Scientific Research Publishing

Open Journal of Statistics, 2018, 8, 793-810
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Detecting Variation in the Rate of Molecular Evolution in Different Lineages of Mammals

Sammer M. Marzouk

The University of Chicago, Chicago, USA
Email: smarzouk@ucls.uchicago.edu

How to cite this paper: Marzouk, S.M. (2018) Detecting Variation in the Rate of Molecular Evolution in Different Lineages of Mammals. *Open Journal of Statistics*, **8**, 793-810.
https://doi.org/10.4236/ojs.2018.85052

Received: April 20, 2018
Accepted: September 17, 2018
Published: September 20, 2018

Abstract

A major research topic within molecular evolution studies is to understand more about the causes of the molecular evolutionary patterns that are recorded within and between taxa. The amount of germ cell divisions in females and males causes the majority of mutations, during DNA replication, that impact molecular evolution. In an XX female and an XY male system of diploid animals, the autosomes come in duplicates, with one copy from the male and female parent. Because of this, the idea that evolution is driven by male mutations has become increasingly more likely. This paper looks at the different male-mutation rates and determines that the male-mutation rate is much higher than female-mutation rates. Ry/a to be approximately 2.2, which means that Ry is approximately −24.2. From software analysis, x was approximated to be about 0.5. And since x and Ry are known, Ry/x was determined to be −49. The results for this paper show the calculated Rx/a and Ry/a are similar to the results of another study, but they are unique in that they produced a relatively high negative number for the Ry/a, which was about −49. This provides evidence that the male-mutation rate is higher than the female-mutation rate. This is interesting because this suggests that, from the data, the mutation rate in males is the defining force in molecular evolution. And because the rate goes beyond the prescribed model, future models of molecular systems will need to consider the rate of male mutations, as well as clarifying this male-mutation rate and calculating the rate of mutation in other sex-determinant systems.

Keywords

Molecular Evolution, Molecular Biology, Genetics, Bioinfomatics

1. Introduction

A major focus of molecular evolutionary studies is to understand more about the

causes of the molecular evolutionary patterns that are recorded within and between taxa. Through studies of sex-linked loci, there has been more recorded information on how important the molecular forces (mutations, genetic drifts, selection, recombination) are relative to one another in changing the molecular variations of a species over time. At first, it was thought that the male-mutation rate should be higher than females due to males having a larger number of germ cell divisions (zygote to gamete) [1] (**Figure 1**). The study of sex-linked genetic diseases has been very influential in the study of the male-to-female mutation ratio (α) [2] [3]. The data from these studies provide support for the hypothesis that males have a higher rate of mutation (in humans). However, the studies are limited in the sample size of genes considered, the types of genes considered, and their relationship to sex-linked gene evolution within other taxa.

Through the assumption that the amount of germ cell divisions in females and males and those mistakes in DNA replication are the majority of mutations that impact molecular evolution, a model was presented that hypothesis replication-driven evolution [4]. Given a XX female and a XY male system of diploid animals, the autosomes come in duplicates, with one copy from the male and female parent. As the chance that a certain pair of autosomes carried by any given male and female is 50%, the mutation rate per generation of autosomes is proportional to $(\alpha + 1)/2$, given a long evolutionary timeline [5]. However, the expected mutation rate per generation for the Y chromosome is α. This is because the Y chromosome is only carried by the male. Females carry two X chromosomes. Males have one X chromosome. As such, the theoretical mutation rate per generation for the X chromosome is represented by $(\alpha + 2)/3$ [3]. Because of all of this, it follows that the ratio of the sex-linked rate to the autosomal mutation rate (let it be called ®) is $Rx/a = (2/3\ (2 + \alpha)/(1 + \alpha))$ for the X-linked genes. For the Y-linked genes, it is $Ry/a = 2\ \alpha/(1 + \alpha)$. In extreme cases of α, which is proportional to the rate of germ cell divisions in males to females, is much greater than unity, then Rx and Ry/a will approach 2/3 and 2, respectively [2].

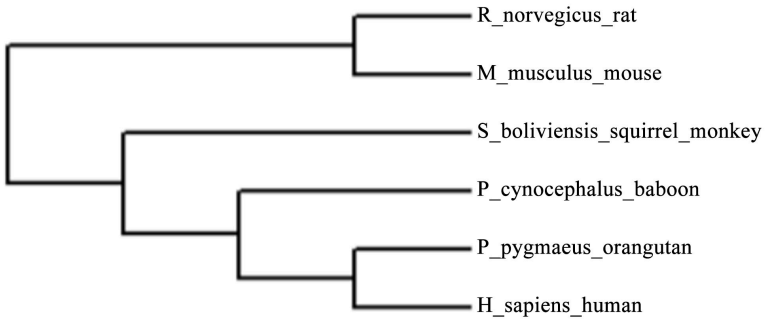

Figure 1. Phylogeny of Species. This phylogeny describes the relationship between the six mammalian species that were looked at over the course of this study. From the phylogeny, we see that the phylogeny splits up into two main clades. And that it subsequently divides into three sub-clades. Rats and mice form their own sub-clade. Humans and orangutans form their own sub-clade. And the Squirrel monkey serves as an in-between amongst these main sub-clades.

In order to test these models, comparisons of rates of synonymous divergence between identified sex-linked and autosomal genes in humans and mice was conducted. These studies calculated that Rx/a = 0.60 and Ry/a = 2.2 [6]. This suggests that α is large. Another study calculated a similar result from Rx/a, which was 0.61 [7]. In this study, the synonymous rates in X-linked and autosomal genes were compared between mice and rates. The results from these studies suggest that the molecular evolution of genes within mammals is a male-driven occurrence. However, manifold variation in synonymous substitution rates between different genes can bias and impact the achieved rate [1] [4]. This is caused by sampling a limited number of non-homologous genes.

There were two studies used in this paper with different amounts of X-linked genes. In the first study discussed, 4 X-linked genes were used. In the second study discussed, 11 X-linked genes were used. In addition, the rates of synonymous substitution of X-linked genes were hypothesized to be less than autosomes. This was mainly due to the fact that the increased impact of selection against deleterious X-linked alleles because of haploidy in males and the smaller population size (defined as one) of the X chromosome would decrease the rate (Charlesworth *et al.* 1987). The impact of the NE can be accounted for relative to drift. But it cannot be accounted for in the increase of effectiveness of weak selective constraints (defined in Nes) on synonymous sites [8].

