Fall 2014

Molecular Evolution of Hominoid Primates: Phylogeny and Regulation

Ranajit Das

MOLECULAR EVOLUTION OF HOMINOID PRIMATES: PHYLOGENY AND REGULATION

A Dissertation

Submitted to the Bayer School of Natural and Environmental Sciences

Duquesne University

In partial fulfillment of the requirements for

the degree of Doctor of Philosophy

By

Ranajit Das

December 2014

MOLECULAR EVOLUTION OF HOMINOID PRIMATES: PHYLOGENY AND REGULATION

By

RANAJIT DAS

23rd June 2014

Dr. Michael Jensen-Seaman Associate Professor of Biological Sciences (Committee Chair)	Dr. Brady Porter Associate Professor of Biological Sciences (Committee Member)
Dr. David Lampe Associate Professor of Biological Sciences (Committee Member)	Dr. Nathan Clark Assistant Professor of Computational and Systems Biology, University of Pittsburgh (Committee Member)
Dr. Philip Reeder Dean, Bayer School of Natural and Environmental Sciences	Dr. Joseph McCormick Chair, Biological Sciences Associate Professor of Biological Sciences

ABSTRACT

MOLECULAR EVOLUTION OF HOMINOID PRIMATES: PHYLOGENY AND REGULATION

By

Ranajit Das

December 2014

Dissertation supervised by Dr. Michael Jensen-Seaman

The complete mitochondrial genome of one eastern gorilla was sequenced to provide the most accurate date for the mitochondrial divergence of gorillas. The most recent common ancestor of eastern lowland and western lowland gorillas existed about 1.9 million years ago, slightly more recent than that of chimpanzee and bonobo. This confirms that the eastern and western gorillas show species level genetic divergence.

Hominoid mating systems differ tremendously. The level of sperm competition varies according to the mating system, which presumably imposes unique selective pressures on the seminal proteins of each species. Cartilage acidic protein 1 (CRTAC1) was identified in our lab as the protein with the largest difference in abundance between human and chimpanzee semen, being found at 142-fold higher in chimpanzee. The coding region of *CRTAC1* is extremely conserved with signature of strong purifying selection. Paradoxically, the *CRTAC1* 'promoter' from human drives transcription

significantly greater than chimpanzee, with or without androgen stimulation. Analyzing H3K27Ac data, a ~2.2kb region was identified as a possible additional *cis*-regulatory element. The *cis*-regulatory region behaved like a silencer and aided in strong transcriptional repression in humans. Although its underlying basis remains elusive, it can be speculated that the differential expression of CRTAC1 between human and chimpanzee seminal plasma results from tissue specific over/under expression of this gene.

The evolutionary history of micro RNAs (miRNAs) within hominoids have remained understudied. The overall goal of this project was to identify the uniquely gained and lost miRNAs and their targets within hominoids. I found 14 miRNAs uniquely gained in humans, the targets of which are associated with brain-associated functions. Older miRNAs were found to be more conserved compared to the newer miRNAs gained within the last 15 million years.

ACKNOWLEDGEMENT

There are a number of people I want to thank for their support during my graduate career at Duquesne University and for making my journey here a memorable one. First, I want to thank my dissertation advisor, Dr. Michael Jensen-Seaman for his support and advice over the years. He has been a terrific mentor and his guidance has been critical for not for the successful completion of my work in his lab but will always be an important guide in my scientific journey ahead.

I want to thank my dissertation committee, Dr. Brady Porter, Dr. David Lampe and Dr. Nathan Clark for their support and valuable advice to my dissertation research and for helping to improve my scientific training at Duquesne University.

I want to thank all the past and present members of the Seaman lab, Dr. Sarah Craig, Scott Hergenrother, Amanda Colvin, Jennifer Vill, Alicia Martinez, and Lindsay Kokoska for their help over the years.

In addition, I want to thank the Selcer, Auron, Castric, McCormick, and Lampe labs at Biological Sciences, Duquesne University for sharing equipment and reagents over the years. Further I would like to thank the Duquesne University Core Sequencing Facility and staff members at the Biological Sciences Main Office for their help and support during my graduate career at Duquesne University.

I want to thank all my family especially my parents, Debjani Das and Laxmikanta Das for their love and unconditional support through all my pursuits. They have not only always been there for me but have always encouraged me to dream bigger and strive harder towards my life goals. Today would not have been possible without their love and blessings.

Finally I want to thank my wife, Dr. Priyanka Upadhyai for her friendship, love, and support and for being my biggest strength over the years. More so, I feel fortunate that we share not only our lives but also a passion for science that has helped surmount many an obstacle over the years.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1: Introduction

1.1 Introduction to Primates

1.1.1 Primates as mammals

According to molecular systematics based classification the placental mammals can be classified into three broad lineages Afrotheria, Xenarthra, and Boreoutheria (Kriegs et al. 2006). Under Boreoutheria, the order Primates is grouped with orders Scandentia, Dermoptera, Rodentia, and Lagomorpha (Fig. 1.1) (Martin 2008). The orders Dermoptera and Primates together form the clade Primatomorpha. Order Dermoptera contains only two extant species of gliding mammals (Stafford 2005). They are commonly known as colugos or 'flying lemurs' (Martin 2008). Colugos are thought to be the closest living sister taxa to Primates (Nie et al. 2008).

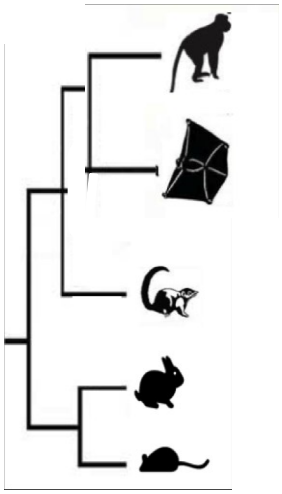


Figure 1.1: Schematic representation of clade Euarchontoglires containing the superorders Euarchonta (Orders Scandentia, Dermoptera, and Primates) and Glires (Orders Rodentia and Lagomorpha)

Primates are a diverse group of mammals. However they share some morphological similarities among each other that make them distinct from other mammals. These characters include stereoscopic vision, which is aided by forward facing

eyes, opposable thumbs and toes, nails instead of claws, slow reproduction rate compared to other similar sized mammals, and extended infancy (Fleagle 1999).

1.1.2 Primate classification

The living primates are grouped into two suborders: Strepsirrhini and Haplorrhini (Groves 2001, Perelman et al. 2011). Some of the most important morphological differences between the two groups include the absence of postorbital closure, and instead the presence of postorbital bar in Strepsirrhini (Fleagle 1999). Strepsirrhinni also have grooming claws and tapetum lucidum (reflective layer in the eye), not seen in Haplorrhinni (Fleagle 1999). A recent primate phylogeny has grouped Strepsirrhini into two infraorders: Lemuriformes (the Madagascar lemurs) and Lorisiformes (lorises and galagos) (Perelman et al. 2011). Perelman et al. 2011 has grouped Haplorrhini primates into two infraorders: Tarsiiformes (tarsiers) and Simiiformes. Simiiformes are further split into two parvorders: Platyrrhini (New World monkeys) and Catarhini (Old World monkeys and apes) (Perelman et al. 2011). The Platyrrhini parvorder contains three New World monkey families Ceboidea, Atelidae, and Pitheciidae, with characteristic prehensile tails. The Catarrhine parvorder is split into two superfamilies: Cercopithecoidea (Old World monkeys) and Hominoidea (apes and humans) (Perelman et al. 2011) (Fig. 2). Cercopithecoidea (Old World monkeys) are further grouped into two subfamilies: Colobinae (langurs, proboscis, African colobus, Asian leaf-monkeys, and snub-nose monkeys) and Cercopithecinae (baboons, mandrills, and macaques) (Groves 2001) (Fig. 1.2). The Hominoidea superfamily is gropued into two families: Hylobatidae (gibbon, siamang) and Hominidae (orangutan, gorilla, chimpanzee, and human). And the family Hominidae is further grouped into two subfamilies: Ponginae

(orangutan) and Homininae (gorilla, chimpanzee, and human) (Perelman et al. 2011) (Fig. 1.2).

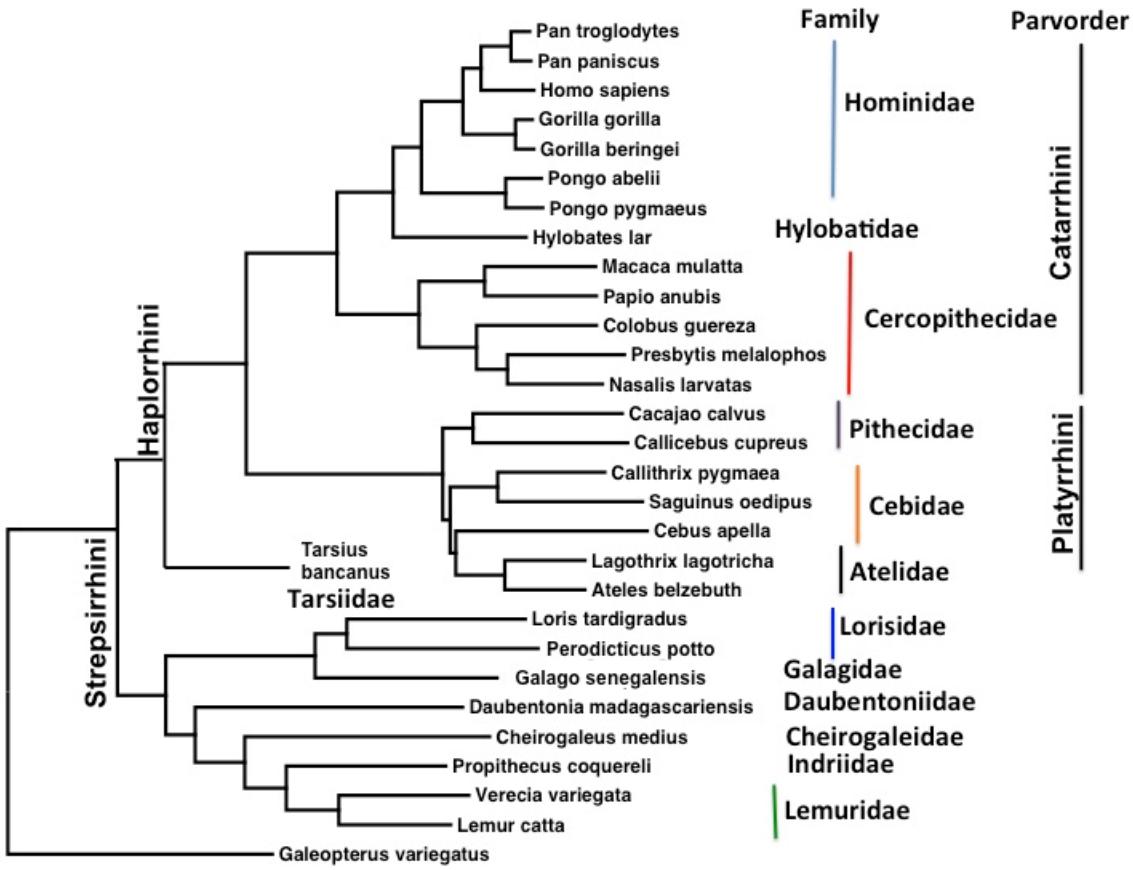


Figure 1.2: Primate phylogenetic tree with colugo as the outgroup. The tree shows various extant families of primates. The phylogeny is reconstructed using the families and parvorders names after Groves (2001) and Perelman et al. (2011).

Haplorhine and Strepshirhine primates split from each other at ~87 Mya. The New World monkeys (Platyrrhini) split from Catarrhine primates at ~43 Mya (Perelman et al. 2011). Old World monkeys and Apes (hominoids) split at ~32 Mya (Perleman et al. 2011). The extant species of the hominoid clade arose at around 20 Mya with gibbons split from hominids. Orangutan split from African apes ~15 Mya, followed by *Gorilla,*

Homo-Pan split ~8 Mya. Humans split from *Pan* species ~6 Mya (See Chapter 2).

1.1.3 Primate habit and habitat

Nonhuman primates are currently found in Africa, Asia, and South America (Fig. 1.3). Primates also occupy restricted areas in Europe (Gibraltar) and North America (southern Mexico), with historically wider ranges in Europe and North America (Fleagle 1999) (Fig. 1.3). The greatest abundance and diversity of Strepsirrhini primates are observed in Madagascar (Fleagle 1999).

