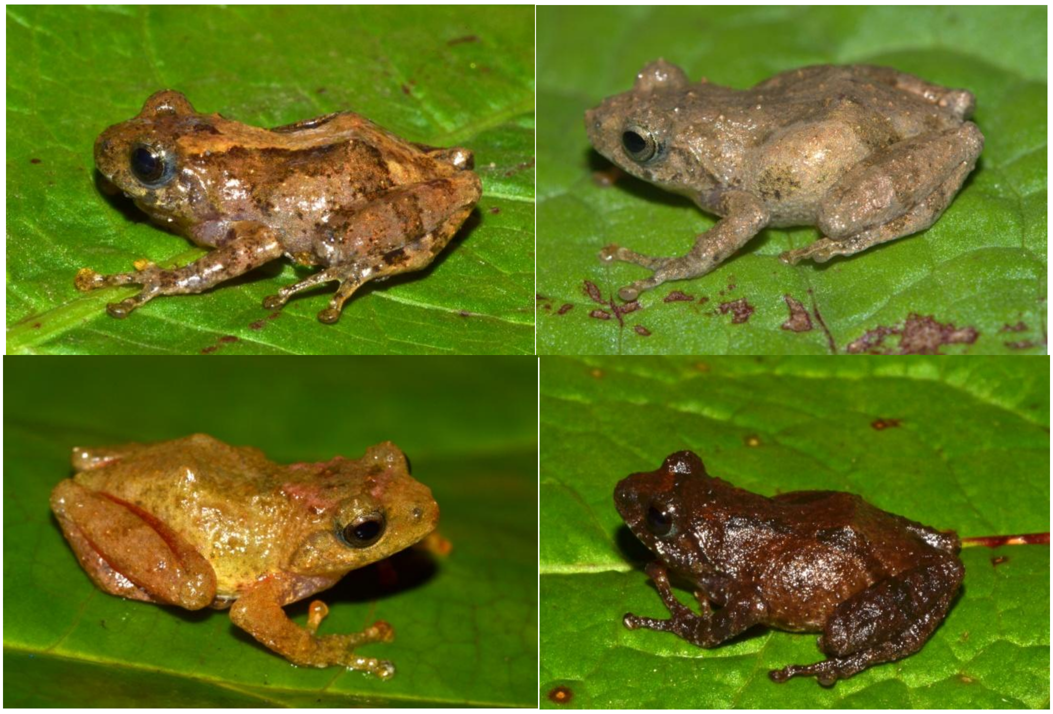

Figure 11: Colour groups in upper elevation samples. Groups 1 to 4 are shown clockwise from top left

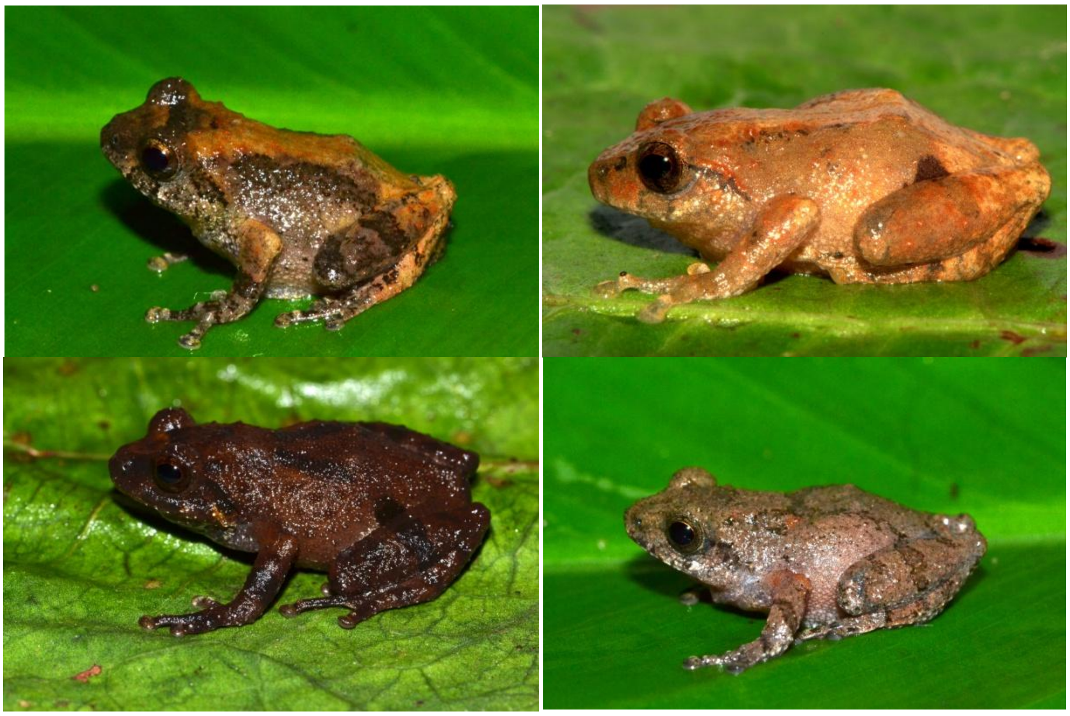

Figure 12: Colour groups in lower elevation samples. Groups 1 to 4 are shown clockwise from top left.

Table 5: The loadings of variables for the discriminant axes for morphometric separation between groups in the upper elevation samples.

Variable	DA1	DA2
HL	187.57	99.689
HW	18.723	-85.434
IN	106.49	6.2511
IOA	78.29	-118.98
IOP	-2.8177	55.975
EN	16.939	120.71
ES	56.768	-31.671
ED	93.275	-130.73
ET	-284.16	-171.98
TYM	286.62	276.39
ARM	-241.72	-3.6826
RU	-172.73	-199.96
LF	177.95	98.256
AG	-86.807	-58.852
FEM	11.468	-51.765
TIB	-29.864	74.948
LT	86.995	-6.0646

The Discriminant Scores were plotted for first two axes (Figure 13). There is clear morphometric separation between the different groups based on external appearance and they have significant divergence in terms of body dimensions even though there is little or no genetic variation (all samples belong to the same species). Based on loadings of the different variables (Table 5), samples in Group 4 can be separated from the remaining three groups by a combination of the large values for Head Length, Inter-narial distance, tympanum diameter and small values of Eye-Tympanum distance and the fore limbs (both Radius-Ulna and Arm lengths). Group 1 differs from the other groups along the second Discriminant Axis as a combination of large values of Radius Ulna, Eye Diameter, Eye-Tympanum distance and Anterior Inter-orbital Distance and small values of Tibia length, Tympanum diameter and Eye-Nostril distance.