In order to address the problem of contrasting synonymous rates between non-homologous genes, this paper will address the problem of male-driven evolutionary phenomen between ZFX and ZFY. ZFX and ZFY are a homologous pair of zinc-finger DNA binding motifs with proteins found on the X and on the Y chromosomes within mammals [9]. In this pair of genes, the ratio of the Y-linked over the X-linked mutation rates is expected to be modeled by the theoretical model Ry/a = $3\alpha/(2 + \alpha)$. This theoretical model approaches 3 when $\alpha >>> 1$ [8] [9]. Also, this paper uses the fact that there exists autosomal duplication of ZFX, names ZFA, in order to verify previous calculations of α based on X/Y-autosome comparison. In addition, to avoid the chance of weak selections on synonymous sites, this paper will investigate sequence divergences of an intron of ZFX and ZFY.

2. Materials and Methods

2.1. Mammal Gene Sequences

DNA sequences from six species of mammals. These species are:

1) Mouse
2) Squirrel Monkey
3) Baboon
4) Orangutan
5) Human
6) Norvegicus Rat

All of the DNA sequences for these species were used from NCBI Genomic

Database. The data was used in a constant FAST-All (FASTA), an extension of FAST-P and FAST-N, format for the paper. The sequences of the species will be included in the supplementary data (Supplementary Data 1). The relationship between these species is shown in Figure 2. From the phylogeny, we see that the phylogeny splits up into two main clades. And that it subsequently divides into three sub-clades. Rats and mice form their own sub-clade. Humans and orangutans form their own sub-clade. And the Squirrel monkey serves as an in-between amongst these main sub-clades.

2.2. MEGA

MEGA (**M**olecular **E**volutionary **G**enetics **A**nalysis) is a software that specializes in analyzing FASTA DNA sequences. The software emphasizes the integration of sequence acquisition with evolutionary analysis. It contains an array of input data and multiple results explorers for visual representation; the handling and editing of sequence data, sequence alignments, inferred phylogenetic trees; and estimated evolutionary distances [10]. The software allows the user the ability to browse, edit, summarize, export, and generate publication-quality captions for their results. MEGA also includes distance matrix and phylogeny explorers as well as advanced graphical modules for the visual representation of input data and output results. The main features of this software used in this paper are the phylogeny construction software and the substitution software [10] [11]. The substitution software will analyze the DNA sequences that are uploaded onto the program, after which, it will calculate the AiC value for each substitution model. The Akaike information criterion (AiC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AiC estimates the quality of each model, relative to each of the other models. Thus, AiC provides a means for model selection. AiC is founded on information theory: it offers an estimate of the relative information lost when a given model

	Accession numbers			
Species	ZFX	ZFY	ZFX intron	ZFY intron
H. sapiens	M30608	M30607	X58925	X58926
P. pygmaeus	X75169	X75176	X58932	X72698
P. cynocephalus	X75174	X75173	X58930	X58931
S. boliviensis	X75175	X75170	X58935	X58936
R. norvegicus	X75171	X75172	X58933	X58934
M. musculus	M32309	M24401	X58927	X58929
	ZFA: X53250			

Figure 2. Accession Number Table. This table describes the ascension numbers for the species that were used in the study. The accession numbers were in GenBank format. In order to access the genetic information of the species, these accession numbers were put into Batch Entrez, published by the NCBI.

is used to represent the process that generated the data. (In doing so, it deals with the trade-off between the goodness of fit of the model and the simplicity of the model.) The model with the lowest AiC value will be the model that will be used to analyze that sequence. See the supplementary information for more specifics on the AiC calculation. After the substitution model was determined, a phylogeny would be created with the gathered information. All of the phylogenies created using the Neighbor-Joining method [12].

2.3. Hyphy

HyPhy (**Hy**pothesis Testing using **Phy**logenies) is an open-source software package for the analysis of genetic sequences (in particular the inference of natural selection) using techniques in phylogenetics, molecular evolution, and machine learning [13]. The paper uses this software to compare the independent and dependent phylogenies through the use of a bootstrap analysis. The bootstrap was run within normal and default parameters. In the bootstrap, the minimum number of simulation recommended was 100. The program allows 100 - 1000 simulations. For this analysis, 550 simulations were run. The simulations calculated the LR value. The LR value is the likelihood ratio, defined as 2 (log L– log L0), where LA is the likelihood for the alternative hypothesis, L0 is the likelihood for the null hypothesis (refer to the documentation for HyPhy) [9] [12]. A simulation in the bootstrap is to pick random sites from the original sequence with replacement, rebuild the phylogenetic tree for two hypotheses, calculate the log likelihood and generate one likelihood ratio. The goal is to see the likelihood ratio from the data fall into the empirical distribution. In the program, the null hypothesis was entered and the alternative hypothesis [7]. A null hypothesis supports the hypothesis that there is no significant difference between specified population; that any observed difference being due to sampling or experimental error. The alternative hypothesis supports the hypothesis that there is a significant difference between specified population and that these differences share a cause. If the p-value is really small, $\sim p < 0.0005$, the null hypothesis is rejected. The possibility of proximal and distal genes evolving independently as the alternative hypothesis. And the possibility of proximal and distal genes evolving independently as the null hypothesis [9]. The MEGA software was also used to align and organize the DNA before the phylogenies were created.

2.4. Phylogeny.fr

The Phylogeny.fr platform transparently chains programs perform complex genomic and proteomic phylogenetic analyses. It is run by the Reseau National de Genopoles. Phylogeny.fr offers three main modes [13]. The “One Click” mode targets non-specialists and provides a ready-to-use pipeline chaining programs with recognized accuracy and speed: MUSCLE for multiple alignments, PhyML for tree building, and TreeDyn for tree rendering. All parameters are set up to suit most studies, and users only have to provide their input sequences to obtain

a ready-to-print tree [4] [7] [12]. The "Advanced" mode uses the same pipeline but allows the parameters of each program to be customized by users. The "A la Carte" mode offers more flexibility and sophistication, as users can build their own pipeline by selecting and setting up the required steps from a large choice of tools to suit their specific needs. Prior to phylogenetic analysis, users can also collect neighbors of a query sequence by running BLAST on general or specialized databases. A guide tree then helps to select neighbor sequences to be used as input for the phylogeny pipeline. This paper uses the advanced mode in order to create a second version of the dependent phylogeny [4] [7] [13].

2.5. Mathematica

Wolfram Mathematica is a modern technical computing system spanning all areas of technical computing—including neural networks, machine learning, image processing, geometry, data science, visualizations, and others. The system is used in many technical, scientific, engineering, mathematical, and computing fields. In this experiment, this software was used in order to look at the similarity index between branches on a phylogeny [14]. It would then go through and calculate the average similarity index between the branches, the clades, and the entire phylogeny. This would eventually lead to the calculation of α [6] [13] [15].