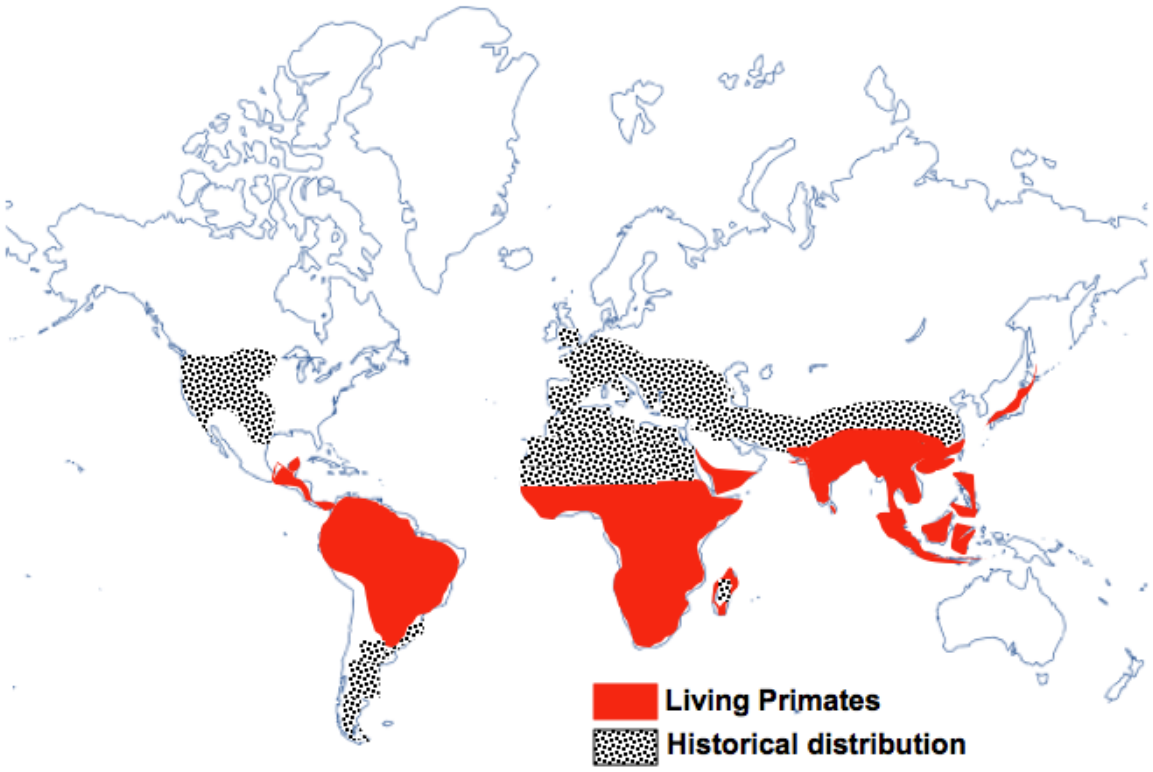


Figure 1.3: The geographical distribution of extant and extinct nonhuman primates (Redrawn after Fleagle 1999)

Primates are found in wide ranges of habitats ranging from deserts to tropical rain forests; however, most primate species are found in tropical forests (Fleagle 1999). Primates also show variation in their diets. Primates are generally of three dietary types:

frugivores (specialize on fruit eating), folivores (specialize on leaf eating), and insectivores (insect feeding) (Fleagle 1999). Species like gibbons show temporal pattern in food preference. They feed on fruits in the morning and leaves at night (Fleagle 1999).

1.2 Sexual selection

Darwin described his idea about sexual selection in his 1871 book '*The Descent of Man and Selection in Relation to Sex'.* His idea of sexual selection developed from the observation that some changes in organisms do not help them to adapt to the environment and hence cannot be explained by natural selection. He thought certain changes evolve to benefit certain individuals over others of the same sex and species in relation to reproduction (Darwin 1871). He divided sexual selection into two categories: operation of sexual selection through competition among members of the same sex for access to members of the other sex (combat) and operation of sexual selection through choice by members of one sex (mostly females) for certain members of the other sex (mostly males) (display). Darwin thought sexual selection through 'display' operates through female choice, where females choose the most striking males to mate with (Darwin 1871).

In more generalized terms sexual selection can be of two types: pre-copulatory sexual selection and post-copulatory sexual selection. Female choice, described by Darwin in his 1871 book, is a type of pre-copulatory intersexual selection. In this type of sexual selection the females choose mates based on male characteristics or displays.

One of the major displays shown by males is auditory display that is, songs and calls (Reviewed in Horth 2007). In the house finch and the European starling the male songs are believed to reflect his quality and are used by females to select mates. Female house finches prefer longer and faster male songs (Nolan and Hill 2004). Female

European starlings like male songs with novel long-bouts (Sockman et al. 2005). Song preference is an important means of female choice among anurans. In the Grey Tree frog (*Hyla versicolor*), females prefer calls of greater duration and sometimes prefer call duration that surpass the normal range exhibited by the species (Ryan 1991). Acoustic displays also play an important role in the sexual selection of insects such as crickets and cicadas. Female preferences for male courtship song parameters in crickets have been documented in several previous studies (Tregenza et al. 2006, Wagner and Reiser 2000, Rantala and Kortet 2003).

Olfactory signals are another important means of female choice. Olfactory signals mainly include sex pheromones, which are chemical signals that can modulate mate choice (Holy et al. 2000). Pheromones are found in all animal taxa but are commonly used by the invertebrates and rodents as a measure of female choice (Reviewed in Horth 2007).

Visual signals are another important means of female choice. The brilliant plumage coloration in peacock, described by Darwin (1871) is the classical example of the visual stimuli. Other examples include bright blue and chestnut coloration of bluebirds, deep-red hues in house finch, and the brilliant mating displays of birds of paradise (Siefferman and Hill 2003, Hill and Farmer 2004, Diamond 1981).

The above-mentioned displays are considered as honest signals of male health and quality (reviewed in Horth 2007). Females choose and mate with healthy males in search of good genes. This preference is passed on to the next generation and the offspring inherit the genes for the mating preference. Thus female choice can potentially lead to

runaway selection, where an increase in male display leads to a selective advantage in males and the females evolve preference for that trait (Fisher 1915).

Another type of pre-copulatory sexual selection is combat (or intrasexual competition), where the members of the same sex fight among each other to get access of the other sex. In this type of sexual selection males evolve specific traits to battle among each other to get the access to the females (Gould and Gould 1989). Male-male combat is seen in wide range of animals. The enlargement of one of the two chelipeds in fiddler crabs is a classic example of male-male combat (Croll and McClintock 2002). Leg spurs in wild turkey are often used fro for male-male combat, when two or more males try to get access to the same female and engage in a battle (Buchholz 1997). Male broad-horned flour beetles develop massively enlarged mandibles, used for male- male combat to get access to the females. Often males with larger mandibles have seen to be better fighters (Okada and Miyatake 2009, Okada et al. 2006). Among mammals the classic example of male weaponry is the antlers of deer and antelopes. In moose, the antlers are used as a weapon for male-male combat and females prefer males with large, symmetric antlers (Rodgers 2001).

Post-copulatory sexual selection is mostly characterized by competition between sperms from two or more males within the female genital tract for fertilizing the egg. This kind of post-copulatory intrasexual selection is known as sperm competition (Parker 1970). The sperm competition often leads to larger testes, as they are required to accumulate a larger mass of seminiferous tubules-the sperm producing tissues, which adds to the larger amount of sperm in the ejaculates (Dixson 1993). Due to high sperm competition existing in their society, the *Pan* species possess the largest testes among the

hominoids relative to their body size (Harcourt et al. 1981). Beside primates, larger testes are found in many promiscuous mammals including some species of bats (Wilkinson and McCracken 2003), cetaceans (Connor 2000), marsupials, and monotremes (Rose et al. 1997). Also, gynandrous mammals show higher testes to body weight ratio compared to the polygynous and monogamous mammals (Fig. 1.4).

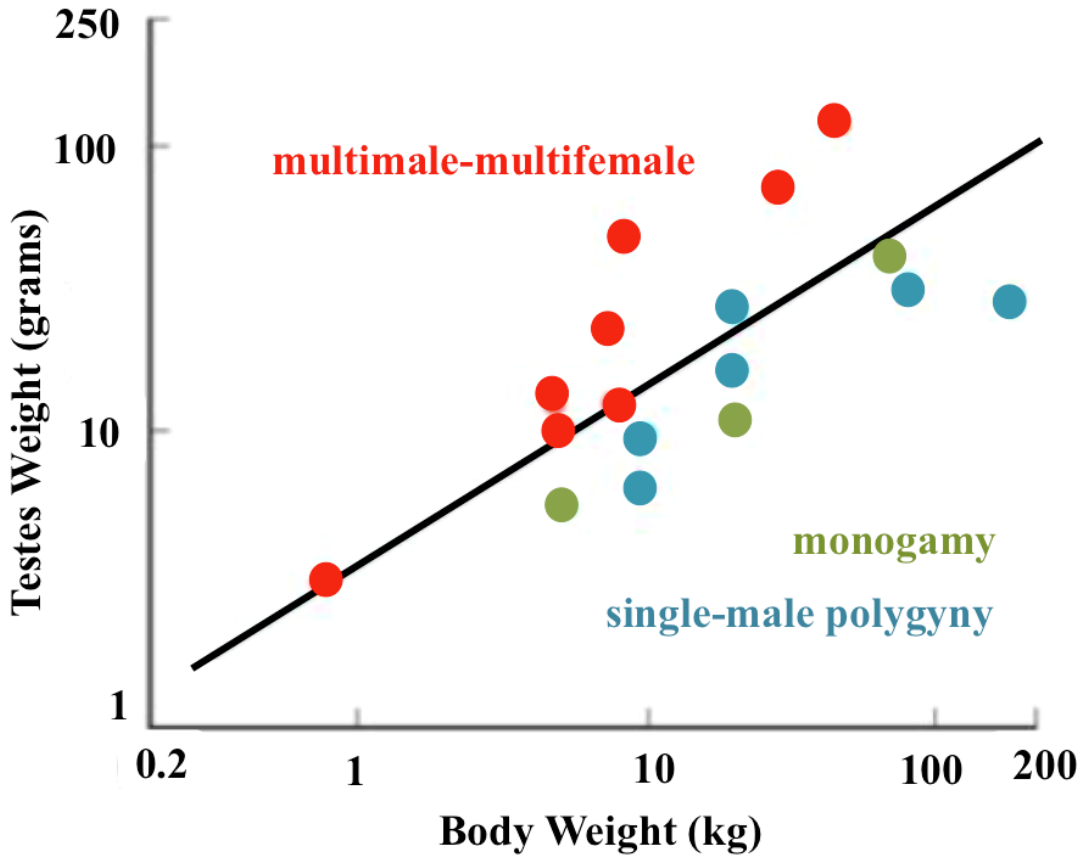


Figure 1.4: Testes to body weight ratio among species with different mating systems. Figure modified from Harcourt et al. 1981 with permission from the journal

High sperm competition may lead to several different modifications in the sperm. For example in chimpanzee with high sperm competition, has a higher sperm motility (Møller 1988), higher sperm concentration in the ejaculates (Møller and Brickhead 1989), and higher ratio of seminiferous tubule to connective tissue (Harvey and Harcourt 1984).

Moreover, the chimpanzee sperm swim faster than its human counterpart (Nascimento et al. 2008), has significantly larger mid-piece volume that contains energy producing mitochondria (Anderson and Dixson 2002), and has higher mitochondrial membrane potential (Anderson et al. 2007). Sperm mid-piece possess mitochondria that supply energy for the flagellar movement. Larger sperm mid-piece indicates the presence larger volume of mitochondria, which in turn assures better sperm motility.

Sperm competition may have also caused several chemical changes in the ejaculates. Chemicals derived from the seminal vesicles and prostate help in seminal coagulation soon after ejaculation and may form a compact structure called copulatory plug or mating plug (Dixson and Anderson 2002). A copulatory plug may help in sperm positioning, prevention of sperm loss, and generation of a physical barrier (Dixson 2012). Firstly by generating a physical barrier, it prevents the entry of other sperm to the female genital tract and secondly, it minimizes the sperm loss and protects the sperm till they reach the uterus (Dixson and Anderson 2002). Among hominoids, the *Pan* species with high sperm competition produce copulatory plugs. Beside primates, copulatory plug formation can be observed in wide variety of animals including kangaroos (Dawson 2012), scorpions (Contreras-Garduno 2006), mice (Ittner and Jürgen 2007), and ground squirrels (Monroe and Koprowski 2012).

Another interesting form of post-mating sexual selection is cryptic female choice. It can be defined as the ability of the females to store and separate sperm from multiple males and regulate paternity by choosing the 'best' sperm for fertilizing their eggs (Eberhard and Cordero 1995). Cryptic female choice can be observed in various egg-laying animals including birds (Wagner et al. 2004), spenodontians (Moore et al. 2009),

gastropods (Beese et al. 2009), arachnids (Welke and Schneider 2009), insects (Ward 2000), and testudines (Holt and Lloyd 2010). In this type of sexual selection, females commonly choose the 'best' sperms on the basis of biochemical signals between the proteins from seminal plasma and female genital tract (Prokupek et al. 2008).

1.3 Eukaryotic gene regulation

Eukaryotic gene regulation is a complex process that includes gene accessibility and transcription, mRNA processing, translation, and post-translational modifications. In this chapter I shall focus my discussion on transcription initiation, the first and arguably the most important steps of eukaryotic gene regulation.

Transcription of eukaryotic genes requires a precise orchestration of a set of interactions among numerous *trans*-acting proteins and DNA sequences (*cis*-regulatory modules). The cis-regulatory modules include different types of regulatory sequences such as promoters, enhancers, and silencers. The promoters can be of two types: upstream promoter elements (UPE) and downstream promoter elements (DPE) (Maston et al. 2006). Eukaryotic genes are transcribed by the enzyme RNA Polymerase II. The *trans*-acting proteins, including the general transcription factors (TFs), cooperate with each other for the optimum binding of the RNA polymerase II to the promoter elements of genes, which may or may not contain a TATA box consensus sequence (Gaston and Jayaraman 2003).