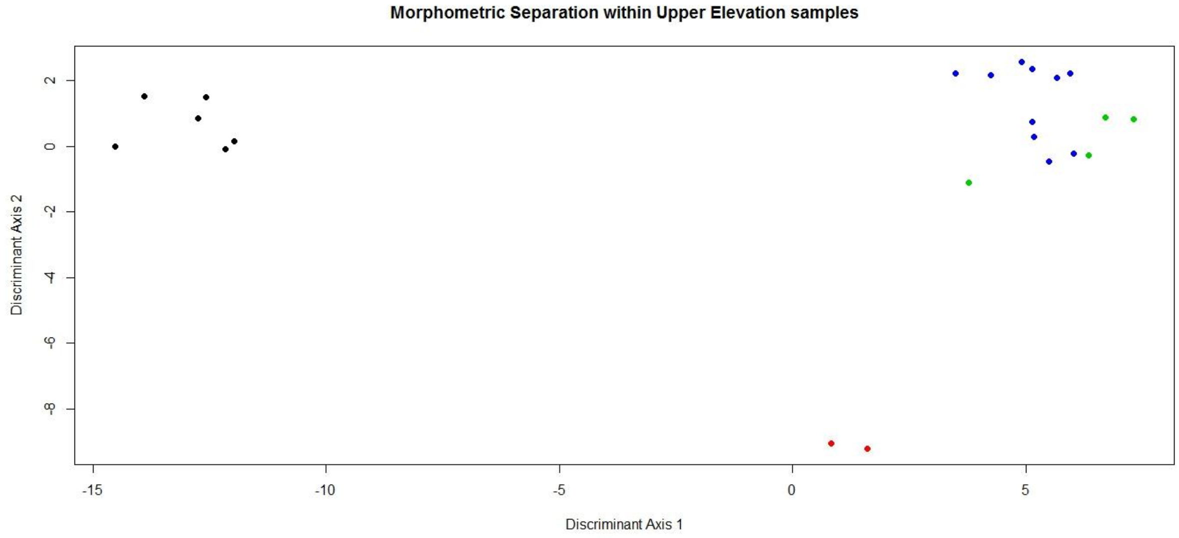

Figure 13: Discriminant Analysis plot showing separation within the upper elevation samples. Colour codes on plot represent the groups as: Group 1-Red, Group 2-Blue, Group 3-Green, Group 4-Black

There is a lot of overlap in morphometric space in the lower elevation species (Figure 14) and there is little divergence with regard to body dimensions. Even though samples belonging to Group 1 have a distinct external appearance which is different from other groups, they do not show any separation in the morphometric space. Based on loadings of the variables (Table 6), the Group 3 which consists of samples belonging to the Khellong species can be differentiated from the rest by a combination of large values for Inter-narial distance, Eye-Snout distance, diameter of Tympanum and length of Tibia and small values for Eye-Tympanum distance and Longest Finger.

Table 6: The loadings of the variables for the discriminant axes for morphometric separation between groups in the lower elevation samples.

Variable	DA1	DA2
HL	13.083	3.9527
HW	1.4747	13.097
IN	-52.215	31.837
IOA	23.234	1.1483
IOP	0.94745	-4.5377
EN	-11.67	-39.977
ES	-25.769	34.312
ED	-8.6675	32.947
ET	24.122	4.6877
TYM	-24.114	-29.027
ARM	-17.945	-17.194
RU	19.632	35.835
LF	45.47	-22.686
AG	-3.1217	-12.151
FEM	11.116	-13.071
TIB	-24.548	1.3738
LT	-2.2073	13.282

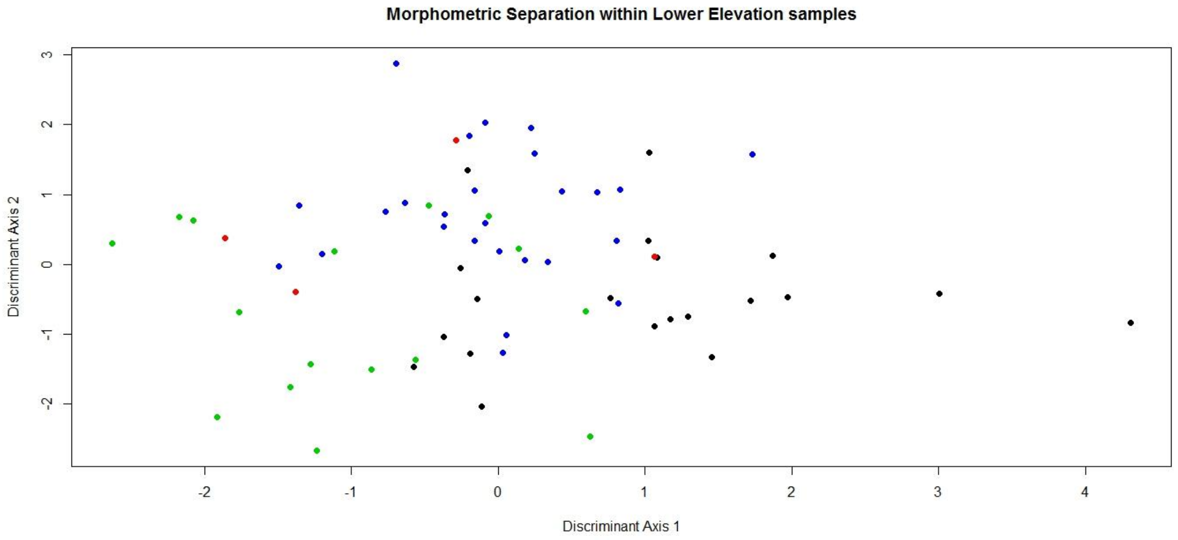

Figure 14: Discriminant Analysis plot showing separation within the lower elevation samples. Colour codes on plot represent the groups as: Group 1-Red, Group 2-Blue, Group 3-Green, Group 4-Black

DISCUSSION

This study reveals genetic structure within the *Philautus* genus in EWS. Though the results have been drawn from a single mtDNA marker, we believe the magnitude of divergence is sufficient to claim that there are three different species in the genus. Within the groups, the genetic variation is very low, for which elevation specific selective sweeps could be a plausible explanation. Selective Sweep is a process of reduction of nucleotide variation in neighbouring regions of a mutation in a DNA sequence due to recent and strong positive natural selection. Selective Sweeps in mitochondria often lead to reduced genetic diversity. Selective Sweep is an example of positive selection, which leads to adaptive evolution and consequently, speciation if it leads to reproductive isolation (Bossuyt & Milinkovitch, 2000). When selective pressure acts on the phenotype of an individual, it leads to changes in the underlying genetic make-up of the organism (Graham & Wilson, 2012). Fixation of beneficial alleles and removal of deleterious ones in the context of the organism's niche thus leads to reduction in the observed genetic diversity.