3. Results

3.1. Experimental Lay-Out

For this experiment, it began by retrieving all coding regions and aligning the correct number of sites. The regions were obtained (either complete cds or partial exons) from Genbank for ZFX and ZFY genes for the following species (**Figure 1**). The phylogeny of the species is as shown below. This phylogeny was reworked into Newick form and saved (**Supplemental 1**). After which, a file was created that combined the ZFX and ZFY gene sequences. After which, these files were converted into FASTA format files with all the sequences (Supplementary Data 1). After which, the MEGA7 software was used in order to align the sequences using the ClustalW operation [15].

After this alignment, the portion of the alignment that contained all the sequences for all the species was selected. This region consisted of about 1346 sites (Supplementary Data 2). Since this experiment involved synonymous/nonsynonymous rates, the area of ~1300 sites was re-aligned so that all of the sequences were in the correct frame.

The MEGA7 software was then used to create the phylogenies based off of these genomic regions. These phylogenies would be the ones used in the HyPhy analysis (**Figure 3** and **Figure 4**). These phylogenies were created using the default options on the MEGA7 software. The phylogenies created from the MEGA7 software were then compared to another set of phylogenies. These phylogenies were made using the phylogeny.fr software. The purpose of making more than one set was to compare the different phylogenies. This would result in

developing on consensus phylogeny (Figure 5 and Figure 6). For the rest of the analysis, the phylogenies found in figures five and six will be the ones used in the HyPhy analysis.

Using this same FASTA file, this was uploaded onto HypHy and partitioned [16]. Since the aim is to calculate the rate of evolution along all the terminal branches, the partition type selected was "codon". For the tree, this paper considered Figure 6 and Figure 7 as the phylogenies that were included in the HyPhy calculations. For the substitution model, the chosen model was GY94_3x4. This was chosen from a best-fit model calculated from the MEGA7 software. The equilibrium parameter is set to "partition." After this, a function was created from these parameters (Supplemental 2). From these parameters, HyPhy calculated the evolutionary rates for all the branches (Supplemental 3). This gives us the rates for rates of nucleotide substitution in the X-linked and Y-linked zinc finger genes.

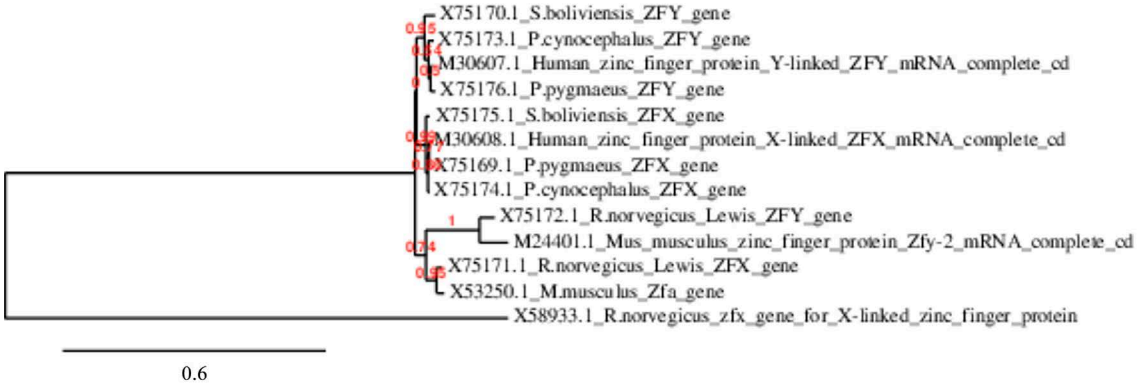

Figure 3. Evolutionary Relationship of taxa of ZFX genes. The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 0.8 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The analysis involved 16 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1585 positions in the final dataset. Evolutionary analyses were conducted in Phylogeny.fr.

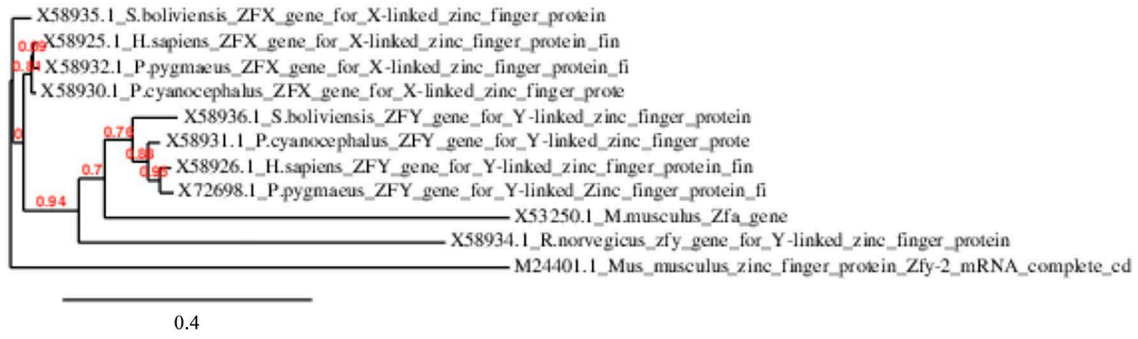

Figure 4. Evolutionary Relationship of taxa of ZFY genes. The evolutionary past was calculated using the Neighbor-Joining method. The tree with the added sum of branch length of 0.89 is visualized above. The computation had 16 nucleotide sequences. All the positions with gaps and/or missing data were removed from the data set. There were 1585 positions in the final computation.