The binding of RNA polymerase II to the promoter elements is a complex process because eukaryotic DNA is packaged as complex chromatin structure. The chromatin packaging is aided by the nucleosome that consists of 147 bp of DNA wrapped around a highly conserved histone protein octamer containing two copies each of the core histones

H2A, H2B, H3 and H4 (Luger et al. 1997, Li and Reinberg 2011). This histone-DNA association hinders the accessibility of DNA to RNA polymerase II and other general transcription factors. Thus chromatin remodeling and covalent modification of the amino terminal 'tails' of histones to alter chromatin compaction are necessary steps for gene accessibility and transcription initiation (Clapier and Cairns 2009, Fischle et al. 2003, Shogren-Knaak et al. 2006).

Transcription initiation involves a set of general transcription factors including TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, which assembles at the core promoter (which may include a TATA box) to form a pre-initiation complex (Fig. 1.5). The formation of pre-initiation complex with TFIID (TBP) binding to TATA box and directs the RNA polymerase II to the transcription start site (TSS) (Fig. 1.5). This general transcription machinery functions in all eukaryotic genes. However, the spatial and temporal control of gene expression is aided by additional *trans*-acting factors called regulatory transcription factors that binds to the *cis*-regulatory modules in a sequence-specific manner and can regulate gene expression from a long distance from the target gene promoter (Lomvardas et al. 2006, Sanyal et al. 2012, Sudou et al. 2012). These regulatory transcription factors can regulate transcription both positively (transcriptional activators) and negatively (transcriptional repressors by controlling the chromatin structure - compaction, covalent modification of histones, and nucleosome positioning. Activator proteins may also help in the formation of pre-initiation complex by interacting with the components of the basal transcription machinery and sub-units of the Mediator complex (Ge et al. 2002, Bhaumik et al. 2004). The activators sometimes recruit additional regulatory proteins known as co-activators that lack DNA binding activity.

Together with the co-activators, activators recruit histone-modifying enzymes such as histone acetyltransferases that helps in chromatin decondensation and transcription initiation (Ogryzko et al. 1996, Akimaru et al. 1997). Similarly, the repressor proteins can recruit co-repressors, which in turn can recruit chromatin-remodeling enzymes that suppress transcription initiation by forming inactive heterochromatins (Li et al. 2007).

The distal regulatory elements of transcription include enhancers, silencers, and insulators. The insulators form a boundary that blocks the interaction between promoters and additional *cis*-regulatory elements (Burgess-Beusse et al. 2002). They are very often found in between promoters and enhancers, and actively participate in high-order nuclear organization together with other *cis*-regulatory elements (Ong and Corces 2011). The term 'enhancer' is often used to include *cis*-regulatory modules that can either promote or antagonize (sometimes called 'silencers') the assembly of the basal transcription machinery at the target gene promoters. These *cis*-regulatory elements regulate transcription in spatial (tissue or gradient) and/or temporal (developmental stage) manner (Ong and Corces 2011). Here I am using the term 'enhancer' for both activators and silencers. Enhancers are often found at long distances away (>10kb) from their target genes and may even be situated on different chromosomes (Lomvardas et al. 2006, Ong and Corces 2011). In case of the highly expressed housekeeping genes the enhancers can be found up to 150kb from the promoters and are involved in looping interactions with the promoters to regulate the target genes (Noordermeer et al. 2008). The looping interactions require proper repositioning of the target loci aided by the repositioning of the nucleosomes (Noordermeer et al. 2008). Techniques like chromatin conformation capture (3C) or the various variations of the technique and/or fluorescent in situ

hybridization (FISH) are employed to identify the right gene specific or tissue specific enhancers by showing physical association between genomic elements within the nucleus (Ong and Corces 2011). Enhancers can also be identified by studying long range looping interactions between enhancers and promoters (Carter et al. 2002, Dekker et al. 2002). Several vertebrate and insect genes such as the β-globin locus, *H19, IGF2, MYB,* and *Abd B* have been identified via loopin interactions (Tolhuis et al. 2002, Murrell et al. 2004, Sipos and Gyurkovics 2005, Degner et al. 2011, Stadhouders et al. 2012). A classic example of looping interaction is shown by the β-globin locus and was identified by the 3C technique. This locus is located ~ 40-60kb away from the active genes, but regulate the specific globin gene, appropriate for a particular developmental stage, by looping interaction (Tolhuis et al. 2002).

In recent years various different non-coding RNAs (ncRNAs) have been identified that can actively participate in gene regulation. These RNAs do not code for protein but instead regulate other mRNAs. Some of these regulatory RNAs include miRNAs (See Chapter 4), enhancer RNAs (eRNA) (Kim et al. 2010), and lincRNAs (Orom et al. 2010). eRNAs facilitates gene activation by interacting with other factors and thus play enhancer like function (Lei and Corces 2006). Long intergenic ncRNAs (lincRNAs) aid in the expression of their neighboring protein-coding genes. They also help in transcription activation by regulating the assembly of transcription factors or other chromatin remodeling enzymes at the promoter (Ong and Corces 2011).

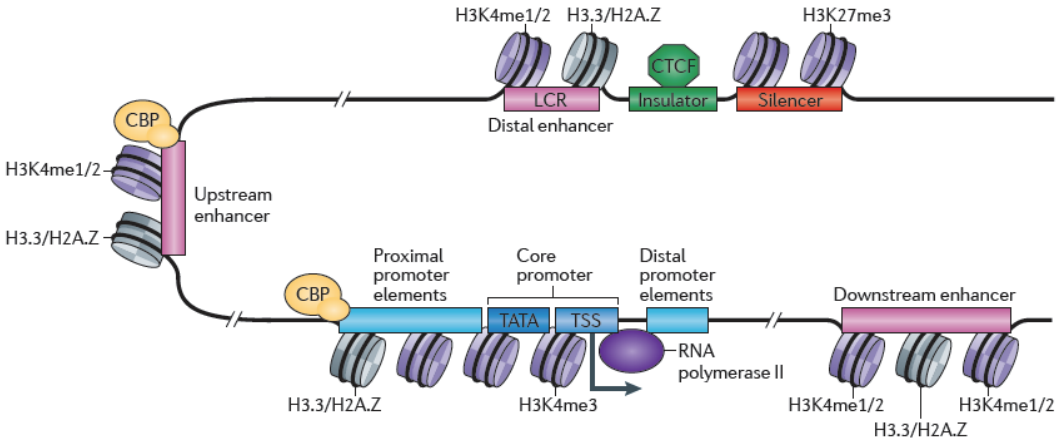


Figure 1.5: Various *cis* and *trans* regulatory elements involved in eukaryotic transcriptional regulation. Figure from Ong and Corces 2011 with permission from the journal

References

Akimaru H, Chen Y, Dai P, Hou DX, Nonaka M, Smolik SM, Armstrong S, Goodman RH, Ishii S 1997. Drosophila CBP is a co-activator of cubitus interruptus in hedgehog signalling. Nature 386:735-738.

Anderson MJ, Chapman SJ, Videan EN, Evans E, Fritz J, Stoinski TS, Dixson AF, Gagneux P 2007. Functional evidence for differences in sperm competition in humans and chimpanzees. Am J Phys Anthropol. 134:274-280.

Anderson MJ, Dixson AF 2002. Motility and the midpiece in primates. Nature 416:496.

Beese K, Armbruster GFJ, Beier K, Baur B 2009. Evolution of female sperm-storage organs in the Carrefour of stylommatophoran gastropods. J Zool Syst Evol Res. 47:49-60.

Bhaumik SR, Raha T, Aiello DP, Green MR 2004. vivo target of a transcriptional activator revealed by fluorescence resonance energy transfer. Genes Dev. 18:333-343.

Buchholz R 1997. Male dominance and variation in fleshy head ornamentation in wild turkeys. J Avian Biol. 28:223-230.

Burgess-Beusse B, Farrell C, Gaszner M, Litt M, Mutskov V, Recillas-Targa F, Simpson M, West A, Felsenfeld G 2002. The insulation of genes from external enhancers and silencing chromatin. Proc Natl Acad Sci. 99:16433-16437.

Carter D, Chakalova L, Osborne CS, Dai YF, Fraser P 2002. Long-range chromatin regulatory interactions in vivo. Nat Genet. 32:623-626.

Clapier CR, Cairns BR 2009. The biology of chromatin remodeling complexes. Annu Rev Biochem. 78:273-304.

Connor RC. 2000. Group living in whales and dolphins. In: Mann J, Connor R, Tyack P, Whitehead H, editors. Cetacean societies:field studies of whales and dolphins.

Contreras-Garduno J, Peretti AV, Alex CA 2006. Evidence that Mating Plug is Related to Null Female Mating Activity in the Scorpion Vaejovis punctatus. Ethology 112:152.

Croll GA, McClintock B 2002. An analysis of cheliped asymmetry in three species of fiddler crabs. Gulf Mex Sci. 2:106-109.

Darwin C. 1871. The descent of man and selection in relation to sex. John Murray, Albemarle Street:London.

Das RH, S. D. Soto-Calderón Id, Dew J. L. Anthony N. M. Jensen-Seaman M. I. 2014. Complete mtDNA sequence of the eastern gorilla (Gorilla beringei) and its implication for African Ape biogeography. J Hered. In press.

Dawson T. 2012. Kangaroos:Csiro Publishing.

Degner SC, Verma-Gaur J, Wong TP, Bossen C, et al. 2011. CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. Proc Natl Acad Sci. 108:9566-9571.

Dekker J, Rippe K, Dekker M, Kleckner N 2002. Capturing chromosome conformation. Science 295:1306-1311.

Diamond JM 1981. Birds of paradise and the theory of sexual selection. Nature 293:257-258.

Dixson AF 1993. Sexual selection, sperm competition and the evolution of sperm length. Folia Primatologica 61:221-227.

Dixson AF. 2012. Primate sexuality:comparative studies of the prosimians, monkeys, apes, and humans. New York, NY:Oxford University Press.

Dixson AL, Anderson MJ 2002. Sexual selection, seminal coagulation and copulatory plug formation in primates. Folia Primatologica 73:63-69.

Eberhard WG, Cordero C 1995. Sexual selection by cryptic female choice on male seminal products-a new bridge between sexual selection and reproductive physiology. Tree 10:493-496.

Fischle W, Wang Y, Jacobs SA, Kim Y, Allis CD, Khorasanizadeh S 2003. Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. Genes Dev. 17:1870-1881.

Fisher RA 1915. The evolution of sexual preference. Eugenics Review 7:184-192.

Fleagle JG. 1999. Primate adaptation and evolution. San Diego, CA:Academic Press.

Gaston K, Jayaraman PS 2003. Transcriptional repression in eukaryotes:repressors and repression mechanisms. Cell Mol Life Sci. 60:721-741.

Ge K, Guermah M, Yuan CX, Ito M, Wallberg AE, Spiegelman BM, Roeder RG 2002. Transcription coactivator TRAP220 is required for PPAR gamma 2-stimulated adipogenesis. Nature 417:563-567.

Gould JL, Gould CG. 1989. Sexual selection. New York, NY:Scientific American Library.

Groves C. 2001. Primate taxonomy. Washington: Smithsonian Institution Press.

Harcourt AH, Harvey PH, Larson SG, Short RV 1981. Testis weight, body weight and breeding system in primates. Nature 293:55-57.

Harvey PH, Harcourt AH. 1984. Sperm competition, testes size, and breeding systems in primates. In: Smith RL, editor. Sperm competition and the evolution of animal mating systems. Orlando: Academic Press.

Hill GE, Farmer KL 2004. Carotenoid-based plumage coloration predicts resistance to a novel parasite in the house finch. Naturwissenschaften 92:30-34.

Holt W, Lloyd R 2010. Sperm storage in the vertebrate female reproductive tract: How does it work so well. Theriogenology 73:713-722.

Holy TE, Dulac C, Meister M 2000. Responses of vomeronasal neurons to natural stimuli. Science 289:1569-1572.

Horth L 2007. Sensory genes and mate choice:Evidence that duplications, mutations, and adaptive evolution alter variation in mating cue genes and their receptors. Genomics 90:159-177.

Ittner LM, Jürgen G 2007. Pronuclear injection for the production of transgenic mice. Nat Prot. 2:1206-1215.

Kim TK, Hemberg M, Gray JM, Costa AM, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182-187.

Kriegs JO, Churakov G, Martin K UJJBJS 2006. Retroposed Elements as Archives for the Evolutionary History of Placental Mammals. PLoS Biology 4:e91.

Lei EP, Corces VG 2006. RNA interference machinery influences the nuclear organization of a chromatin insulator. Nat Genet. 38:936-941.

Li B, Carey M, Workman JL 2007. The role of chromatin during transcription. Cell 128:707-719.

Li G, Reinberg D 2011. Chromatin higher-order structures and gene regulation. Curr Opin Genet Dev. 21:175-186.

Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R 2006. Interchromosomal interactions and olfactory receptor choice. Cell 126:403-413.

Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389:251-260.

Martin RD 2008. Colugos:obscure mammals glide into the evolutionary limelightJ Biol. J Biol. 7:13.

Maston GA, Evans SK, Green MR 2006. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet. 7:29-59.