In a neutrally evolving population, the standing genetic variation results in roughly a normal distribution of the genetic distances and a star-shaped haplotype network. When histograms of the genetic distances within each species were plotted, we get distributions that are highly positively skewed. In a population with high genetic variation, one would expect roughly normal distribution of genetic distances within the population. Negative values of Tajima's D for all three species suggest purifying selection acting on them, which removes genetic variations introduced into the population through drift and mutations.

The sharp altitudinal range separation between the Bompu and Sessni species indicates that even in absence of physical barriers, abiotic factors could restrict ranges of species and be the driver for maintaining sharp species boundaries.

The three genetic groups have a significant level of morphometric divergence between them, though the Sessni and Khellong species have a lot of overlap. Sessni species showing a decreasing trend along Discriminant Axis 2 (Figure 10) with increase in elevation could be a case of character displacement due to competition. Plasticity in phenotypic characteristics helps the species to survive better when there is competition.

The different colour groups within the Bompu species show high morphometric divergence even though they are all very closely related genetically. This suggests that there has been considerable phenotypic divergence without much change in the genetic make-up at the mitochondrial locus used in this study. An alternate explanation could be that there is phenotypic plasticity in spite of similar genotypes. There could be reproductive isolation between these different groups, thereby forming reproductively isolated populations. Over evolutionary timescales, these different groups have the potential to accumulate mutations and diverge at the molecular level. Information on their ecology and habitat use could also reveal if ecological divergence accompanies phenotypic divergence.

The lower elevation species (Sessni and Khellong species) have little morphometric divergence in spite of them being genetically divergent. This result is in contrast to the results obtained for the upper elevation samples, where there was morphometric divergence of similar genotypes. Sequence data from multiple markers could possibly help us explain these observations better, because the particular locus that has been genotyped for this study could be unique in being divergent between these two species.

The genetic divergence between the different lineages is almost 7% which is sufficient to designate them as three different species. Inspite of unclear morphometric separation, there are certain characteristics that could be diagnostic to keying out specimens in the field which do not require molecular data. Specimens belonging to the Bompu and Sessni lineages have a mean SVL of Bompu – 20 ± 2.9 mm and mean SVL of Sessni – 20 ± 2.3mm). However, the individuals in the Sessni lineage have a black stripe across their lower belly on the ventral part (sometimes coming up to the dorsal part) which is not present in the Bompu lineage. The Khellong lineage comprises individuals which are slightly smaller (mean SVL = 18 ± 0.9 mm) and do not have any specific markings on their body.

There are some similarities of the results from this study to what is known of the *Eleutherodactylus* genus from Central America. *Eleutherodactylus* is the most species rich genus among all vertebrates (Crawford & Smith, 2005). The genus has direct developing frogs that do not undergo a tadpole phase in their development. The degree of endemism is so high in *Eleutherodactylus* that the location of capture is many times enough to identify specimens to the species level. The genus has almost 700 reported species and species still being discovered with new locations being explored. This genus of frogs is found from Florida in north to Ecuador in the South. In a 4 year period between 1999-2002, the number of species discovered within this genus was almost one per month. *Eleutherodactylus* are also characterised by high phenotypic variability among populations and little morphometric divergence among species. They are mostly arboreal species with large digital tips.

Based on findings from Western Ghats in India and from Sri Lanka, *Philautus* are also a highly species rich genus with very high degree of endemism and geographical structure over small geographical scales. Other aspects of their biology in which they resemble *Eleutherodactylus* are arboreal habitat, direct development without a tadpole stage, high phenotypic variability within populations and little morphometric separation between species.

Data on abiotic variables could possibly explain which environmental components are vital in deciding the niches of the species. Due to bad quality of sequences, certain samples had to be left out of the analysis. As a result, the representation of samples from below 1000m asl was not adequate. Re-sequencing for these samples could possibly reveal novel haplotypes belonging to the lower elevations and also a better understanding of the overlap in altitudinal distribution between the Sessni and Khellong samples.

Conclusive inferences cannot be drawn using results obtained from a single gene, because different regions of the genome are under varying selection pressure and therefore, gene trees are not always congruent. Data from more mitochondrial genes would reveal patterns of genetic differentiation at other mitochondrial loci. Because mitochondrial DNA is a single molecule that does not undergo recombination, it is expected that the results would more or less be congruent across different loci with variations in the scale of difference depending on the strength of selection acting on each locus.

Further, data from a nuclear gene could confirm if the low genetic diversity is due to a strong mitochondrial selective sweep or due to recent population extinction. Recent population extinction would mean a low genetic diversity within the populations across all genetic markers whereas markers where selective sweep is not operational would not show low diversity.

Results from another elevational profile could reveal if individuals belonging to similar altitude ranges from over a large geographical range are all of the same genetic composition, which would mean there are very strong altitude specific selective pressures acting in the landscape. However, getting entirely divergent haplotypes in new sampling locations would suggest there is extremely high structure within the genus and there are lineages which could qualify as species over very short geographic ranges.