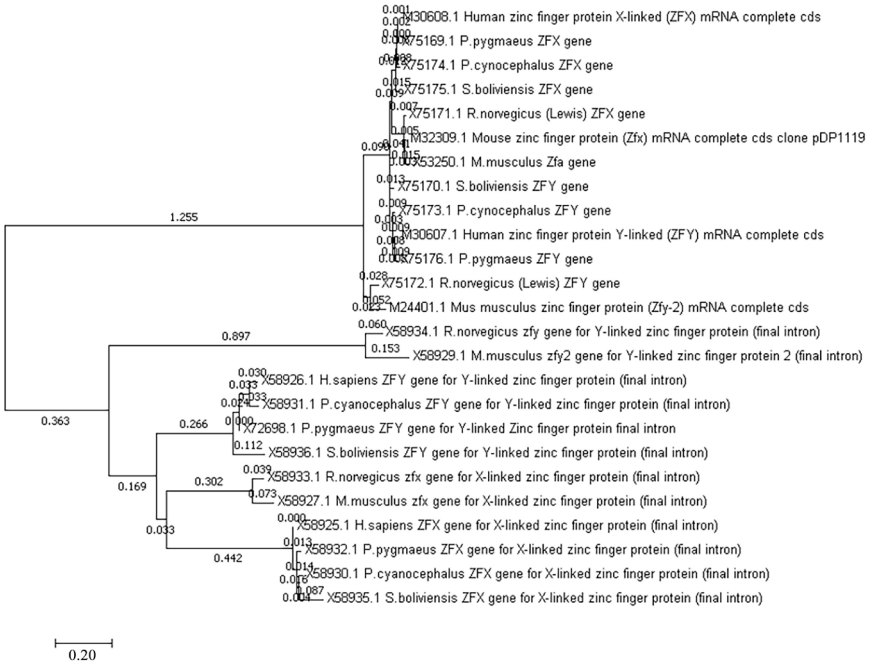

Figure 5. Evolutionary relationships of taxa using all genetic samples. The evolutionary history was averaged using the Neighbor-Joining method. The phylogeny with the average of branch length was 4.79392861, which is the average branch length in the tree. The computational analysis required 25 nucleotide sequences. All the possibilities that contained gaps and missing data were eliminated. There 610 positions within the terminal dataset.

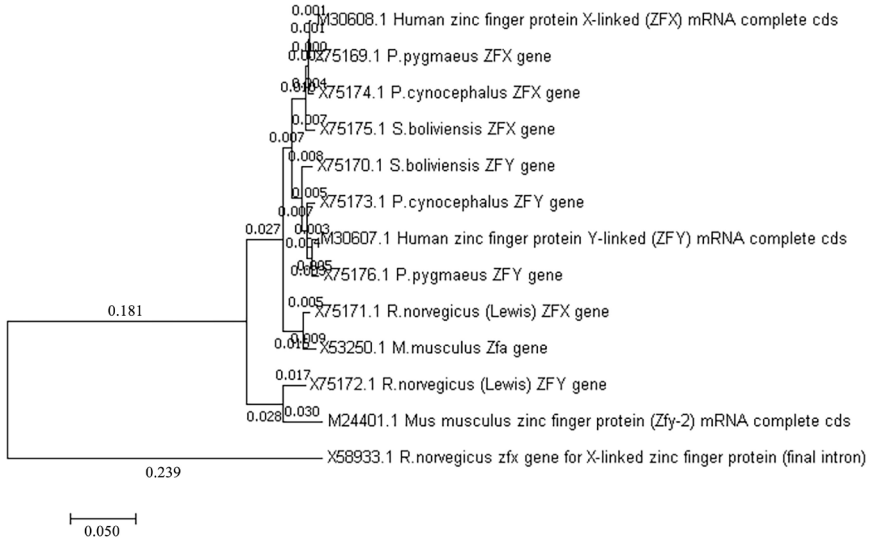

Figure 6. Evolutionary relationships of taxa using intron genetic samples. The evolutionary history was estimated using the Neighbor-Joining method. The tree of choice was determined with the average sum of branch length method. This was calculated to be 0.61897064, and it is demonstrated from the figure above. The analysis took advantage of 13 nucleotide sequences. All the possible positions containing gaps of missing data and/or information were eliminated. There were 846 recorded positions in the final dataset from the intron genetic samples. Evolutionary computation analyses and calculations were done using the MEGA7 software.

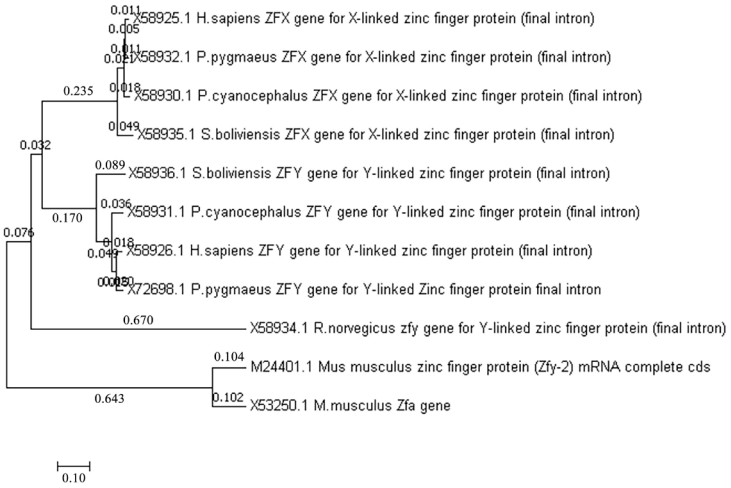

Figure 7. Evolutionary relationships of taxa using non-intron genetic samples. The cumulative evolutionary history was detected and computed using the Neighbor-Joining method. The tree of best fit was one with an average sum of branch length = 2.37315259. The analysis was conducted from a set of 11 nucleotide sequences. Codon positions included developed from 1st + 2nd + 3rd + Noncoding strands. All positions with gaps and missing data were removed from the dataset before the calculated rates of average change. There were a total of 676 positions in the terminally calculated dataset. Evolutionary analyses were conducted in MEGA7.

The next step was evaluating the male-driven evolution hypothesis using ZFX and ZFY intron sequences. For this process, a similar process was conducted as with the gene samples above. For this, the phylogeny was used in the HyPhy software was **Figure 6**. The final evolutionary rates are shown in **Supplemental 4** and **Supplemental 5**.

3.2. Calculation of Evolutionary Coefficients

After the conclusion of the HyPhy analysis, the information of the genetic similarities was then saved and uploaded to Mathematics. Using the data analysis software available, the synonymous and nonsynonymous rates for the intron zinc finger genetic samples were calculated (**Figure 8** and **Figure 9**) [17].

Mathematica was then used to calculate the Ry/x ratio. It would then go through and calculate the average similarity index between the branches, the clades, and the entire phylogeny. This would eventually lead to the calculation of α [18] [19]. Since the Rx/a equation is known, it can be re-written so that Rx = (((2/3)(2 + a)(a))/(1 + a)) [20]. From the Mathematica function, α was approximately −10.874575054845. However, for the rest of the paper, this will be referred to as −11. But in the calculations, the exact value was used. From this, Rx was calculated 0.592947503738, but it will be rounded to 0.6. We see that $\alpha \sim -11$ and Rx ~ 0.6 [21] [22].

Since we know α, we can calculate Ry/a to be approximately 2.2, which means that Ry is approximately −24.2 [23]. From the Mathematica software, x ~ 0.5 [24]. Since we know what x and Ry to be, Ry/x to be −49 [20] [22] [24].