Møller AP 1988. Ejaculate quality, testes size and sperm competition in primates. J Hum Evol. 17:479-488.

Møller Ap BTR 1989. Copulation behavior in mammals:evidence that sperm competition is wide-spread. Biol J Linn Soc. 38:119-131.

Moore JA, Daugherty CH, Godfrey SS, Nelson NJ 2009. Seasonal monogamy and multiple paternity in a wild population of a territorial reptile (tuatara. Biol J Linn Soc. 98:161-170.

Munroe KE, Koprowski JL 2012. Copulatory Plugs of Round-Tailed Ground Squirrels (Xerospermophilus tereticaudus). Southwest Nat. 57:208-210.

Murrell A, Heeson S, Reik W 2004. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. Nat Genet. 36:889-893.

Nascimento JM, Shi LZ, Meyers S, Gagneux P, Loskutoff NM, Botvinick EL, Berns MW 2008. The use of optical tweezers to study sperm competition and motility in primates. J R Soc Interface 5:297-302.

Nie W, Fu B, O'Brien PCM, Wang J, et al. 2008. flying tree-shrews? Molecular cytogenetic evidence for a Scandentia-Dermoptera sister clade. BMC Biology 6:18.

Nolan PM, Hill GE 2004. Female choice for song characteristics in the house finch. Anim Behav. 67:403-410.

Ogryzko VV, Schiltz RL, Russanova V, Howard Bh NY 1996. The transcriptional coactivators p300 and CBP are histone acetyltransferases. Cell 87:953-959.

Okada K, Miyanoshita A, Miyatake T 2006. Intra-sexual dimorphism in male mandibles and male aggressive behavior in the broadhorned flour beetle Gnatocerus cornutus. Insect Behav J. 19:457-467.

Okada K, Miyatake T 2009. Genetic correlations between weapons, body shape and fighting behaviour in the horned beetle Gnatocerus cornutus. Anim Behav. 77:1057-1065.

Ong CT, Corces VG 2011. Enhancer function:new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 12:283-293.

Orom UA, Derrien T, Beringer M, Gumireddy K, et al. 2010. Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46-58.

Parker GA 1970. Sperm competition and its evolutionary consequences in the insects. Biol Rev. 45:525-567.

Perelman P, Johnson WE, Roos C, Seuánez HN, et al. 2011. A molecular phylogeny of living primates. PLoS Genetics 7:e1001342.

Prokupek A, Hoffmann F, Eyun SI, Moriyama E, Zhou M, Harshman L 2008. An evolutionary expressed sequence tag analysis of Drosophila spermatheca genes. Evolution 62:2936-2947.

Rantala MJ, Kortet R 2003. Courtship song and immune function in the field cricket Gryllus bimaculatus. Biol J Linn Soc. 79:503-510.

Rodgers A. 2001. Appearance and characteristics of Moose:Voyager Press.

Rose RW, Nevison CM, Dixson AF 1997. Testes weight, body weight and mating systems in marsupials and monotremes. J Zool. 243:523-531.

Ryan MJ 1991. Sexual selection and communication in frogs. Trends Ecol Evol. 6:351-355.

Sanyal A, Lajoie BR, Jain G, Dekker J 2012. The long-range interaction landscape of gene promoters. Nature 489:109-113.

Shogren-Knaak M, Ishii H, Sun JM, Pazin MJ, Davie JR, CL P 2006. Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science 311:844-847.

Siefferman L, Hill GE 2003. Structural and melanin coloration indicate parental effort and reproductive success in male eastern bluebirds. Behav Ecol. 14:855-861.

Sipos L, Gyurkovics H 2005. Long-distance interactions between enhancers and promoters. FEBS Journal 272:3253-3259.

Sockman KW, Gentner TQ, Ball GF 2005. Complementary neural systems or the experience-dependent integration of mate choice cues in European Starlings. J Neurobiol. 62:72-81.

Stadhouders R, Thongjuea S, Andrieu-Soler C, Palstra RJ, et al. 2012. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. EMBO Journal 31:986-999.

Stafford BJ. 2005. Order Dermoptera. In: Wilson DE RD, editor. Mammal species of the world Johns Hopkins University Press. p. 110.

Sudou N, Yamamoto S, Ogino H, Taira M 2012. Dynamic in vivo binding of transcription factors to cis-regulatory modules of cer and gsc in the stepwise formation of the Spemann-Mangold organizer. Development 139:1651-1661.

Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat w 2002. Looping and interaction between hypersensitive sites in the active β-globin locus. Mol Cell. 10:1453-1465.

Tregenza T, Simmons LW, Wedell N, Zuk M 2006. Female preference for male courtship song and its role as a signal of immune function and condition. Anim Behav. 72:809-818.

Wagner RH, Helfenstein DE 2004. Female choice of young sperm in a genetically monogamous bird. Proc Roy Soc. 271:134-137.

Wagner WE, Reiser MG 2000. The importance of calling song and courtship song in female mate choice in the variable field cricket. Anim Behav. 59:1219-1226.

Ward PI 2000. Cryptic female choice in the yellow dung fly Scathophaga sterocoraria. Evolution 54:1680-1686.

Welke K, Schneider JM 2009. Inbreeding avoidance through cryptic female choice in the cannibalistic orb-web spider Argiope lobata. Behav Ecol. 20:1056-1062.

Wilkinson GS, McCracken GF. 2003. Bats and balls:sexual selection and sperm competition in the Chiroptera. In: Kunz TE, Fenton MB, editors. Bat ecology. Chicago, IL:University of Chicago Press. p. 128-155.

Chapter 2: Evolutionary history of gorillas inferred from complete mitochondrial genome sequences

[This is a pre-copyedited, author-produced PDF of an article accepted for publication in Journal of heredity following peer review. The version of record Das R, Hergenrother SD, Soto-Calderón ID, Dew JL, Anthony NM, Jensen-Seaman MI (2014) Complete mtDNA sequence of the eastern gorilla (Gorilla beringei) and its implication for African Ape biogeography is available online at: http://jhered.oxfordjournals.org/content/early/2014/09/04/jhered.esu056.full.pdf+html, DOI: 10.1093/jhered/esu056]

2.1 Introduction

2.1.1 Brief introduction to gorilla phylogeography

Gorilla is one of the three living African ape genera (other two are *Homo* and *Pan*) restricted to equatorial Africa (Fig 2.1). Currently gorillas are found in four fragmented populations. The largest population (the western gorilla) is found scattered around in Republic of Congo, Gabon, and Equatorial Guinea. A smaller population called Cross River gorilla are found in southwest Cameroon. Two gorilla populations are found in the Democratic Republic of Congo. The larger population is called Eastern Lowland gorilla. The other population is found in Virunga volcanic mountains of the Democratic Republic of Congo, Uganda, and Rwanda (mountain gorilla) (Fig. 2.1).

Western lowland gorilla *(G. g. gorilla)*
Cross-river gorilla *(G. g. dielhi)*
Mountain gorilla *(G. b. beringei)*
Eastern lowland gorilla *(G. b. graueri)*

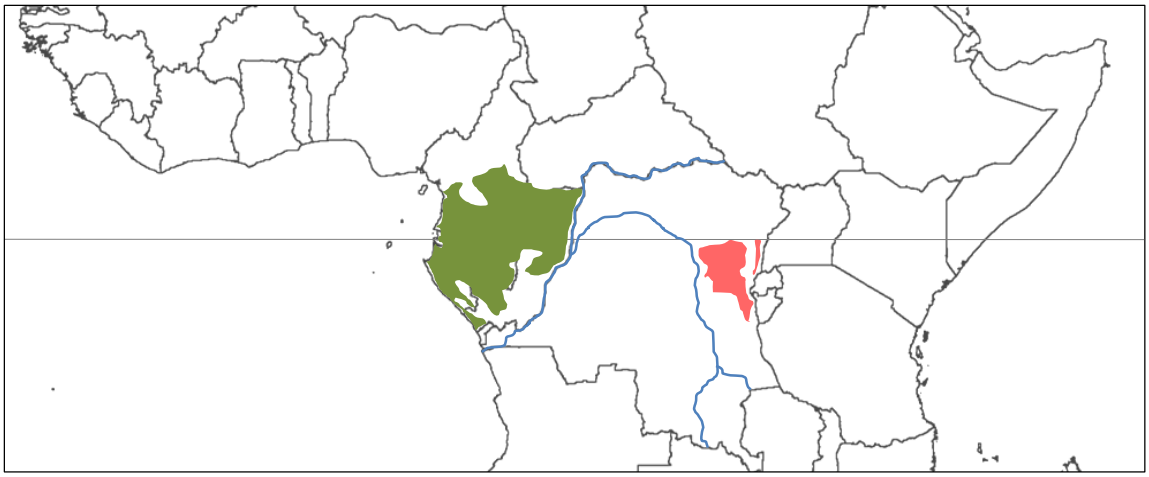


Figure 2.1: Current geographical locations of four gorilla populations. Populations are marked with red boundaries

The two larger gorilla populations (Eastern Lowland and Western Lowland) are separated by nearly 1000 km (Jensen-Seaman and Kidd 2001). Large rivers like Congo, and Ubangui along with open woodlands and savannas surrounding these rivers are potential barriers to the gene flow between the two gorilla populations. Traditionally, the living gorilla populations were considered a single species (*Gorilla gorilla*) with three recognized subspecies (*G. g. gorilla, G. g. beringei, G. g. graueri*) (Groves 2003). Currently many authors consider gorillas to be two species: (1) the western gorilla (*G. gorilla*) comprising two subspecies: the western lowland gorilla (*G. g. gorilla*) and the Cross River gorilla (*G. g. diehli*), and (2) the eastern gorilla (*G. beringei*) comprising two subspecies: the mountain gorilla (*G. b. beringei*) and the eastern lowland gorilla (*G. b. graueri*) (Groves 1996, 2001; Sarmiento and Butynski 1996).

An analysis on the *HV1* region of D-Loop of all available gorilla mtDNAs revealed that gorillas mainly belong to four haplogroups A, B, C, and D (Anthony et al. 2007). The greatest genetic divergence was found to exist between the eastern (haplogroups A and B) and western (haplogroup C and D) gorillas (Anthony et al. 2007). The authors found that the western lowland gorillas ($\theta = 0.047$) are ~2 times more diverse than the eastern lowland gorillas ($\theta = 0.029$). Another study on the D-Loop DNA sequence diversity in several populations of eastern gorillas revealed that haplotypes from eastern gorillas belong to two distinct clades (Jensen-Seaman and Kidd 2001). One clade exclusively included the individuals from the eastern lowland gorillas and the other exclusively included individuals from the mountain gorilla population. Very low level of genetic diversity was found within each clade (Jensen-Seaman and Kidd 2001). This study indicated that the eastern lowland gorillas are reciprocally monophyletic and genetically distinct from mountain gorillas. Although both of the above-mentioned studies were performed on a large sample size, they only focused on ~300bp *HV1* region of D-Loop. So, the genetic diversities and split dates obtained in these studies may not be very robust and conclusive. Since the entire mitochondrial genome was never sequenced from the eastern lowland gorilla before the current study, all previous studies only relied on the D-Loop and/or parts of mtDNA. The use of the entire mtDNA for the prediction of genetic diversity and split times makes the current study better, robust and more conclusive than all previous studies.

2.1.2 Bayesian inference of phylogeny

Bayesian statistics was originally proposed in 18th century. In the mid 20th century Felsenstein (1968) first proposed its utility for phylogeny reconstruction, but construction

of Bayesian phylogeny became popular only recently. It was used in the 1990s (Rannala and Yang 1996, Mau 1996, Li 1996) for phylogeny reconstruction, and since then many authors have shown interest in Bayesian approaches (Huelsenbeck et al. 2001, 2002, Lewis 2001). This popularity is largely due to the availability of various computer programs for phylogeny reconstruction using Bayesian approach including BEAST v1.7.5 (Drummond and Rambaut 2007) and MrBayes v3.2.2 (Huelsenbeck and Ronquist 2001).

2.1.2.1 Introduction to Bayesian statistics

Bayes' theorem is a type of conditional probability, which differs from conventional statistics since it includes 'prior knowledge' for hypothesis testing. Mathematically Bayes' theorem is expressed as:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Where, P(B|A)/P(B) is the likelihood of an event and P(A) is a prior knowledge (prior probability) concerning the event. The outcome is the posterior probability of the event.

For example, suppose a new cancer-determining test has 99% likelihood that it will determine true cancer positives (and 99% likelihood that it will determine true negatives). In other words there is 1% likelihood that the test will give a false negative. The test also has a 0.1% likelihood of giving false positive. So, solely from the likelihood framework, the test has ~99% chance of being correct. Now suppose we add a prior knowledge [P(A)] that only 0.1% people in the population can have that kind of cancer. There are 100,000 individuals in the population, and so there will be 100 cancer victims and 99 of them will be identified as cancer-positives (since the test is 99% accurate). Out of 99,900 healthy individuals 0.1% (~100) individuals will be tested false positives (since

the test determines 0.1% false-positives). So, from the Bayesian framework, the accuracy of the test goes down from being 99% to ~50% (99/(100+99) = 99/199). Incorporation of prior knowledge thus can greatly influence the outcome of an event.