REFERENCES

1. Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., … Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*(2), 229–46. doi:10.1111/j.1420-9101.2012.02599.x
2. Athreya, R. (2006). Eaglenest Biodiversity Project – I.
3. Avise, J. C. (1978). Variances and frequency distributions of genetic distance in evolutionary phylads, *40*, 225–237.
4. Biju, S. D., & Bossuyt, F. (2009). Systematics and phylogeny of Philautus Gistel , 1848 (Anura , Rhacophoridae) in the Western Ghats of India , with descriptions of 12 new species, *1848*(Table 1), 374–444.
5. Bossuyt, F., & Milinkovitch, M. C. (2000). Convergent adaptive radiations in Madagascan and Asian ranid frogs reveal covariation between larval and adult traits. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6585–90.
6. Cadena, C. D. (2013). Ecological speciation along an elevational gradient in a tropical passerine bird ?, *26*, 357–374. doi:10.1111/jeb.12055
7. Canestrelli, D., Bisconti, R., Sacco, F., & Nascetti, G. (2014). What triggers the rising of an intraspecific biodiversity hotspot? Hints from the agile frog. *Scientific Reports*, *4*, 1–9. doi:10.1038/srep05042
8. Chapple, D. G., Hoskin, C. J., Chapple, S. N. J., & Thompson, M. B. (2011). Phylogeographic divergence in the widespread delicate skink (Lampropholis delicata) corresponds to dry habitat barriers in eastern Australia. *BMC Evolutionary Biology*, *11*(1), 191. doi:10.1186/1471-2148-11-191
9. Cheviron, Z. a, & Brumfield, R. T. (2009). Migration-selection balance and local adaptation of mitochondrial haplotypes in rufous-collared sparrows (Zonotrichia capensis) along an elevational gradient. *Evolution; International Journal of Organic Evolution*, *63*(6), 1593–605. doi:10.1111/j.1558-5646.2009.00644.x
10. Choudhury, A. (2003). Birds of Eaglenest Wildlife Sanctuary and Sessa Orchid Sanctuary , Arunachal Pradesh , India, *19*, 1–13.
11. Correa, C., Pastenes, L., Sallaberry, M., & Veloso, A. (2010). Phylogeography of Rhinella spinulosa (Anura : Bufonidae) in northern Chile, *31*, 85–96.

12. Crawford, A. J., & Smith, E. N. (2005). Cenozoic biogeography and evolution in direct-developing frogs of Central America (Leptodactylidae : Eleutherodactylus) as inferred from a phylogenetic analysis of nuclear and mitochondrial genes, *35*, 536–555. doi:10.1016/j.ympev.2005.03.006
13. Dinesh, K. P., Radhakrishnan, C., Channakeshavamurthy, B. H., & Kulkarni, N. U. (2015). A Checklist of Amphibians of India, (January), 1–13.
14. District, W. K., Pradesh, A., Agarwal, I., Mistry, V. K., & Athreya, R. (2010, A preliminary checklist of reptiles of Eaglenest Wildlife Sanctuary *17*(2), 81–93.
15. Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, *107*(1), 1–15. doi:10.1038/hdy.2010.152
16. Elmer, K. R., Dávila, J. a, & Lougheed, S. C. (2007). Applying new inter-individual approaches to assess fine-scale population genetic diversity in a neotropical frog, Eleutherodactylus ockendeni. *Heredity*, *99*(5), 506–15. doi:10.1038/sj.hdy.6801025
17. Evans, B. J., Kelley, D. B., Tinsley, R. C., Melnick, D. J., & Cannatella, D. C. (2004). A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. *Molecular Phylogenetics and Evolution*, *33*(1), 197–213. doi:10.1016/j.ympev.2004.04.018
18. Evolutive, B. (2000). Absence of evidence for isolation by distance in an expanding cane toad (Bufo marinus) population : an individual-based analysis of microsatellite genotypes, 1905–1909.
19. Forks, G. (2001). Microsatellite variation and fine-scale population structure in the wood frog (Rana sylvatica), 1087–1100.
20. García-R, J. C., Crawford, A. J., Mendoza, A. M., Ospina, O., Cardenas, H., & Castro, F. (2012). Comparative phylogeography of direct-developing frogs (Anura: Craugastoridae: Pristimantis) in the southern Andes of Colombia. *PloS One*, *7*(9), e46077. doi:10.1371/journal.pone.0046077
21. Graham, R. I., & Wilson, K. (2012). Male-killing Wolbachia and mitochondrial selective sweep in a migratory African insect. *BMC Evolutionary Biology*, *12*(1), 1. doi:10.1186/1471-2148-12-204

22. Hebert, P. D. N., Cywinska, A., Ball, S. L., & Jeremy, R. (2003). Biological identifications through DNA barcodes, (September 2002), 313–321. doi:10.1098/rspb.2002.2218

23. Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics (Oxford, England)*, *22*(23), 2971–2. doi:10.1093/bioinformatics/btl505

24. Hickerson, M. J., Carstens, B. C., Cavender-Bares, J., Crandall, K. a, Graham, C. H., Johnson, J. B., Yoder, a D. (2010). Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution*, *54*(1), 291–301. doi:10.1016/j.ympev.2009.09.016

25. Hoffman, E. a, & Blouin, M. S. (2004). Evolutionary history of the northern leopard frog: reconstruction of phylogeny, phylogeography, and historical changes in population demography from mitochondrial DNA. *Evolution; International Journal of Organic Evolution*, *58*(1), 145–59. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15058727

26. Holsinger, K. E. (2010). Tajima ' s D , Fu ' s F S , Fay and Wu ' s H , and Zeng et al .' s E Introduction Fay and Wu ' s H.

27. Hurst, G. D. D., & Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings. Biological Sciences / The Royal Society*, *272*(1572), 1525–34. doi:10.1098/rspb.2005.3056

28. Johnson, K. P., Walden, K. K. O., & Robertson, H. M. (2013). Next-generation phylogenomics using a Target Restricted Assembly Method. *Molecular Phylogenetics and Evolution*, *66*(1), 417–22. doi:10.1016/j.ympev.2012.09.007

29. Kotaki, M., Kurabayashi, A., Matsui, M., Kuramoto, M., Djong, T. H., & Sumida, M. (2010). Molecular phylogeny of the diversified frogs of genus Fejervarya (Anura: Dicroglossidae). *Zoological Science*, *27*(5), 386–95. doi:10.2108/zsj.27.386

30. Lansing, E. (2005). The evolution of species distributions: Reciprocal transplantations across species ranges, *59*(8), 1671–1684.

31. Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies, 1817–1828. doi:10.1101/gr.159426.113.