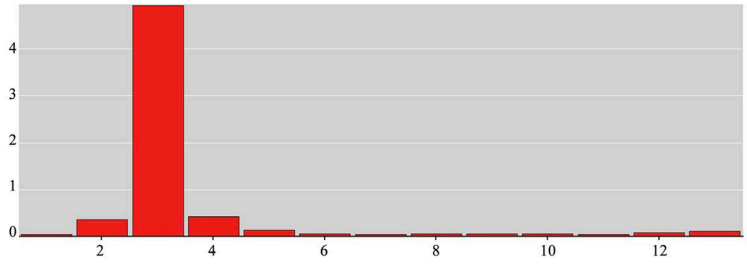

Figure 8. Quantification of Branch Lengths of Indox/Indux Genetic Profiles. This figure shows the quantification and visualization of the data that was gained from Supplemental 4. The x-axis is the mean of the sample group, and the y-axis shows the actual numerical count of the branch lengths. This shows that a branch length of mean 3 with a relationship coefficient of 5 was the most prevalent in the phylogeny.

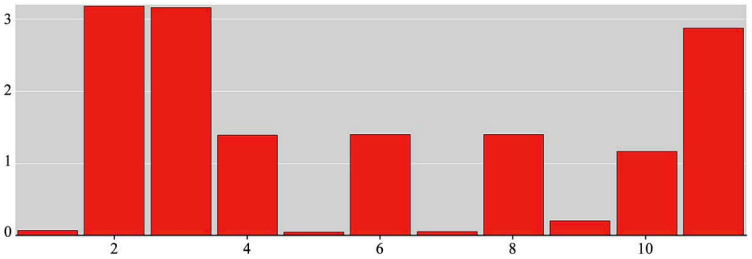

Figure 9. Quantification of Branch Lengths of Zinc Finger Genetic Profiles. This figure shows the quantification and visualization of the data that was gained from Supplemental 5. The x-axis is the mean of the sample group, and the y-axis shows the actual numerical count of the branch lengths. This shows that a branch length of mean 3 with a relationship coefficient of 3.5 and a branch mean of 3 with a branch coefficient of 3.5 were the most prevalent in the phylogeny.

3.3. Analysis of the Phylogenies

In the Intron phylogeny, we see that there are five sub-clades that are formed (Figure 4). In the uppermost sub-clade, we have M30608, X75196.1, Z75174.1, and X75175.1 (Figure 4). In the second sub-clade, which is a part of the same general clade as sub-clade one, has X75170.1, X75173.1, M30607.1, and X75176.1 (Figure 4). In the third sub-clade, X75171.1 and X53250.1 were related together (Figure 4). The fourth sub-clade has X75172.1 and M24401.1 (Figure 4). And the final sub-clade is a single sample called X58933.1 (Figure 4). This is an interesting phylogeny because it placed the M308, M244, and M306 within clades that were mainly inhabited by X and Y genetic profiles. This is even more interesting as X58933.1 was placed as outlier in this phylogeny (Figure 4). And throughout all demonstrates phylogenies, X58933 was placed as an outlier or as a genetic sample that is ancestral different from most samples.

Zinc Finger phylogeny divides up into four main sub-clades. The first clade is further divided into two clades (Figure 6). The top clade has X58925.1, X58930.1, X58935, and X58932. The second clade has X58938.1, X58931.1, X58926.1, and X72698.1 (Figure 6). The third clade has only one protein, X58934.1 (Figure 6). And the final clade is made up of M24401.1 and X53250.1 We see that different Zinc finger genes are related, even though they might not

be the same type of classification. For example, we see that M24401.1 be more closely related to X53520.1. It would be more reasonable for it to be closely related to X58934.

In the cumulative phylogeny, we see that there are six main sub-clade that are formed (Figure 5). In the first one, we have X58935.1-X58923.1 (Figure 5). In the second sub-clade, we have X58927.1 and X58933.1 (Figure 5). In the third sub-clade, we have X58926.1-X58936.1 (Figure 5). In the fourth clade, X58934.1 and X58929.1 are considered ancestors (Figure 5). In the fifth sub-clade, X58972.1 and M24401.1 are considered related (Figure 5). In the sixth sub-clade, X58976.1-X58908.1 are considered related (Figure 5). Again, in this cumulative phylogeny, we also see some unforeseeable overlap between genetic species that we would not have considered to be related. For example X58931.1 and M24401.1 form their own separate sub-clade. This is interesting because it would have been through that X58931.1 would be apart of the larger sub-clade above with, with members of the X5 family. A similar thing could have been expected from the M24401.1 sample.

3.4. Analysis of Tree Length Averages

From the tree length averages, we went on to quantify them and performed a statistical analysis on the samples (Supplemental 4, Supplemental 5). From this, we got Figure 8 and Figure 9. In Figure 8, we see that a tree with a mean length of 3 and with value of 5 is the most common type of relationship among the phylogenies (Figure 8). There also appears to be a left-skewing of the data. This might come from the fact that there are more samples of this data set that have length averages between 1 - 3. This was the phylogeny from the Zinc Finger gene (Figure 8). For the Intron gene family, we see that there is more of a standard distribution to the results. We see that the most common relationship among the clades is when the averages of the branches is between 2 - 3, inclusive, and the branch length constant 3.3333. There are other groups that have larger and influential datasets within this graph (Figure 9). Looking at trees that are an average of 12 long, with a branch length of 2.6777, this appears to be a major deviation from the results of the Zinc finger, which didn't have many branches above the length of 4 (Figure 9). And in the intron graph, there appears to be a pattern in the middle (Figure 9). We see that the branch goes from 1.33, to 0, to 1333, to 0. In general this graph appear to have standard distribution (Figure 9).

4. Discussion

The results obtained by this paper are similar to results from other studies conducted. We calculated Rx to be 0.592947503738, but it will be rounded to 0.6. We see that $a \sim -11$ and Rx ~ 0.6 [25]. Since we know a, we can calculate Ry/a to be approximately 2.2, which means that Ry is approximately −24.2. From the Mathematica software, x ~ 0.5. Since we know what x and Ry to be, Ry/x to be −49 [24]. The results this paper produced for Rx/a and Ry/a are similar to the

results of another student. Our results are unique in that we did get a relatively high negative number for the Ry/a, which was about −49 (**Supplemental 3**). This provides evidence that the male-mutation rate is higher than the female-mutation rate [13] [15] [24]. This is interesting because this suggests that, from the data, the mutation rate in males is so high that it actually goes off of the limit prescribed by the mathematical model. This means that we would need more data to accurately describe the male-mutation rate.

The potential problems and limitation that come from using the ZFX/ZFY system for evaluating male-driven evolution hypothesis is that the ZFX/SFY system needs to be considered in conjunction with other genetic systems for the most accurate data. The reason why looking at the ZFX/ZFY system is limiting is because of the fact that this is not the only evolving and mutation system in the human body. By limiting our point of comparison to just one system, we cannot make a conclusion about which gender has the higher mutation rate. It would be better to look at a large range of genetic systems that are involved in the reproduction process for the best quality data on this process.