2.1.2.2 Bayesian phylogeny vs. other phylogenies: pros and cons

Conventionally, there are three basic methods of phylogeny reconstruction: Distance methods (Neighbor joining, UPGMA), Maximum parsimony, and Maximum likelihood. Distance-matrix methods calculate genetic distance from multiple sequence alignments using non-parametric methods. They are simplest to implement, and do not require the nucleotide substitution model to be specified. Maximum parsimony (MP) is another simple method for phylogeny reconstruction. It considers a tree to be 'most preferred' that includes minimum number of evolutionary changes to explain the given data. More advanced methods include maximum likelihood (ML) method. It uses the optimality criterion to determine the best tree, and apply a nucleotide substitution model to estimate the phylogenetic tree. The relative pros and cons of these methods have been debated for long time (Faith 1985, Swofford and Olsen 1990, Kunhner and Felsenstein 1994, Huelsenbeck 1995, Farris et al. 1996, Lewis 1998, Steel and Penny 2000). One of the well-known problems with maximum parsimony is 'Long-branch attraction'. It suggests that when rates of evolutionary changes vary greatly among branches, maximum parsimony method may not be the best choice (Felsenstein 1978; Siddall 1998). Since maximum likelihood method is dependent on proper choice of nucleotide substitution model, it often becomes inconsistent if proper model is not chosen (Farris 1999). The distance methods are highly susceptible to evolutionary rate variation and thought to perform less efficiently than both parsimony and likelihood based methods (Huelsenbeck

and Hillis 1993). However, ML and MP have often found to perform with similar efficiency and generate identical tree topologies (Reed et al. 2002, Kimball et al. 2003).

In terms of phylogeny reconstruction, Bayesian inference is quite similar to ML method. Like ML this method also includes a likelihood function (see section 2.1.2.1) and depends on proper selection of nucleotide substitution model. So, like ML, this method is also susceptible to proper evolutionary model selection. However, the aspect that sets Bayesian method apart from ML is the application of prior knowledge regarding the relationship among taxa (for e.g. older separation events, newer separation events), for phylogeny reconstruction. This is done by explicitly stating a prior probability distribution (e.g. normal vs. lognormal vs. exponential) before phylogeny reconstruction. Uniform priors allow one to set up an upper and lower bound of a certain parameter (like constant population size). Normally distributed priors allow the parameter to select values from a normal distribution with certain mean and standard deviation. Log normally distributed priors, similarly, allow the parameter to select value from a log-normal distribution (Drummond and Rambaut 2007). Several authors (Ho 2007, Bjork et al. 2011) think lognormally distributed priors perform better than normally distributed priors when using fossil calibration points for dating the tree as it will sample values from the more distant past more frequently than recent. Log-normally distributed priors are ideal for small populations that are genetically highly structured with 'unreal' increase in effective population size (Drummond and Rambaut 2007). Proper selection of priors can assist in determining true phylogenies, but improper prior selection can lead to inaccurate estimation of phylogeny (Archibald et al. 2003).

2.1.2.3 Technical details of Bayesian phylogeny

Bayesian phylogeny reconstruction involves Markov chain Monte Carlo (MCMC) method (Metropolis et al. 1953; Green 1995). MCMC is a class of algorithm that samples from probability distributions based on constructing a Markov chain. The Markov chain undergoes transition from one state to the other in a 'memoryless' way and the next step depends only on the current step, not on the sequence of events before it. MCMC simulation provides a sophisticated and computationally efficient way of approximating posterior probabilities of trees and other parameters (Huelsenbeck et al. 2002).

The posterior probability describes the probability of trees, considering prior probabilities, model parameters and the data itself. The posterior probability of a tree is calculated through the following steps. Firstly, all trees from species s are labeled from 1 (τ_1) to B(s) (τ_{Bs}), where B(s) is the number of possible trees for s. The data (for e.g. DNA sequences) are denoted by X. So, for i^{th} tree the posterior probability ($Pr[\tau_i|X]$) is calculated as:

$$\Pr[\tau_i \mid \mathbf{X}] = \frac{\Pr[\mathbf{X} \mid \tau_i] \times \Pr[\tau_i]}{\sum_{j=1}^{B(s)} \Pr[\mathbf{X} \mid \tau_j] \times \Pr[\tau_j]},$$

where, $Pr(X|\tau_i)$ is the likelihood of tree i, $Pr(\tau_i)$ is the prior probability of tree i, and the denominator is a normalizing constant that involves a summation over all B(s) possible trees. $B(s) = (2s-3)!/[2^{s-2}(s-2)!]$ for rooted trees (adapted from Huelsenbeck et al. 2002).

The likelihood value $Pr(X|\tau_i)$ depends on several different parameters like values of the parameters in the substitution model (θ), and the lengths of the branches on the tree (ν) (expected number of substitutions per site). So, $Pr(X|\tau_i)$ is expressed as:

$$\Pr[\mathbf{X} \mid \tau_i] = \int_{v_i, \theta} f(\mathbf{X} \mid \tau_i, \upsilon_i, \theta) f(\upsilon_i, \theta) d\upsilon_i \, d\theta$$

where, $f(v_i, \theta)$ is the prior probabilities of branch lengths and model parameters (adapted from Huelsenbeck et al. 2002).

Distribution of trees is generated as the major product of Bayesian phylogeny. The posterior probability distribution can be summarized into a tree-like form in several different ways (Archibald et al. 2003). One of the commonly used trees is maximum posterior probability estimate of phylogeny (MAP) (Huelsenbeck et al. 2002, Archibald et al. 2003). It is the single tree with maximum probability. Another commonly used tree is the majority rule consensus tree that summarizes the distribution of all generated trees (Archibald et al. 2003). All Bayesian approaches summarize the results into 95% credibility interval (Highest posterior density, HPD) for all parameters of interest using posterior probability distribution. This approach is analogous to 95% confidence intervals in standard statistics (Huelsenbeck et al. 2002).

2.1.3 Introduction to mitochondrial genome in respect to phylogenetics

Mitochondrial DNA (mtDNA) is a circular DNA found inside mitochondria, the 'power house' of the cell. mtDNA is maternally inherited through mothers ovum. Mammalian mtDNA is ~16kb long. On average each human mitochondrion contains 5 mtDNA molecules (range 1-15) (Satoh and Kuroiwa 1991). MtDNA has two strands: the guanine rich heavy strand (H strand) and the cytosine rich light strand (L strand). MtDNA contains 13 protein-coding genes. Of these 13 genes, 12 (*ATP6, ATP8, COI, COII, COIII, Cytb, ND1, ND2, ND3, ND4, ND4L,* and *ND5*) are found on the heavy strand and *ND6* is found on the light strand. All 13 proteins encoded by mtDNA in

mammals participate in the electron transport chain. mtDNA has 22 transfer RNA (tRNA) genes, and 2 rRNA genes for the large (16S) and small (12S) subunits of ribosomal RNAs. Besides the coding regions, the compact mitochondrial genome has a non-coding region called the control region or 'D-loop' region. A large part of the control region is hypervariable with a high mutation rate, which makes it useful for studying phylogenetic relationships below species level (Larizza et al. 2002).

The first advantage of using mtDNA for phylogeny reconstruction is its inheritance pattern. It helps to monitor the transmission of the molecule along a single line (San Mauro et al. 2006). Another advantage of maternal inheritance of mtDNA is lack of recombination in the mitochondrial genome (Gillham 1994, Rokas et al. 2003). As a result, the mtDNA molecule can maintain its integrity through generations and that makes it a great choice for phylogenetic and population genetic studies (Birky 2001). Finally, the high mutation rate of animal mitochondria makes it an ideal choice for constructing phylogeny for closely related species. As mentioned before, the hypervariable region of the control region can be used to reveal conspecific variation (Larizza et al. 2002). Recent studies have revealed that entire mtDNA sequences can not only provide high resolution for reconstructing a robust phylogeny (Ingman et al. 2000, Miya et al. 2001, Delisle and Strobeck 2005, Yu et al. 2007, Krause et al. 2008, Zhang et al. 2008, Morin et al. 2010) but also can help in determining accurate dates of divergence events within a phylogeny (Schrago and Russo 2003, Yu et al. 2007, Rohland et al. 2007, Matsui et al. 2009, Krause et al. 2008, Zhang et al. 2008, Morin et al. 2010). Entire mtDNA sequences have been used successfully for reconstruction of high-resolution phylogeny with divergence dates in primates including prosimians (Matsui et al. 2009),

gibbons (Chan et al. 2010), orangutans (Xu and Arnason 1996), chimpanzees (Bjork et al. 2011), and humans (Ingman et al. 2000).

However, there are some problems associated with mtDNA-based phylogenies. In recent past some authors have argued that the evolution of mtDNA is not neutral (Ballard and Whitlock 2004, Hurst and Jiggins 2005). There are many instances that show the effect of direct and indirect selection on mtDNA, making the use of mtDNA as a marker for genomic history unreliable (Hurst and Jiggins 2005). Another major problem for mtDNA based phylogeny construction, especially for apes, is the presence of nuclear DNA segments in mtDNA ('numts') due to translocation events between nuclear and mitochondrial DNAs (Thalmann et al. 2004). Gorillas have been found to have several 'numts' integrated in their mitochondrial genome (Garner and Ryder 1996, Jensen-Seaman et al. 2004, Anthony et al. 2007). Due to the presence of 'numts' in the ape genomes, the phylogeny based on short mtDNA sequences can be unreliable. The only solution to avoid this problem is to use long-range amplification of large DNA fragments (Thalmann et al. 2004) and sequence the entire mtDNA instead of small mtDNA fragments.

2.2 Methods

This project is a part of a larger collaborative project. Several people contributed in this project. The entire mtDNA of a western lowland gorilla ("Chipua") was sequenced by Scott Hergenrother in Seaman Lab. Dr. Michael Jensen-Seaman helped in writing the manuscript and the manuscript was edited by Dr. N. Anthony and Dr. I. Soto-Caldron. I sequenced an eastern lowland gorilla ("M'kubwa"), analyzed the data (including tree building and dating), wrote the manuscript, and submitted the sequences to the GenBank.

2.2.1 DNA sequencing

The entire mtDNA was sequenced from a wild-born male eastern lowland gorilla ("M'kubwa"; *G. beringei graueri*). The tissue sample was obtained from the Coriell Cell Repositories, Hamden, NJ (ID PR00206). According to the International Gorilla studbook, M'kubwa was captured from the eastern Democratic Republic of the Congo in 1953, and died in the Houston Zoo in 2004 (Studbook ID 9907, Niekisch 2011). The mitochondrial genome was PCR-amplified in three overlapping fragments of 7.1, 7.5, and 5.5kb by my colleagues using long range Taq Polymerase. The amplified PCR products were purified using Wizard SV columns (Promega, Madison, WI). The purified products were sequenced with multiple primers on both strands. Sequencing was repeated several times until the entire mtDNA is covered from both directions. DNA sequencing was carried out using the BigDye v3.1 sequencing kit on an ABI3100 Avant and an ABI3130 automated capillary sequencer (Applied Biosystems). For sequence editing, creating contigs, and generating consensus sequences SeqMan program of the LaserGene package (DNA-Star, Madison, WI) was used.

2.2.2 Primate mtDNA Sequence Alignments

The Eastern gorilla (M'kubwa) sequence was aligned with the entire mtDNA sequences from Chipua, one additional Western gorilla (Genbank accession number NC_011120), four chimpanzees (*Pan troglodytes troglodytes*, HM068587; *Pan troglodytes schweinfurthii*, HM068591; *Pan troglodytes verus*, HM068593; *Pan troglodytes ellioti,* HM068585), two bonobos (GU189676 and GU189674), human (J01415.2), Neanderthal (FM865411), two orangutans (*Pongo abelii*, NC_002083 and *Pongo pygmaeus*, NC_001646), gibbon (*Hylobates lar*, NC_002082), and macaque

(*Macaca mulatta*, NC_005943). Thus a total of 16 taxa were used for the phylogenetic analysis. For the ND5 only study, the sequence of a mountain gorilla (AF240447) was used.

Additionally, a D-loop phylogeny was generated only for gorillas. For this purpose twelve previously published sequences of the first hypervariable region (HV1) of the D-loop were used in addition to the two gorillas sequenced in our lab (M'kubwa and Chipua). The sequences came from two mountain gorillas (AY530103, AF089820), two eastern lowland gorillas (AF050738, AF187549), and eight western lowland gorillas (AY530138, AY530141, AY530128, AY530132, AY530118, AY530112, AY530119, AY530120).

The 15 complete mtDNA genomes were aligned with the ClustalW online server (Larkin et al. 2007). Two datasets were used for the analysis. One dataset contained 12 guanine-rich protein-coding genes located on the heavy strand of the mtDNA (10887 nucleotides), following the approach discussed in Raaum *et al.* (2005). The other dataset contained the entire mitochondrial genomes except the D-loop (15599 nucleotides). For the ND5 only analysis, a data-subset of 1812 nucleotides was used. Finally for the D-loop dataset 261 nucleotides of HV1 of D-loop was used.

2.2.3 Protein coding genes, tRNA and rRNA analysis

The mtDNA from one human (J01415.2), one chimpanzee (HM068587), one bonobo (GU189674), and the two novel gorilla mtDNAs from Chipua and M'kubwa were divided into individual protein coding genes, tRNA, and rRNA genes. The protein coding, tRNA and rRNA genes and translated amino acid sequences were aligned using ClustalW online server (Larkin et al. 2007).

