Estimating Substitution Trends in the Mitochondrial Genome of Malagasy Snakes

Renee Jean-Baptiste (TYWLS), Crystal Lin (TYWLS), Odelia Lorch (HSMSE), Ava Mohajer (The Dalton School), Stephanie Pinto (Forest Hills HS), Elena Urquiola (Hunter HS)
Carolyn Sy (Helen Fellow, Education), Dr. Frank T. Burbrink (Department of Herpetology)

Background

Mitochondrion: organelle that produces ATP energy for organisms to function (*fig. 1*).

Mitochondrial genome: circular DNA in mitochondria that contain 13 protein-coding genes. (*fig.2*)

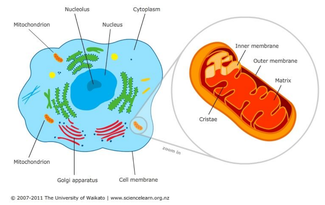


Figure 1. Animal cell and mitochondrion

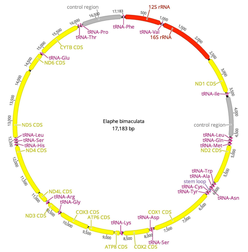


Figure 2. Snake mitochondrial genome

Control region: location at which replication begins and mtDNA begins unzipping (*fig. 3a*). This is where mtDNA spends the greatest amount of time single-stranded. Snakes have two control regions (*fig. 2*); other vertebrates have only one.

WANCY region: location at which outer strand of mtDNA begins being filled in (*fig. 3b*). This is where mtDNA spends the least amount of time single-stranded.

Substitutions: mutations that can accumulate when mtDNA is left single-stranded. Substitution rates are proportional to the amount of time left single-stranded.

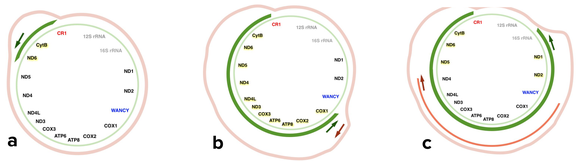


Figure 3. Replication

Tools and Methods

- Geneious
  - Assembled genomes from reads
- FUBAR
  - Calculated substitution rates and selection probabilities
- Python
  - Processed CSV files
- R, RStudio
  - Data manipulation and visualization
  - Packages: BayesianFirstAid, ape, circlize

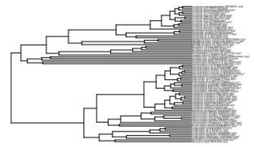
Data


Mimophis malfahensis


Madagascarophis colubrinus


Dromicodryas quadrilineatus

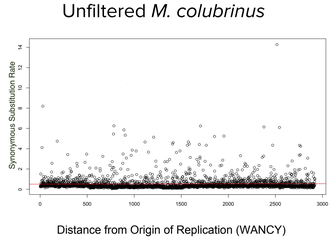
Subfamily Pseudoxyrhophiinae

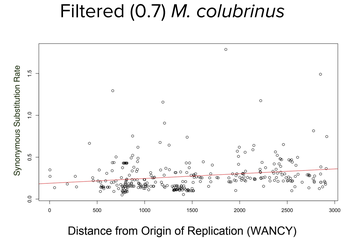


Full mitochondrial genomes of 80 snake species from Madagascar.
Taxa were studied due to their isolated population from a single historical migration.

Results

Is there a linear relationship between substitution rates and distance from control region like in other vertebrates?

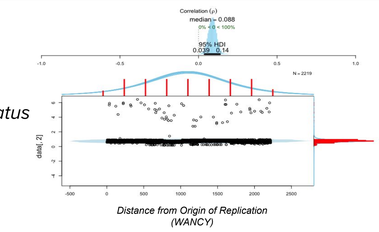


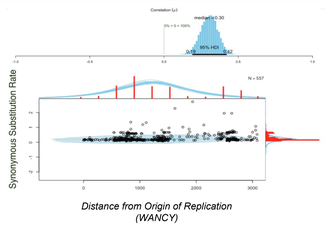


As genes get further from WANCY, we observe a positive trend in substitution rates.
Plots were filtered based on probability of positive and negative selection.

Is this a general trend? Does it occur both for individual species and between multiple species?

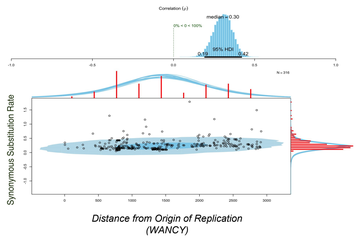
D. quadrilineatus
12 individuals
corr = 0.088

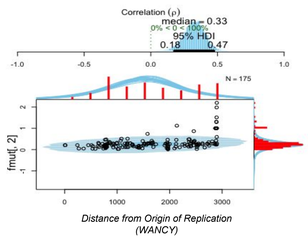




M. mahfalensis
11 individuals
corr = 0.30

M. colubrinus
19 individuals
corr = 0.30





Subfamily *Pseudoxy.*
80 species
corr = 0.33

Bayesian correlation plots were used to determine statistical significance of our data.

Remarks

- Found significant positive correlation between distance from CR and synonymous rate in all groups tested except *D. quadrilineatus*
- Interspecies shows stronger trend than intraspecies -- Bayesian correlation is higher
- Substitution rates become hard to track if individuals are too closely related (*D. quadrilineatus*)
- Supports replication model from the first control region to WANCY region
- Next steps: how does the second control region affect the substitution rates in snake mitochondrial DNA?

Code

```
library(ape)
library(geiger)
tree <- read.tree(file.choose()) #choose dated_tree732

alltips <- tree$tip.label

goodtips <- c("Alluaudina_bellyi_RAN43680_seq1", "Alluaudina_mocquardi_
#tips we need - make a list

badtips <- setdiff(alltips, goodtips) #tips we don't need - make a lis

newtree <- drop.tip(tree, badtips)
newtree$tip.label
plot.phylo(newtree, cex=0.2)
write.nexus(newtree, file = "fubar_tree_final")
```

Generate and prune phylogenetic trees for FUBAR analysis.

```
library(BayesianFirstAid)

plot_rates <- function(csv_file, tolerance=0.90) {
  x <- read.csv(csv_file)
  x <- x[which(x[,8]<tolerance),] # filter neg. selection
  x <- x[which(x[,7]<tolerance),] # filter pos. selection
  distance <- x[,2]
  alpha <- x[,4]
  plot(distance, alpha,
       xlab="Distance from OL (codons)",
       ylab="Synonymous substitution rate", ) # scatter plot
  line<-lm(x[,4]~x[,2])
  fit<-bayes.cor.test(x$Distance,x$alpha)
  plot(fit) # bayesian correlation plot
}
```

Create scatter plots and Bayesian correlation plots.

```
library(circlize)

foutfinal<- read.csv(file = "foutfinal.csv")

# Create data
data = data.frame(
  factor = 1,
  x = foutfinal[,1],
  y = foutfinal[,2]
)

# General parameters
circos.par("track.height" = 0.4)

# Initialize chart
circos.initialize(factors = data$factor, x = data$x )

# Build the regions.
circos.trackPlotRegion(factors = data$factor, y=data$y, panel.fun = functic
  circos.axis(labels.cex=0.5, labels.font=1, lwd=0.8, h="bottom", directior
})
```

Create circular visualization of the rates across the genome.

Acknowledgements

Our team would like to thank Dr. Louise Crowley and Mali'o Kodis from BridgeUp, and Lauren Vonnahme, Arianna Kuhn, and Dr. Marcelo Gehara from the Herpetology Department.