Another issue that arises is whether or not the mutations that come from the system are influential or not in the reproduction process. The hypothesis is that higher male-mutation rates contribute more to the evolution of female-mutation rate. This causes male-mutation driven evolution. However, in order for this to be the case, there needs to be a way of measuring the impact of mutations. Not every mutation is the same. As shown here, this paper shows that the male-mutation rate is higher than the female-mutation rate. However, there needs to be a way of showing that male mutations are influential or preferred in the reproductive process.

The major advantages of using synonymous rates in these two rates are influential in the male mutation process. And by looking at these rates, it is possible to get an accurate depiction of α. In addition, synonymous rates are universal among species, meaning that it is easier to compare synonymous rates from different species than it is to do the same with intron rates [14] [18] [24].

The major disadvantage of using synonymous rates is that selection restricts nonconservative synonymous rates. This is important to consider because this causes a statistical bias. In order to calculate the most accurate α, it is necessary to look at both nonconservative and conserved synonymous rates. By only looking at one, the data becomes skewed.

The advantage of looking at intron rates is that intron may be related to mRNA movement and chromatin assembly. And the length of introns matters when it comes to evolution. This means that looking at the intron might give more insight into the evolutionary impact that male mutations do [12] [20] [25].

The major disadvantage of intron rates is that they are limited to specific species and are variable. All of the completely sequenced eukaryotic harbor introns in the genomic structure, whereas no prokaryotes identified so far carry introns. Second, the amount of total introns varies in different species [23]. Third, the

length and number of introns vary in different genes, even within the same species genome. Fourth, all introns are copied into RNA by transcription and DNAs by replication processes, but intron sequences do not participate in protein-coding sequences [19] [23].

I believe that using intron rates is more beneficial when comparing samples among the same species or similarly related sister species. However, for the most part, I believe it more useful and beneficial to use synonymous rates when comparing many different species that might or might not be related. For this experiment, the synonymous rates might better reflect the true male-mutation rate [19].

The difference in male-mutation rates between rodents and primates is a generation-time effect that comes from a meaningful biological different that has increased overtime. First, the difference between rodent and primate α is best explained by the generation-time effect. As males produce gametes continuously throughout adulthood, the number of germline cell divisions increases with paternal age; the number of female germline cell divisions is, by contrast, insensitive to age. Species with longer generation times, therefore, experience greater discrepancies in the number of germline cell divisions between the sexes. Given the difference in generation times, we expect α to be higher in humans than in rats. Moreover, if α has evolved with generation time, then the human-rat α represents a long-term average over the deep branches connecting the two species. Most of the divergence history between humans and rat is undoubtedly characterized by short generation times and, hence, small α, as the long generation times of hominoids evolved recently. For the same reason, estimating α from human–macaque divergence might not accurately reflect α for recent human or chimpanzee molecular evolutionary history. Indeed, genome-scale data show that male-biased mutation is considerably lower in Old World monkeys than in hominoids. Second, two notably low estimates of α from X and Y chromosome sequence data in humans A study noted that the repeat-based analysis failed to correct for multiple substitutions and made the false assumption that repetitive elements in the same subfamily are the same age [14] [16] [21].

For future research directions, male-driven evolution can be tested by looking at other sex determination systems in larger evolutionary contexts. For example, using the data from the ZFX/ZFY study conducted here, it would be beneficial to look at the SMCY/SMCY system. This is another sex determination system that has more information on the male-mutation rate. In addition, looking at the mitochondrial and chromosomal DNA sequences within an evolutionary context would also produce information that could further test the male-driven evolution hypothesis [14].

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Haldane, J.B.S. (1947) A Test for Homogeneity of Records of Familial Abnormalities. *Annals of Eugenics*, **14**, 339-341. https://doi.org/10.1111/j.1469-1809.1947.tb02412.x

[2] Wolfe, K.H., and Sharp, P.M. (1993) Mammalian Gene Evolution: Nucleotide Sequence Divergence between Mouse and Rat. *JME*, **37**, 441-456.

[3] Ketterling, R.P., Ricke, D.O., Wurster, M.W. and Sommer, S.S. (1993) Deletions with Inversions: Report of a Mutation and Review of the Literature. *Human Mutation*, **2**, 53-57. https://doi.org/10.1002/humu.1380020110

[4] Anisimova, M. and Gascuel, O. (2006) Approximate Likelihood Ratio Test for Branches: A Fast, Accurate and Powerful Alternative. *Systematic Biology*, **55**, 539-552. https://doi.org/10.1080/10635150600755453

[5] Baer, A. (1988) Human Genetics. Problems and Approaches. Springer-Verlag, New York.

[6] Castresana, J. (2000) Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, **17**, 540-552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

[7] Chang, B.H.-J., Shimmin, L.C., Suyue, S.-K., Hewett-Emmett, D. and Li, W.-H. (1994) Weak Male-Driven Evolution in Rodents. *PNAS*, **91**, 827-831. https://doi.org/10.1073/pnas.91.2.827

[8] Charlesworth, B., Coyne, J.A., and Barton, N.H. (1987) The Relative Rates, of Evolution of Sex Chromosomes and Autosomes. *The American Naturalist*, **130**, 113-146. https://doi.org/10.1086/284701

[9] Chevenet, F., Brun, C., Banuls, A.L., Jacq, B. and Chisten, R. (2006) TreeDyn: Towards Dynamic Graphics and Annotations for Analyses of Trees. *BMC Bioinformatics*, **7**, 439. https://doi.org/10.1186/1471-2105-7-439

[10] Dereeper, A., Audic, S., Claverie, J.M. and Blanc, G. (2010) BLAST-EXPLORER Helps you Building Datasets for Phylogenetic Analysis. *BMC Evolutionary Biology*, **10**, 8. https://doi.org/10.1186/1471-2148-10-8

[11] Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M. and Gascuel, O. (2008) Phylogeny.fr: Robust Phylogenetic Analysis for the Non-Specialist. *Nucleic Acids Research*, **36**, W465-469. https://doi.org/10.1093/nar/gkn180

[12] Dorit, R.L., Akashi, H. and Gilbert, W. (1995) Absence of Polymorphism at the ZFY Locus in the Human Y Chromosome. *Science*, **268**, 1183-1185. https://doi.org/10.1126/science.7761836

[13] Du, J. (1998) Singular Integral Operators And Singular Quadrature Operators Associated with Singular Integral Equations. *Acta Mathematica Scientia*, **18**, 227-240. https://doi.org/10.1016/S0252-9602(17)30757-9