32. Meegaskumbura, M., Manamendra-arachchi, K., Schneider, C. J., & Pethiyagoda, R. (2007). New species amongst Sri Lanka's extinct shrub frogs (Amphibia: Rhacophoridae: Philautus), *15*, 1–15.
33. Monsen, K. J., & Blouin, M. S. (2004). Extreme isolation by distance in a montane frog Rana cascadae, 827–835.
34. Nuñez, J. J., Wood, N. K., Rabanal, F. E., Fontanella, F. M., & Sites, J. W. (2011). Amphibian phylogeography in the Antipodes: Refugia and postglacial colonization explain mitochondrial haplotype distribution in the Patagonian frog Eupsophus calcaratus (Cycloramphidae). *Molecular Phylogenetics and Evolution*, *58*(2), 343–352. doi:10.1016/j.ympev.2010.11.026
35. Programme, F. C., & Delhi, N. (2006). Biodiversity Significance of North East India, 1–71.
36. Rice, A. M., Rudh, A., Ellegren, H., & Qvarnström, A. (2011). A guide to the genomics of ecological speciation in natural animal populations. *Ecology Letters*, *14*(1), 9–18. doi:10.1111/j.1461-0248.2010.01546.x
37. Rodríguez, F., Pérez, T., Hammer, S. E., Albornoz, J., & Domínguez, A. (2010). Integrating phylogeographic patterns of microsatellite and mtDNA divergence to infer the evolutionary history of chamois (genus Rupicapra). *BMC Evolutionary Biology*, *10*, 222. doi:10.1186/1471-2148-10-222
38. Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. a, … Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews. Genetics*, *15*(3), 176–92. doi:10.1038/nrg3644
39. Vásquez, D., Correa, C., Pastenes, L., Palma, R., & Méndez, M. a. (2013). Low phylogeographic structure of Rhinella arunco (Anura: Bufonidae), an endemic amphibian from the Chilean Mediterranean hotspot. *Zoological Studies*, *52*(1), 35. doi:10.1186/1810-522X-52-35

Appendix A: Standardisation of Next Generation Sequencing methods

Phylogeography mostly deals with evolutionary relationships at the population level. At this level, the fraction of base pairs in the genome that differ would be significantly less compared to the entire genome than between species. Therefore, longer DNA sequences are needed for inferring relationships. *Philautus* are characterised by high intra specific variation which almost verges on the level of inter specific variation in their phenotypic characters. The lack of morphometric divergence suggests they may have radiated recently. Lineages that have separated from each other only very recently have relatively few differences between their genome sequences. Therefore, to answer questions in recently diverged taxa, large amounts of genomic data is required, which is laborious and expensive using traditional PCR methods. Phylogenomics and Next-Generation Sequencing technologies offer promise for carrying out such studies by providing large amounts of sequence data dn low cost per base. NGS approach will not only yield a more accurate view of how the genome as a whole is evolving, but will also increase the likelihood that loci evolving at many different rates are sampled, thereby enabling us to ascertain the complex nature of evolutionary forces in speciation.

One class of phylogenomic target enrichment methods involves sequence capture of nuclear regions flanking conserved regions. Therefore, using long range PCR, the Exon Primed Intron Crossing (EPIC,) approach can be used for target enrichment. In this approach, exons (which are conserved in organisms across large timescales) are used as priming sites to obtain amplicons that include the intervening intron sequences. Exons, being protein coding sections are conserved across evolutionary timescales and are therefore ideal to use as priming sites. The introns, being non-coding are much less conserved and are therefore informative loci for phylogenetic studies at shallower timescales. An advantage of using the EPIC strategy is that the sequence gives data for both conserved and variable sections (having both the exon and intron fragments) which helps in examining genetic variation at the intraspecific and inter-specific level simultaneously, particularly helpful when studying species complex. We adapted the EPIC approach for investigating phylogeography.

Our NGS workflow comprises the following steps

- Target enrichment
    a. Identify long range primers
    b. Standardise long range PCR (5-12 kb)
- Fragmentation of PCR amplicons to a final size of 500bp, using small volumes of PCR products.
- Ligation of NGS adapters.

Under this project, I tried to standardise the fragmentation of long range PCR products under the constraint of small volumes (corresponding to PCR products) than recommended in literature.

Target Enrichment is the first step in library preparation for Next Generation Sequencing where different methods are used to capture the genomic region of interest. Widely used strategies for target enrichment include Reduced Representation Libraries and RNA-seq methods.

We adopted long range PCR in this case using the Exon Primed Intron Crossing strategy for target enrichment. Long range PCR typically refers to the amplification of DNA fragments in excess of 5 kb, using a blend of thermo-stable DNA polymerases that allows for the amplification of longer fragments than those that can be achieved with a single enzyme.

The lab had been involved in identifying long range primers for amplifying mitochondrial and nuclear loci. In a previous project (BIO 410, 2013, Mishra, Habib & Athreya), I had identified priming regions for nuclear loci to be amplified using the EPIC approach (rhodopsin, tyrosinase and BDNF). This was done by comparing full genome sequences of *Xenopus tropicalis*, *Danio rerio*, *Anolis carolinensis, Gallus gallus* and *Homo sapiens*. Other members of the lab had designed Long Range PCR primers and standardised a procedure to amplify 11kb long fragments of mitochondria.

I used the following PCR protocol to generate amplicons for fragmentation

10 ul PCR reaction mixture was used for long range amplification reactions – 5ul NEB Taq Hotstart 2X Mastermix, 1 ul of BSA (5ug/ul), 1 ul of each the primers F57 and F33 (10mM), 1 ul template (30ng/ul), 1ul PCR pure water to make up the 10ul reaction volume.

The primers used were
F57 (5’-GTGTCCCACCCHACTAGAGGAGCCTG-3’) and
F33 (5’-AACCATGGTRGYGAGGAATTAGCAGT-3’)

The positions of these primers on the frog mitochondrial genome (*Rhacophorus schlegelli* genome, which is in the same family as *Philautus*) are shown in Figure A1. The genes contained in this pair of primers are – ND4, ND4L, ND3, COIII, ATP6, ATP8, COII, COI, ND2 and ND1

PCR was performed under the following conditions - 95 ºC for 30s followed by 35 cycles of 94 ºC at 10s (denaturation), 57ºC for 60s (annealing), 65ºC for 570s (extension) followed by a final extension of 65ºC for 600s min and then hold at 4ºC.