The ClustalW file containing the alignment of 12 protein-coding gene sequences from chimpanzee, bonobo, M'kubwa, and Chipua were converted into phylip (.phy) format using ALTER web-based server. PHYLIP v3.695 package (Felsenstein 1993) was then used for further analysis. 1000 bootstrap pseudoreplicates of the alignment were generated using the SEQBOOT program, implemented in PHYLIP v3.695 package. Then pairwise genetic distances were calculated for all 1000 psudoreplicates using DNADIST program. Using DNADIST output file, 1000 genetic distances were generated for chimpanzee-bonobo and Eastern-Western gorilla using UNIX command lines. The genetic distances were compared to find out how many genetic distance pairs show chimpanzee-bonobo distances greater than Eastern-Western gorilla and vice-versa. All genetic distances from both pairs were plotted as a scatterplot using GraphPad Prism v6 (GraphPad Software).

The same .phy file ("12-gene" 4 species alignment file) was re-analyzed using the 'ape' package (Paradis et al. 2004) implemented in R v3.0.2. The genetic distances and variances between both chimpanzee-bonobo and Eastern-Western gorilla sequences were first calculated with 'dist.dna' program using F81 nucleotide substitution model, and then repeated with TN93 model.

Pairwise K_a/K_s values between chimpanzee and bonobo, and Eastern and Western gorilla were calculated using 'seqinr' package (Charif and Lobry 2007) implemented in R v3.0.2. All R codes used in this section are shown in Appendix 1.5.

The pairwise genetic distances for all 13 protein coding genes separately, t-RNAs, r-RNAs, D-loops, transition and transversion rates, and rate heterogeneity parameters were estimated using MEGA v5.2.2 (Tamura et al. 2011).

Additionally, the number of transitions and transversions between Chipua mtDNA sequence to another Western gorilla (NC_001645), M'kubwa, chimpanzee (HM068587), human (J01415.2), orangutans (NC_002083), gibbon (*Hylobates lar*, NC_002082), and macaque (*Macaca mulatta*, NC_005943) mtDNA sequences were calculated using Kimura-2-parameter model in MEGA v5.2.2.

2.3.4 Phylogenetic analysis

2.3.4.1 Bayesian approach

To infer Bayesian phylogeny, I used Bayesian Markov Chain Monte Carlo approach (MCMC) implemented in BEAST v1.7.5 (Drummond and Rambaut 2007) and MrBayes v3.2.2 (Huelsenbeck and Ronquist 2001). The BEAST input file was generated using BEAUTi v1.7.5 (Drummond and Rambaut 2007). The BEAST files used for all analyses are shown in Appendix 1.1. Uncorrelated lognormal relaxed molecular clock was used to allow evolutionary rate to vary from branch to branch. This approach has been previously (Drummond et al. 2006) shown to provide a better estimate of time to most recent common ancestor (tMRCA) over the strict molecular clock that does not allow the evolutionary rate to vary among branches. SRD06 model of nucleotide substitution was used to partition the nucleotide data by codon position, so that the third codon position can differ from position 1 and 2. This model has been successfully used previously for reconstruction of Bayesian phylogeny of chimpanzee (Bjork et al. 2011). The rate of evolution was calibrated using lognormally distributed priors as described in Raaum et al. (2005) with lognormal means of zero and lognormal standard deviation of 0.56. As discussed by several authors (Ho 2007, Bjork et al. 2011), lognormally distributed priors work better than both normally distributed and exponentially distributed

priors, when using fossil calibration for dating the tree. A lognormal curve with mean of zero and standard deviation of 0.56 will be skewed to the right with a longer tail. Due to this kind of shape it will sample values from the more distant past more frequently than time representing nearer past. Unlike previous studies (Raaum et al. 2005 and Bjork et al. 2011), I offset these distributions only at the robustly fossil-supported internal node of human-chimpanzee split by 5 Ma. The same offset point for human and chimpanzee split was used in Bjork et al. (2011) to ensure that the median values of the distribution equals the expected 6 Ma split. Same offset point was used for the ND5 only phylogeny construction.

MCMC simulation ran for 10 million generations for both datasets. After running, 10% of the generations from each run were discarded as 'burnin'. The maximum clade credibility (MCC) tree was identified and annotated using TreeAnnotator v1.7.5 (Drummond and Rambaut 2007). Nodes with posterior probabilities exceeding 90% ($P > 0.9$) were used for tree building. The MCC tree generated by TreeAnnotator v1.7.5 was visualized and dated using FigTree v. 1.4 (Rambaut 2014). Tracer v1.5 (Rambaut and Drummond 2009) was used for summarizing the posterior estimates of the various parameters sampled by the Markov Chain. tMRCA means, medians, 95% highest posterior density (95% HPD) intervals (all in Ma) and effective sample sizes (ESS) were calculated using Tracer.

For MrBayes MCMC study, the ClustalW alignment of both 12-gene dataset and without D-loop dataset was converted into nexus format. The nucleotide substitution model was set to GTR model, which was suggested to be the best model by jModelTest v2.1.4 (Guindon et al 2003, Darriba et al 2012). jModelTest out put is shown in

Appendix 1.3. The analysis was run for 1 million generations with a sampling frequency of 10 to get at least 1,000 samples from the posterior probability distribution. The parameter values and trees were summarized using 'sump burnin' and 'sumt burnin' commands respectively, taking 25% of the samples from the posterior probability distribution in both cases. The tree output was visualized using AWTY (Wilgenbusch 2014).

2.3.4.2 Maximum Likelihood approach

The ClustalW alignments of both 12-gene and without D-loop datasets were used for phylogeny construction by Maximum likelihood approach. A Model test was performed using jModelTest v2.1.4 (Guindon et al 2003, Darriba et al 2012) to determine the best fitting nucleotide substitution model for the datasets. Akaike's information criteria with correction (AICc) and Bayesian information criteria (BIC) values were used to determine the best model, considering the smaller the values of AICc and BIC, the better the model is. MEGA v5.2.2 (Tamura et al 2011) was used to construct the maximum likelihood tree. 1000 bootstrap resampling were performed for both trees. Additionally, Parsimony and a distance tree were also drawn with both datasets with 1000 bootstrap resamplings, using MEGAv5.2.2. For maximum parsimony tree Subtree-Pruning-Regrafting (SPR) search method was employed. For the distance tree Neighbor-joining method was employed with Kimura-2-parameter model.

2.3 Results

2.3.1 DNA sequencing

Complete mitochondrial genome sequences were obtained for the eastern lowland

gorilla ("M'kubwa", *Gorilla beringei graueri*) and deposited in GenBank (accession number KF914213). In the M'kubwa mtDNA sequence, in a poly-C stretch of the second hyper-variable region (HV2) (16,219 – 16,230), the exact number of cytosines could not be determined, presumably due to polymerase stutter. Following previous publications (Garner and Ryder 1996; Thallmann et al. 2005), the sequence was therefore submitted as …$AAC_{12}ACT$… in this region and annotated as undetermined. All sequences appeared to be derived from authentic mitochondrial DNA, not numts. No apparently heterozygous sites were observed. Further more, there were no premature stop codons or indels seen within protein-coding regions of the mtDNA.

2.3.2 Dating species splits using mtDNA

jModelTest suggested TIM2+I+G (with AICc and BIC scores of 93,434 and 93,711 respectively) to be the best nucleotide substitution model. Since this model is not implemented in MEGA v5.2.2 and MrBayes v3.2.2, the second best model, GTR+I+ G (with AICc and BIC scores of 93,438 and 93,729 respectively) was used for phylogenetic analysis. As mentioned before, two datasets were created for the analyses. The "12-gene" dataset containing the concatenated sequences of the 12 protein-coding genes located on the heavy strand (10,887 nucleotides), following Raaum *et al.* (2005). The other, referred to as the "complete mtDNA" dataset, contains the entire mitochondrial genome sequence except the D-loop (15,599 nucleotides). Both the 12-genes (Fig. 2.2) and the complete mtDNA sequence (Fig. 2.3) alignments produced identical tree topologies, with the combined Bayesian posterior probability of 1.000. MrBayes generated identical tree topology (Fig. 2.4).

The maximum likelihood tree too had identical topology, with all nodes supported in 99% or more of bootstrap replicates (Fig. 2.5). Maximum parsimony (Fig. 2.6) and Neighbor-joining trees (Fig. 2.7) are also shown. All trees were drawn using macaque as the outgroup, but for aesthetic reason not shown on the trees.

Phylogenetic analysis of the novel *HV1* sequences along with all previously published gorilla *HV1* sequences confirms that wild-born M'kubwa is an eastern lowland gorilla belonging to haplogroup B and that the mtDNA of Chipua belongs to the D3 haplogroup (Fig. 2.9), following the *HV1* haplotype nomenclature of Anthony et al. (2007).