[14] Edgar, R.C. (2004) MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research*, **32**, 1792-1797. https://doi.org/10.1093/nar/gkh340

[15] Greuel, G.-M., Pfister, G. and Schönemann, H. (2006) Singular 3.0. A Computer Algebra System for Polynomial Computations. Centre for Computer Algebra, University of Kaiserslautern, Kaiserslautern. http://www.singular.uni–kl.de

[16] Greuel, G.-M. and Pfister, G. (2002) A Singular Introduction to Commutative Algebra. Springer, Berlin. https://doi.org/10.1007/978-3-662-04963-1

[17] Guindon, S. and Gascuel, O. (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, **52**, 696-704. https://doi.org/10.1080/10635150390235520

[18] Hatyashida, H., Kuma, K. and Miyata, T. (1992) Intrachromosomal Gene Conversion as a Possible Mechanism for Explaining Divergence Patterns of ZFY-Related Genes. *Journal of Molecular Evolution*, **35**, 181-183.

[19] Lanfear, J. and Holland, P.W.H. (1991) The Molecular Evolution of ZFY-Related Genes in Birds and Mammals. *Journal of Molecular Evolution*, **32**, 310-315. https://doi.org/10.1007/BF02102189

[20] Malaspina, P., *et al*. (1990) The Human Y Chromosome Shows a Low Level of DNA Polymorphism. *Annals of Human Genetics*, **54**, 297-305. https://doi.org/10.1111/j.1469-1809.1990.tb00385.x

[21] Miyata, T., Hatyashita, H., Kurna, K., Mitsuyasu, K. and Yasunaga, T. (1987) Male-Driven Molecular Evolution: A Model and Nucleotide Sequence Analysis. *Cold Spring Harbor Symposia on Quantitative Biology*, **52**, 863-867.

[22] Shimmin, L.C., Chang, B.H.-J. and Li, W.-H. (1993) Male-Driven Evolution of Sequences. *Nature*, **362**, 745-747. https://doi.org/10.1038/362745a0

[23] Shimmin, L.C., Chang, B.H.-J., Hewett-Enunett, D. and Li, W.-H. (1993) Potential Problems in Estimating the Male-to-Female Mutation Ratio from DNA Sequence Data. *Journal of Molecular Evolution*, **37**, 160-166. https://doi.org/10.1007/BF02407351

[24] Shimmin, L.C., Chang, B.H.-J. and Li, W.-H. (1994) Contrasting Rates of Nucleotide Substitution in the X- and Y-Linked Zinc Finger Genes. *Journal of Molecular Evolution*, **39**, 569-578. https://doi.org/10.1007/BF00160402

[25] Stokes, D.G. and Perry, R.P. (1995) DNA-Binding and Chromatin Localization Properties of CHD1. *Molecular and Cellular Biology*, **15**, 2745-2753. https://doi.org/10.1128/MCB.15.5.2745

Supplementary Data

Supplemental 1. Newick phylogeny of zinc finger and intron sequences.

```
((((((('M30608.1 Human zinc finger protein X-linked (ZFX) mRNA complete
cds':0.00131645,X75169.1_P.pygmaeus_ZFX_gene:0.00032429):0.00213707,X75174.1_P.cynocephalus_ZFX_gene:0.00780198):0.00591123,X75175
.1_S.boliviensis_ZFX_gene:0.01532804):0.01192762,('X75171.1 R.norvegicus (Lewis) ZFX gene':0.00670358,('M32309.1 Mouse zinc finger protein
(Zfx) mRNA complete cds clone
pDP1119':0.00513205,X53250.1_M.musculus_Zfa_gene:0.01490778):0.00337724):0.04050972):0.00881072,(X75170.1_S.boliviensis_ZFY_gene:0.01
286322,(X75173.1_P.cynocephalus_ZFY_gene:0.00886358,('M30607.1 Human zinc finger protein Y-linked (ZFY) mRNA complete
cds':0.00909234,X75176.1_P.pygmaeus_ZFY_gene:0.00926128):0.00803199):0.00845412):0.00289818):0.09046867,('X75172.1 R.norvegicus (Lewis)
ZFY gene':0.02751163,'M24401.1 Mus musculus zinc finger protein (Zfy-2) mRNA complete cds':0.05212709):0.02342578,(('X58934.1
R.norvegicus zfy gene for Y-linked zinc finger protein (final intron)':0.06014926,'X58929.1 M.musculus zfy2 gene for Y-linked zinc finger protein
2 (final intron)':0.15295979):0.89689279,((((('X58926.1 H.sapiens ZFY gene for Y-linked zinc finger protein (final intron)':0.02984941,'X58931.1
P.cyanocephalus ZFY gene for Y-linked zinc finger protein (final in-
tron)':0.03302199):0.03260246,X72698.1_P.pygmaeus_ZFY_gene_for_Y-linked_Zinc_finger_protein_final_intron:0.00000000):0.02386518,'X5893
6.1 S.boliviensis ZFY gene for Y-linked zinc finger protein (final intron)':0.11208143):0.26578143,(('X58933.1 R.norvegicus zfx gene for X-linked
zinc finger protein (final intron)':0.03905238,'X58927.1 M.musculus zfx gene for X-linked zinc finger protein (final in-
tron)':0.07344376):0.30167258,('X58925.1 H.sapiens ZFX gene for X-linked zinc finger protein (final intron)':0.00000000,('X58932.1 P.pygmaeus
ZFX gene for X-linked zinc finger protein (final intron)':0.01298981,('X58930.1 P.cyanocephalus ZFX gene for X-linked zinc finger protein (final
intron)':0.01387674,'X58935.1 S.boliviensis ZFX gene for X-linked zinc finger protein (final in-
tron)':0.08690421):0.00400487):0.01564675):0.44233224):0.03293563):0.16891097):1.61776924);
```

This table describes the Newick form of the Zinc finger and Intron sequences. This format was given by the MEGA7 software, which converted the sequence into Newick format. This was used in the lab to build the phylogeny for the HyPhy software. The ZFX gene on the X chromosome is structurally similar to a related gene on the Y chromosome. It encodes a member of the krueppel C2H2-type zinc-finger protein family. The full-length protein contains an acidic transcriptional activation domain (AD), a nuclear localization sequence (NLS) and a DNA binding domain (DBD) consisting of 13 C2H2-type zinc fingers. Studies in mouse embryonic and adult hematopoietic stem cells showed that this gene was required as a transcriptional regulator for self-renewal of both stem cell types, but it was dispensable for growth and differentiation of their progeny. Multiple alternatively spliced transcript variants encoding different isoforms have been identified, but the full-length nature of some variants has not been determined. The ZFY gene is on the Y chromosome that is similar to the ZFX gene in every other regard.