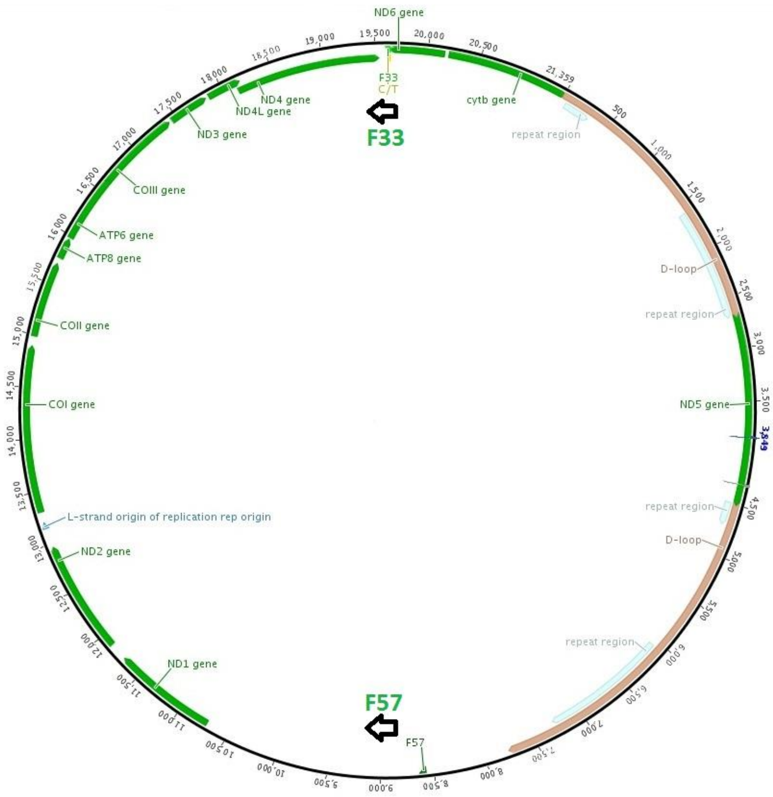

Figure A1: Positions of the primers F33 and F57 in the frog mitochondrial genome

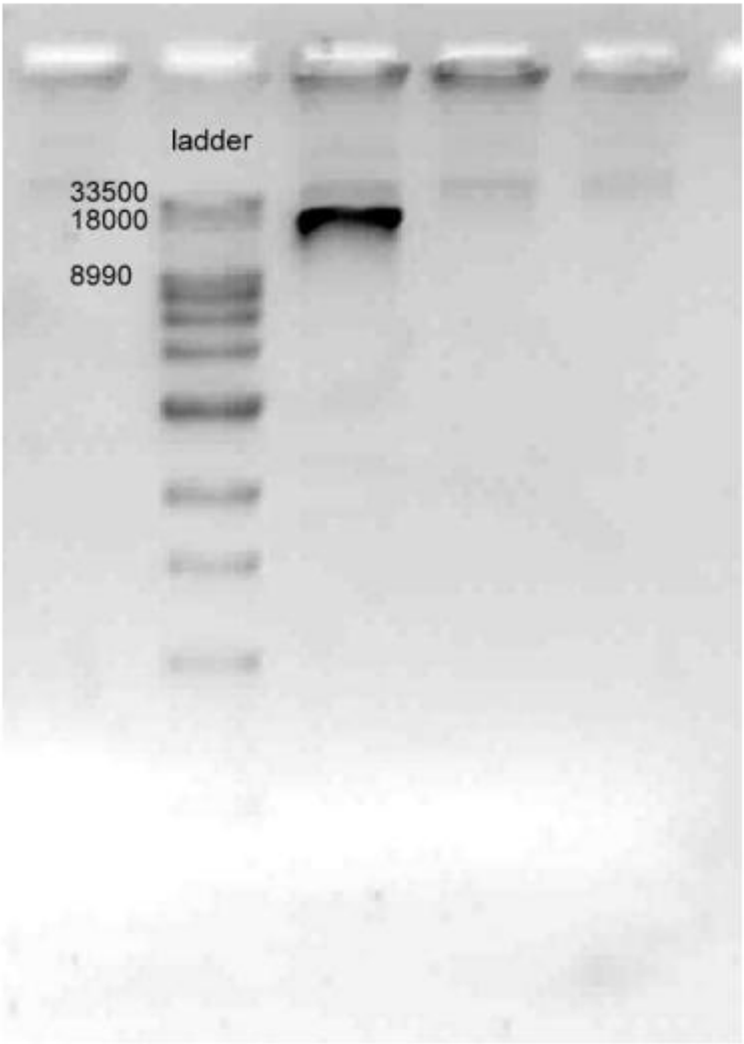

Figure A2: Gel image for PCR products using primer pair F33-F57

Initially, using the prescribed PCR mixture, amplification was not observed. However, amplification was observed after 1mM of extra Mg^{2+} was added to the reaction in form of $MgSO_4$ (Figure A2). Magnesium ions increase the activity of the polymerase. Additives can be used to increase the activity of the polymerase as one of the solutions to troubleshoot PCRs. However, addition of other additives such as Trehalose did not result in amplification.

However, the amplification of mitochondrial genome using long range PCR failed after a few rounds of successful amplification. Modifying the Mg^{2+} concentration or addition of additives such as Trehalose did not help..

The following methods were used to troubleshoot the failed PCR but to no effect:

1. Dilution of template – Using too much total DNA results in packed DNA in the confined space of the reaction vessel and can lead to false priming and even poor DNA synthesis due to the obstructed diffusion of large *Taq* polymerase molecules. Therefore, the DNA template going into the reaction was reduced to 10-15 ng by using 0.5ul of 30ng/ul template.
2. Using freshly extracted DNA as template – The quality of template should be good (devoid of nicks in the strands) for successful long range amplification. Freshly extracted DNA using charge-switch extraction protocol was used as template for the long range PCR reactions. Freezing and thawing the template DNA could lead to breaks that could compromise amplification. Successful amplification was obtained sporadically but not in a consistent manner.
3. Reduction of primer concentration – Using a primer concentration much higher than the optimum could lead to majority of primers forming dimers within themselves, causing the amplification to fail. It can also lead to non-specific binding.

Standardisation of fragmentation for small amounts of DNA: The second step in library preparation involves shearing the DNA into smaller fragments for the sequencing platform. Different sequencing platforms have different DNA fragment length requirements.

Most of the standard protocols from manufacturers prescribe use of at least 50µl of PCR product. Anticipating the processing a large number of samples and consequent need to reducing PCR costs, we decided to investigate standardisation of a fragmentation protocol involving 10ul volumes corresponding to PCR reactions. we explored the possibility. Our goal was to fragment the PCR amplicons to a distribution centred around 500bp for Illumina Mi-Seq platform.

Bioruptor uses sonication to shear DNA. The conditions that determine the length of DNA obtained after shearing are (1) length of each sonication cycle (2) number of sonication cycles (3) ratio of the PCR product and 1xTE Buffer and (4) volume of the sonication tube.

We used 10µL of PCR product and 20µL of TE buffer for fragmentation. The cycle length and number of cycles were varied. Our results suggest that an optimum distribution of fragment lengths (centred at ~500bp) is obtained using 10ul PCR product and 20ul 1x TE buffer in 0.2ml PCR tubes after sonication for 9 cycles of 30s each at low power frequency settings in Diagenode Bioruptor (Figure A3).