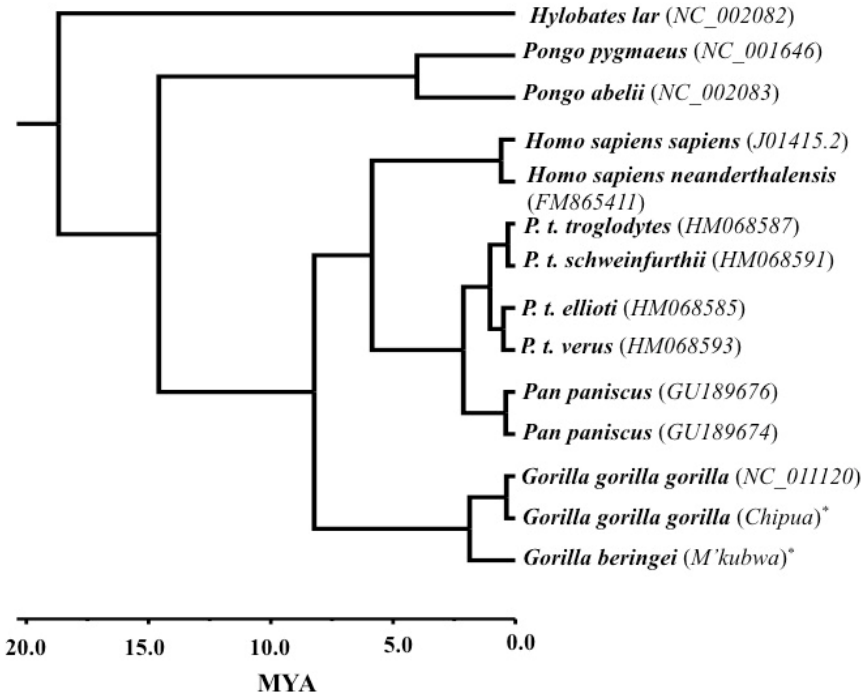


Figure 2.2: "12-gene" based maximum credibility tree generated by BEAST v1.7.5

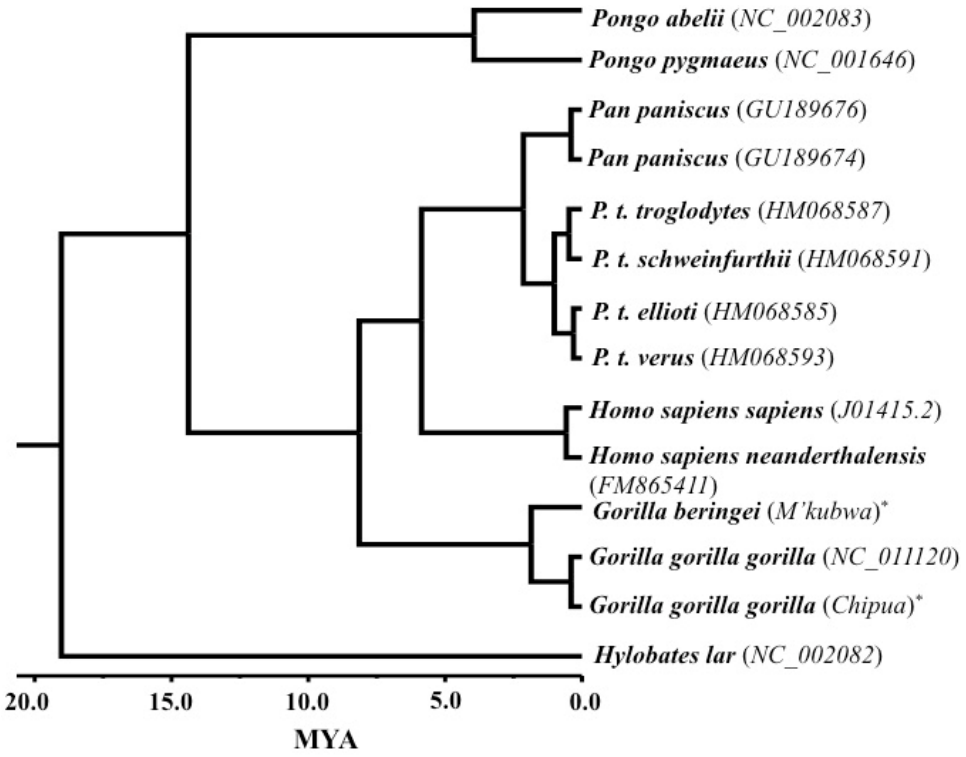


Figure 2.3: "complete mtDNA" based maximum credibility tree by BEAST v1.7.5

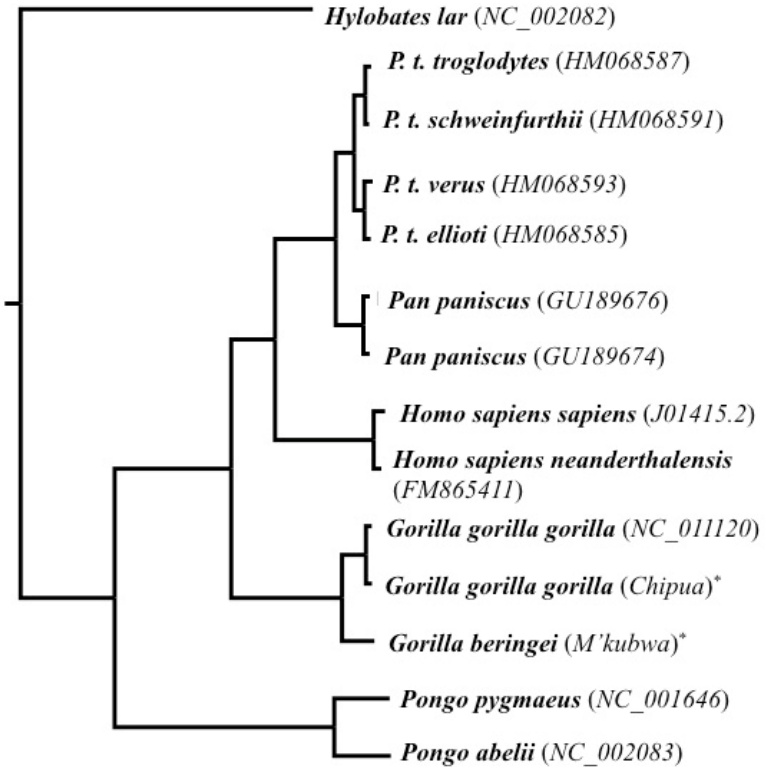


Figure 2.4: "12-gene" based hominoid phylogeny generated by MrBayes v3.2.2

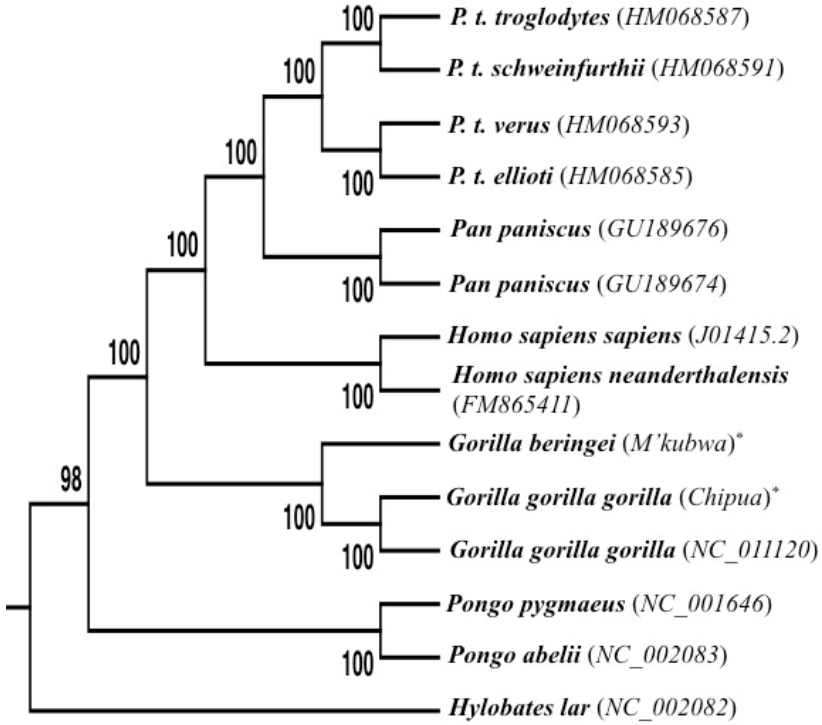


Figure 2.5: "12-gene" based Maximum Likelihood tree generated by MEGA v5.2.2

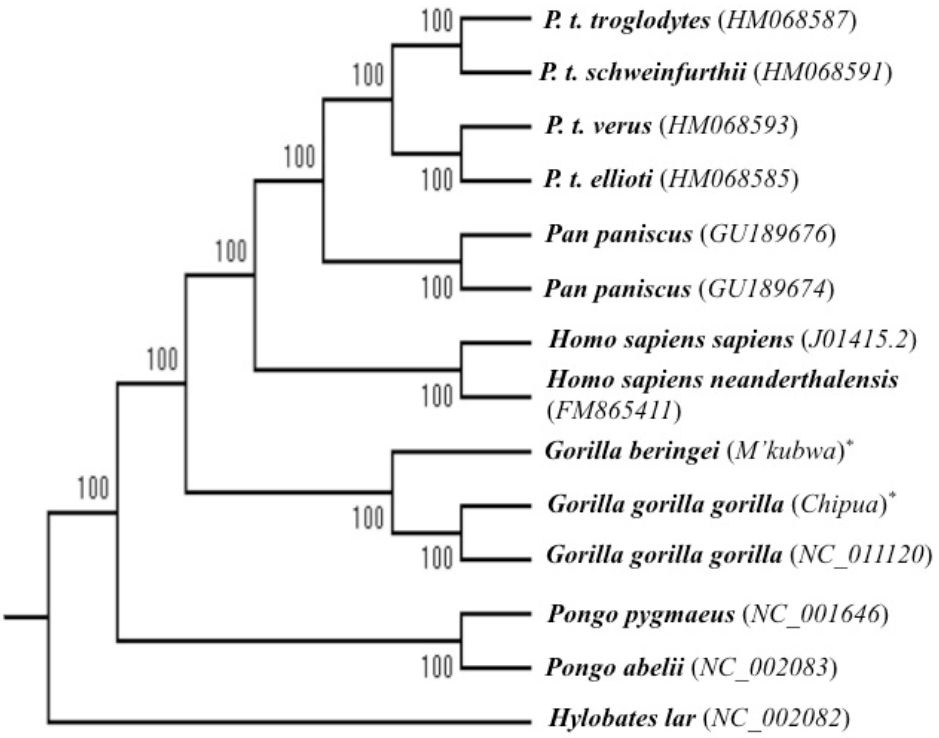


Figure 2.6: "12-gene" based Maximum Parsimony tree generated by MEGA v5.2.2

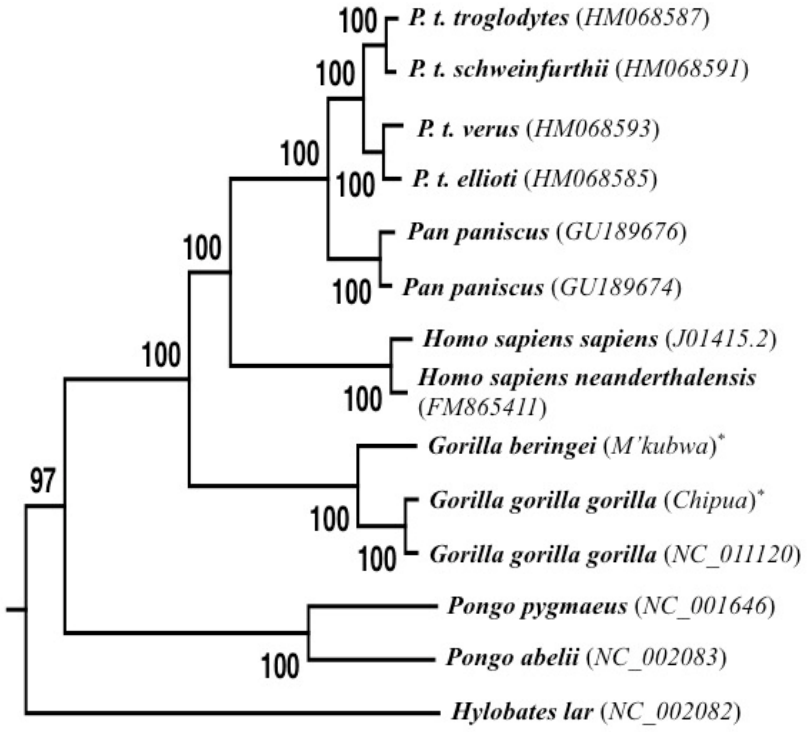


Figure 2.7: "12-gene" based Neighbor-joining tree generated by MEGA v5.2.2

Using a Bayesian MCMC approach the divergence between the gorilla species occurred nearly two million years ago at 1.926 (1.487 – 2.416) Mya, slightly more recently than the chimpanzee-bonobo split at 2.173 (1.729 – 2.691) Mya, calibrated with a human-chimpanzee divergence of 6 Mya (Table 2.1).

Table 2.1: Dates with confidence intervals

	12 heavy strand genes		**Whole mtDNA minus D-loop**	
Taxon divergence	**tMRCA[a]**	**95% HPD[b]**	**tMRCA[a]**	**95% HPD[b]**
Old World monkey - Hominoid	32.215	25.288 – 40.411	32.535	25.339 - 40.758
Gibbon-Hominid	19.068	15.423 - 23.432	19.280	15.314 - 23.507
Pongo-African Apes	14.853	12.053 - 18.163	14.537	11.657 - 17.763
Pongo pygmaeus-P. abelii	4.090	3.192 – 5.213	3.989	3.015 - 5.089
Gorilla-*Homo*/*Pan*	8.396	7.063- 10.192	8.280	6.919 – 10.003
Homo-Pan	5.983	5.200 - 7.058	5.982	5.197 - 7.082
Pan troglodytes-P. paniscus	2.163	1.742 - 2.691	2.172	1.715 - 2.679
P.t.troglodytes/P.t. schweinfurthii-P.t.verus/P.t.ellioti	1.054	0.824 - 1.330	1.027	0.803 - 1.304
Gorilla gorilla-G. beringei	1.900	1.456 - 2.397	1.895	1.438 - 2.391
Deepest root within Western gorilla	0.370	0.258 - 0.494	0.404	0.284 - 0.531

Human-Neanderthal	0.591	0.428 - 0.762	0.587	0.430 - 0.758

[a]Time to most recent common ancestor, in Myr.
[b]95% highest posterior density.

The deepest split within western gorillas is between our novel Chipua sequence and previously published genomes, occurring relatively recently, less than 400,000 years ago. The subspecies level diversification within chimpanzee of 1.055 (0.831-1.332) Mya is much newer than Eastern-Western gorilla divergence time. Interestingly, the deepest root within Western gorilla marginally overlaps with human-Neanderthal split time of 0.590 (0.429 – 0.758) Mya. Estimated divergence dates from the complete mtDNA genome (excluding the D-loop) are very similar to the 12-gene data set (Table 2.1).

I in turn used the estimated date of 6 Mya for human and chimpanzee divergence as a calibration point and a previously published mountain gorilla *NADH5* sequence to estimate the time of the eastern lowland (*G. b. graueri*) and mountain (*G. b. beringei*) gorilla divergence at 0.378 (0.04-0.864) Mya, slightly more recent than the deepest split within *G. gorilla* (Figure 2.8).

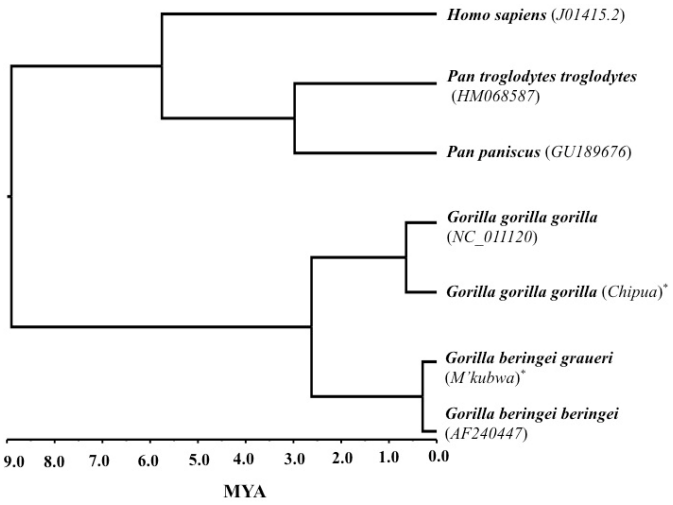


Figure 2.8: ND5 based maximum credibility tree generated by BEAST v1.7.5

The dates obtained from HV1 were substantially higher than *NADH5*; when calibrated with an eastern-western gorilla divergence of 2 Mya, the deepest split within western gorillas is 1.775 Mya (1.176-2.599 Mya; median 1.685 Mya) and the *G. graueri* vs. *G. beringei* split time is 0.929 Mya (0.764-2.028; median 0.803 Mya).