Supplemental 2. Branch lengths of indox and zinc finger genes in figure six.

```
columnHeaders = {{"Length","Branch;1;10;11;12;13;2;3;4;5;6;7;8;9"}};
tableData = {
{0.05010068441289961}
{0.3714184652550603}
{4.931518988834103}
{0.4299480197153737}
{0.1522065179520734}
{0.06345844241541707}
{0.05006735938485853}
{0.06327432886419679}
{0.0617543634374266}
{0.0633619613606832}
{0.05118810449541206}
{0.08467604239294882}
{0.1194693255807737}
};
OpenWindow (CHARTWINDOW,{{"Branch Length Distribution for New_Tree"}
        {"columnHeaders"}
        {"tableData"}
        {"Bar Chart"}
        {"Index"}
        {"Length"}
        {"(null)"}
        {"(null)"}
        {"(null)"}
        {"0"}
        {""}
```

{"0;0"}
{"10;1.309;0.785398"}
{"Times:14:0;Times:12:0;Times:14:2"}

{"0;0;13816530;16777215;0;0;6579300;11842740;13158600;14474460;0;3947580;16777215;15670812;6845928;16771158;2984993;9199669;7018159;1460610;16748822;11184810;14173291"}
{"16,0,0"}
},
"416;459;70;70");

This table describes the lengths of the branches of the two types of genes that are looked at during this project. In addition, they are shown above in HTML format so that they can be reproduced on the HypHy software. This represents the values from the phylogeny represented in figure six.

Supplemental 3. Parametric values from optimization of figure six.

Parameter ID,Value,Constraint
New_Tree,,
sequence_4_part_Shared_TVTS,0.3014325675649091,
New_Tree.1.nonSynRate,0.06901588413547036,
New_Tree.1.synRate,0.05053588807565249,
New_Tree.10.non SynRate,3.188961312233401,
New_Tree.10.synRate,3.052103709022195,
New_Tree.11.non SynRate,3.16449024503411,
New_Tree.11.synRate,3.869432618145147,
New_Tree.2.non SynRate,1.393360696267385,
New_Tree.2.synRate,1.098712163770307,
New_Tree.3.non SynRate,0.05083488694089981,
New_Tree.3.synRate,0.06135334632311153,
New_Tree.4.non SynRate,1.406608296268738,
New_Tree.4.synRate,0.9622893298051612,
New_Tree.5.non SynRate,0.05782245062023443,
New_Tree.5.synRate,0,
New_Tree.6.non SynRate,1.403747977122309,
New_Tree.6.synRate,0.9678068097761779,
New_Tree.7.non SynRate,0.2118662914766112,
New_Tree.7.synRate,0.231417634535939,
New_Tree.8.non SynRate,1.170726194290785,
New_Tree.8.synRate,1.228942532197198,
New_Tree.9.nonSynRate,2.88131994242897,
New_Tree.9.synRate,2.881509938437538,

This table shows the values of the optimization from the HyPhy software. These values represent the phylogeny shown in figure six. The optimization covered both Synonymous rates (synRate) and Non-synonymous rates (non SynRate).

Supplemental 4. Parametric values from optimization of figure seven.

{
{New_Tree,,}
{sequence_2_part_Shared_TVTS,0.1398566456958525,}
{New_Tree.1.nonSynRate,0.046079846011786,}
{New_Tree.1.synRate,0.05010181347408959,}
{New_Tree.10.non SynRate,0.532757503965848,}
{New_Tree.10.synRate,0.3714563320244949,}
{New_Tree.11.non SynRate,5.523357746603065,}
{New_Tree.11.synRate,4.931020919638585,}
{New_Tree.12.non SynRate,0.6293587946200982,}
{New_Tree.12.synRate,0.4298926777986825,}
{New_Tree.13.non SynRate,0.257933820813617,}
{New_Tree.13.synRate,0.1521883666308918,}
{New_Tree.2.non SynRate,0.1526668208524836,}

```
{New_Tree.2.synRate,0.06345357370253404,}
{New_Tree.3.non SynRate,0.04882866273434684,}
{New_Tree.3.synRate,0.05006407701062028,}
{New_Tree.4.non SynRate,0.1384273534142994,}
{New_Tree.4.synRate,0.0632724542529494,}
{New_Tree.5.non SynRate,0.06910128421774275,}
{New_Tree.5.synRate,0.06174732411099753,}
{New_Tree.6.non SynRate,0.1357582495034291,}
{New_Tree.6.synRate,0.06334848811915411,}
{New_Tree.7.non SynRate,0.09042285724557725,}
{New_Tree.7.synRate,0.05117526113499771,}
{New_Tree.8.non SynRate,0.1364094758369981,}
{New_Tree.8.synRate,0.08468395876463589,}
{New_Tree.9.nonSynRate,0.2126365008187848,}
{New_Tree.9.synRate,0.1194518767592083,}
}
```

This table shows the values of the optimization from the HyPhy software. These values represent the phylogeny shown in figure seven. The optimization covered both Synonymous rates (synRate) and Non-synonymous rates (non SynRate).

Supplemental 5. Branch lengths of indox and zinc finger genes in figure seven.

```
columnHeaders = {{"Length","Branch;1;10;11;2;3;4;5;6;7;8;9"}};
tableData = {
{0.05053588807565249}
{3.052103709022195}
{3.869432618145147}
{1.098712163770307}
{0.06135334632311153}
{0.9622893298051612}
{1e-009}
{0.9678068097761779}
{0.231417634535939}
{1.228942532197198}
{2.881509938437538}
};
OpenWindow (CHARTWINDOW,{{"Branch Length Distribution for New_Tree"}
                {"columnHeaders"}
                {"tableData"}
                {"Bar Chart"}
                {"Index"}
                {"Length"}
                {"(null)"}
                {"(null)"}
                {"(null)"}
                {"0"}
                {""}
                {"0;0"}
                {"10;1.309;0.785398"}
                {"Times:14:0;Times:12:0;Times:14:2"}

        {"0;0;13816530;16777215;0;0;6579300;11842740;13158600;14474460;0;3947580;16777215;15670812;6845928;16771158;2984993;9199669;7018159;1460610;16748822;11184810;14173291"}
                {"16,0,0"}
                },
                "416;459;70;70");
```

This table describes the lengths of the branches of the two types of genes that are looked at during this project. In addition, they are shown above in HTML format so that they can be reproduced on the HypHy software. This represents the values from the phylogeny represented in figure seven.