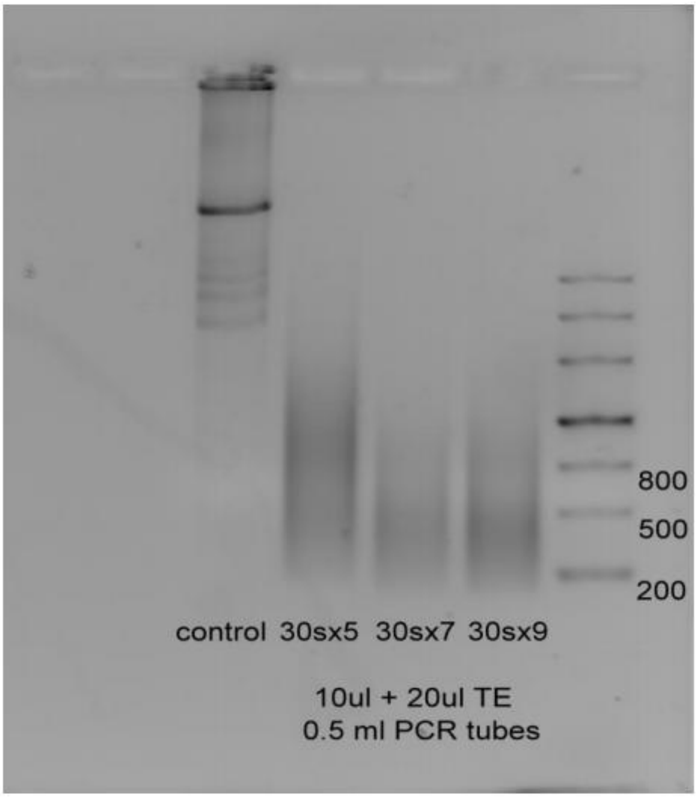

Figure A3: Fragmentation results using Diagenode Bioruptor

After fragmentation, the next step in library preparation involves the selection of fragments of a particular length (~500bp in this case as we intend on using Illumina Platform). After size selection, the fragments are ligated to adapters before sequencing. These adapters bind to the flow cell of the sequencing platform.

Gel extraction of DNA:

DNA can be extracted from the desired base pair length range by running PCR products against a ladder and excising the gel containing the DNA wherever necessary. This procedure was used to isolate and extract the required fragment length of the fragmented amplicon. This can be used as an alternative to the expensive size selection method that is a part of the NGS workflow. However, it is necessary that the DNA extracted using this method can be used for other downstream procedures. To test this, I did a series of gel extractions of a known 800bp PCR amplicon to be followed by a second PCR.

Appendix B: Preparing accurate datasets for analysis using trace files

Chromatograms obtained from the sequencer are important to analyse DNA sequence quality. Only clear sharp peaks of a single base in the chromatogram are reliable for use in analysis. If the peaks are not strong, or there is a signal for two different bases at the same position, these positions should be removed before analysis. For inferring relationships between evolutionary lineages based on DNA sequences, the DNA sequences should be very accurate. The first 25 bases or so give peaks that are crowded and not well resolved. The sequence quality starts deteriorating after approximately the first 600-900 bases. Homopolymer regions often lead to bad quality base calls on the downstream side due to enzyme slippage (Figure B2).

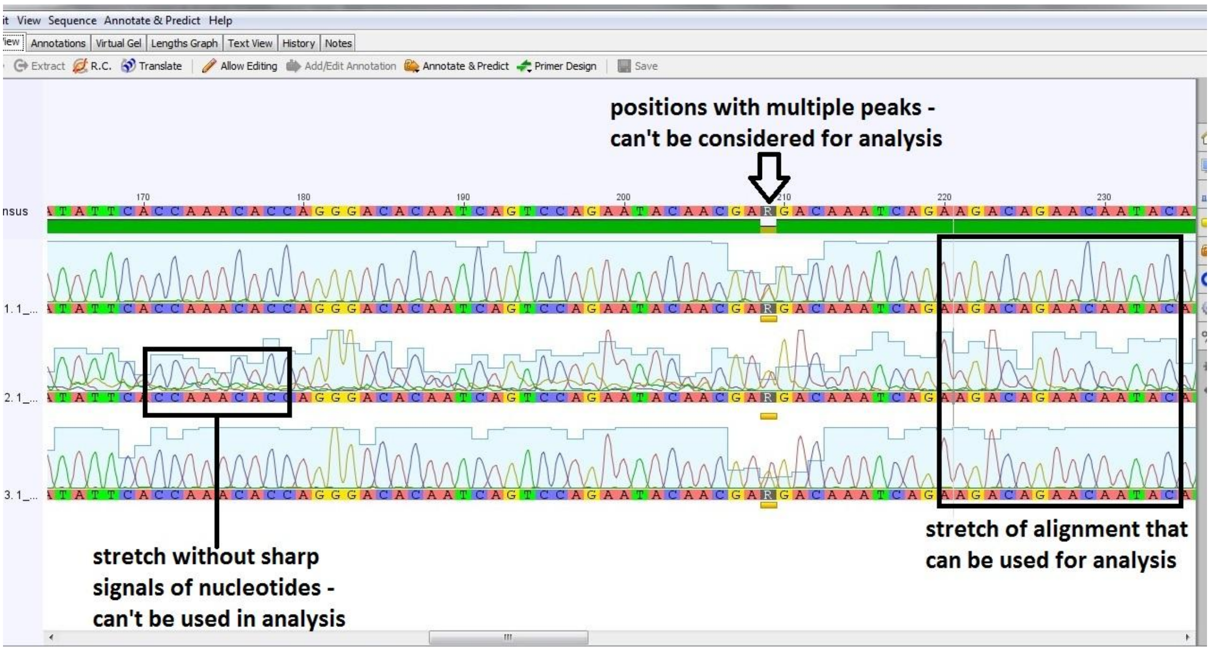
it View Sequence Annotate & Predict Help
iew Annotations Virtual Gel Lengths Graph Text View History Notes
Extract R.C. Translate Allow Editing Add/Edit Annotation Annotate & Predict Primer Design Save
positions with multiple peaks -
can't be considered for analysis
170
180
190
200
210
220
230
nsus
ATATTCACCAAACACCAGGGACACAATCAGTCCAGAATACAACGARGACAAATCAGAAGACAGAACAATACA
1. 1_...
2. 1_...
3. 1_...
stretch without sharp
signals of nucleotides -
can't be used in analysis
stretch of alignment that
can be used for analysis

Figure B1: Geneious view showing regions in the alignment that can be selected for analysis based on the chromatograms

Preparing the dataset: The chromatograms were examined by eye. Positions with strong signals for multiple nucleotides for atleast one sequence were removed. Sequences with long stretches of low quality were removed from the analysis altogether. Positions without a sharp peak for any single nucleotide were removed. Finally, the dataset prepared was 700bp long after removing positions with conflict. A sample of how to manually examine chromatograms in Geneious is given in Figure B1

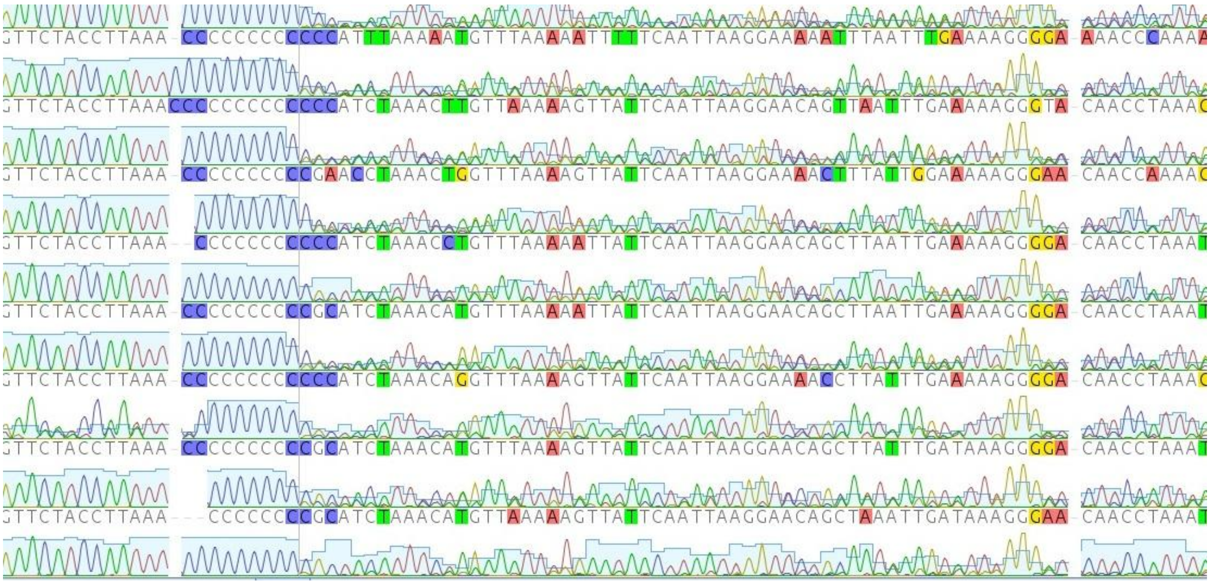

Figure B2: Cytosine homopolymer region leads to low quality sequence reads downstream due to enzyme slippage

While the visual quality checks of the base calls are feasible for small datasets, it was felt that an automated command-line program would be more suitable for high throughput analysis for large sequences. Not only would it reduce the human effort, but would also implement an objective quantitative quality criterion for identifying good base calls.

The output by the sequencer have information that can be used to construct the chromatograms corresponding to each of the four bases (A, T, G and C). Further, using R one could automate the process to find the ratio of highest peak amplitude to the second highest peak amplitude (thereby introducing an objective criterion for sequence quality check), which is a proxy for the sequence quality at that particular nucleotide position. The DATA.1 to DATA.4 data fields can be used to construct the chromatograms for bases A, G, T and C respectively. P1AM is the value of base with highest amplitude and P2AM is the value of base with second highest amplitude. The following R script was written to find the ratio of highest amplitude to second highest amplitude as indicator of sequence quality. High values of this ratio identify unambiguous base calls in the sequence.

```
> library(sangerseqR)
> x <- read.abif(file.choose())
> amp2<- as.character(x@data$P2AM.1)
> amp2[amp2 == "0"] ="1"
> seqQual=x@data$P1AM.1 / x@data$P2AM.1
```

Figure B3 shows the ratio of highest base amplitude to second highest base amplitude plotted across the nucleotide positions. Using R commands, one can zoom in on the region of interest. A filter can be implemented to select the positions that have a quality above a user-defined value. A large number of input sequences can be processed at a single time without manual intervention.

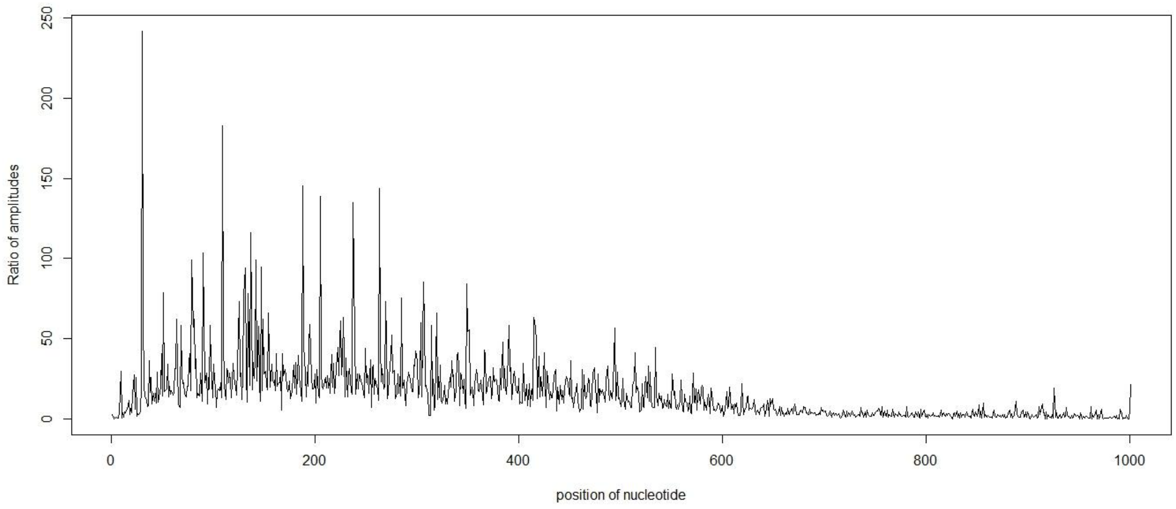

Figure B3: Plot showing ratios of highest base amplitude to second highest base amplitude as indicator of sequence quality the given nucleotide position.