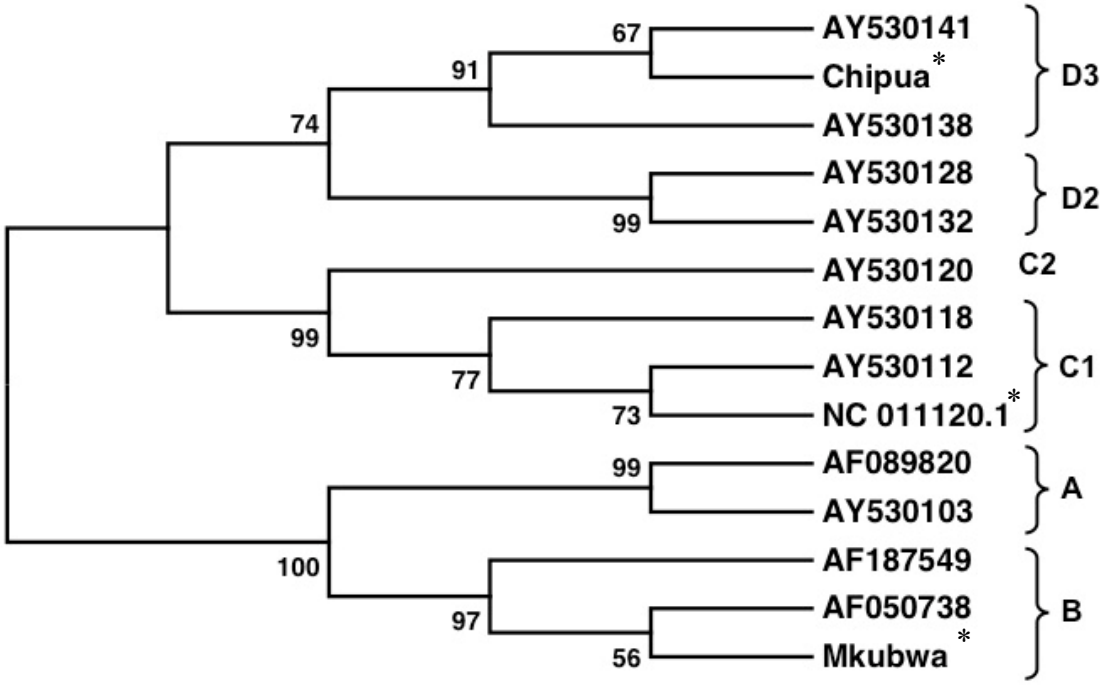


Figure 2.9: Consensus maximum likelihood tree (GTR+I+G model) of gorilla *HV1* sequences, showing the placement of complete genomes (indicated with an asterisk) with respect to previously described haplogroups. Percentage of bootstrapped replicates supporting each node (out of 1,000 bootstraps) are shown.

2.3.3 Additional analysis of Eastern-Western gorilla split time based on Great Ape Genome Project Data

Since the initiation of this project, partial mitochondrial genome sequences became available from 31 additional gorillas, as part of Great Ape Genome Project (GAGP) http: //biologiaevolutiva.org/greatape/ (Prado-Martinez et al. 2013). I modified

my complete mtDNA data by adding nine additional gorillas (including one additional Eastern gorilla) and removing the Western gorilla (NC_011120) used in previous calculations.

I first randomly picked eight additional gorillas from GAGP database and reconstructed the phylogeny with one chimpanzee (*Pan troglodytes troglodytes*) and one human using BEAST v1.7.5. The phylogeny came up with a more ancient split date for Eastern and Western gorilla of 2.291 Mya (1.772-2.971, median=2.229). The deepest split among Western gorillas was ~510,000 years ago.

Then I calculated genetic distances among all gorillas published in GAGP database using maximum composite likelihood model in MEGA v5.2.2. Then I picked nine gorillas from GAGP database showing higher genetic distance among each other. Using these nine additional gorillas I reconstructed the phylogeny with the same chimpanzee (*Pan troglodytes troglodytes*) and human using BEAST v1.7.5. The phylogeny came up with even more ancient split date for Eastern and Western gorilla of 2.507 Mya with a broader confidence interval (1.762-3.861, median=2.293). The deepest split among Western gorillas was ~530,000 years ago.

Then I added the three other chimpanzee subspecies (*P. t. verus, P. t. ellioti and P. t. schweinfurthii*), and Neanderthal to the analysis and reconstructed the phylogeny using BEAST v1.7.5. This time the split date for Eastern and Western gorilla was almost similar to the last analysis (2.494 Mya) but the confidence interval was braoder (1.779-4.700; median= 2.264). The deepest split among Western gorillas remained at ~530,000 years ago.

Finally, I repeated the analysis with all primates that were used in section 2.3.2.

But the Western gorilla (NC_011120) was replaced with the nine Western gorillas that were used in the previous analysis. The phylogeny was reconstructed in BEAST v1.7.5. The results were identical to section 2.3.2 (Table 2.2) with identical tree topology (Fig. 2.10). The split date for Eastern and Western gorilla was 1.94 Mya and the confidence interval became narrower and the median went down too (1.508-2.462; median = 1.909). The deepest split among Western gorillas also came up to be younger at ~420,000 years ago. All trees and BEAST files are added in Appendix 1.

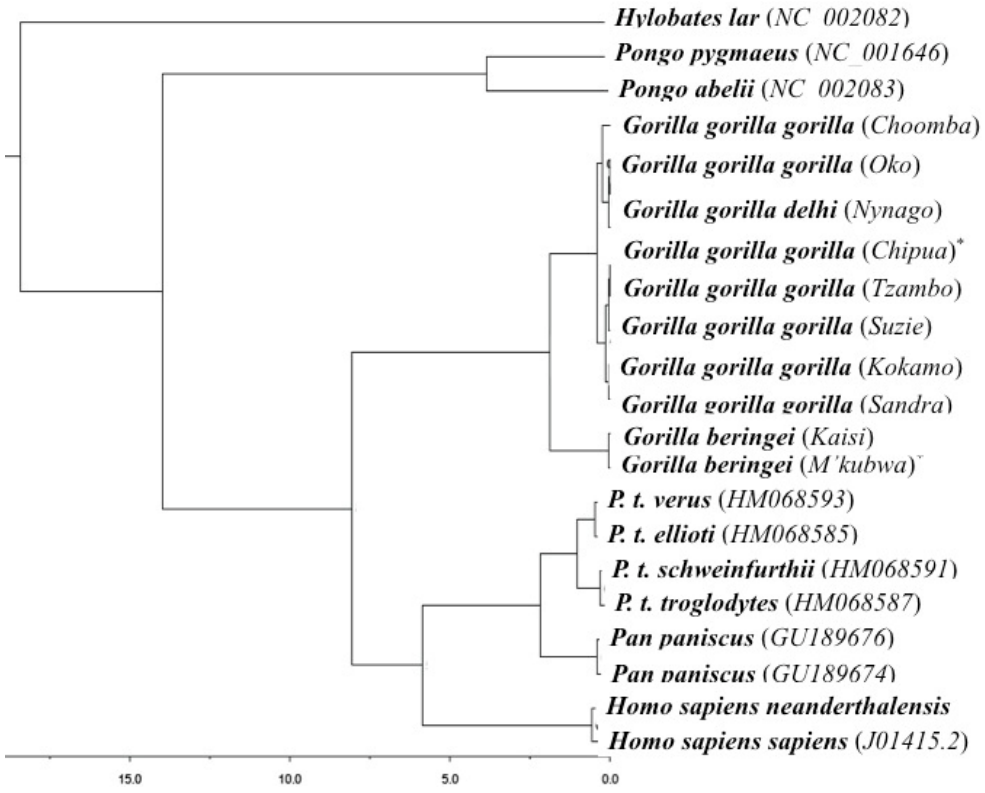


Figure 2.10: "complete mtDNA" based maximum credibility tree with GAGP gorillas by BEAST v1.7.5

Table 2.2: Comparison of Split dates before and after the addition of GAGP gorillas

	Whole mtDNA minus D-loop (with GAGP gorillas)		**Whole mtDNA minus D-loop (without GAGP gorillas)**	
Taxon divergence	**tMRCA**	**95% HPD**	**tMRCA**	**95% HPD**
Old World monkey - Hominoid	31.268	23.940 – 39.165	32.535	25.339 – 40.758
Gibbon-Hominid	18.863	15.109 – 23.030	19.280	15.314 – 23.507
Pongo-African Apes	14.332	11.495 – 17.590	14.537	11.657 – 17.763
Pongo pygmaeus-P. abelii	3.962	3.019 – 4.987	3.989	3.015 – 5.089
Gorilla-*Homo*/*Pan*	8.269	6.876 – 9.946	8.280	6.919 – 10.003
Homo-Pan	5.997	5.197 – 7.082	5.982	5.197 – 7.082
Pan troglodytes-P. paniscus	2.189	1.744 – 2.714	2.172	1.715 – 2.679
P.t.troglodytes/P.t. schweinfurthii-P.t.verus/P.t.ellioti	1.035	0.801 – 1.293	1.027	0.803 – 1.304
Gorilla gorilla-G. beringei	1.940	1.508 – 2.462	1.895	1.438 – 2.391
Deepest root within Western gorilla	0.425	0.319 – 0.543	0.404	0.284 – 0.531
Human-Neanderthal	0.596	0.439 – 0.775	0.587	0.430 – 0.758

2.3.4 Chimpanzee-Bonobo and Eastern-Western gorilla comparison

In the gene-by-gene comparison using the 13 protein-coding genes, eight of the *Pan* species splits are older than the *Gorilla* species splits (Table 2.3). The tRNA and rRNA divergence is nearly identical for both groups. Interestingly, the D-loop also, showed nearly identical genetic divergence.

Table 2.3: Chimpanzee/bonobo vs. Eastern/Western Gorilla genetic distance

Locus[a]	Chimp-Bonobo distance[b]	Western-Eastern gorilla distance[b]	Chimp-Bonobo Ka/Ks	Western-Eastern Ka/Ks
ATPase6	0.040	0.054	0.317	0.321
ATPase8	0.018	0.015	0.300	0.408
COI	0.031	0.021	0.078	0.075
COII	0.028	0.033	0.000	0.116
COIII	0.042	0.031	0.118	0.170
Cytb	0.050	0.050	0.206	0.060
NADH1	0.051	0.040	0.091	0.144
NADH2	0.049	0.038	0.151	0.189
NADH3	0.027	0.036	0.143	0.313
NADH4	0.042	0.037	0.113	0.211
NADH4L	0.045	0.033	0.052	0.035
NADH5	0.062	0.058	0.253	0.173
NADH6	0.049	0.047	0.106	0.000
tRNAs	0.018	0.016	-	-
12S rRNA	0.014	0.008	-	-
16S rRNA	0.024	0.025	-	-
D-Loop	0.105	0.103	-	-

[a]Genetic distance results were calculated separately for each gene between Chipua and M'kubwa, with the exception of the 22 tRNAs, which were concatenated into one sequence for analysis.
[b]Genetic distance calculated with MEGA v5.2.2 using maximum composite likelihood model

All genes in the mitochondrial genome showed similar ($P = 0.35$, Wilcoxon Rank test) genetic divergence between chimpanzee and bonobo, and Eastern and Western gorilla (Table 3). The overall genetic distance, using "12 genes" dataset, between chimpanzee-bonobo pair was 0.046; the same for Eastern-Western gorilla was 0.042 using Maximum Composite Likelihood Model in MEGA V5.2.2. Similar results were obtained using the 'dist.dna' function in R. According to TN93 model chimpanzee-bonobo distance was 0.043 and Eastern-Western distance was 0.040.while F81 model suggested chimpanzee-bonobo distance to be 0.042 and Eastern-Western distance to be 0.039.

Out of 1000 bootstrapped pairwise genetic distances created for the *Pan* and for the *Gorilla* species, in 971 cases chimpanzee-bonobo genetic distances were greater than Western-Eastern gorilla genetic distances. For the rest 29 cases Western-Eastern gorilla genetic distances were greater than chimpanzee-bonobo distances. There was no identical value for any given pair (Fig.2.11).

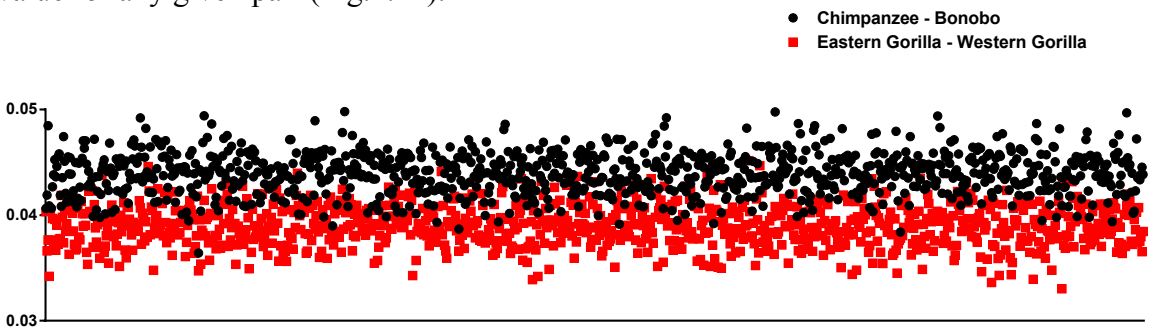


Figure 2.11: The mitochondrial genetic distance between *Gorilla* species (Chipua and M'kubwa) compared to that of the *Pan* species, based on 1,000 bootstrapped replicates of the 12-gene dataset

Similar 1,000 bootstrapped replicates of pairwise sequence alignments created for the *Pan* subspecies and for the *Gorilla* species. There was no overlap between the two (Fig. 2.12). Out of 1000 bootstrapped pairwise genetic distances, in all 1000 cases *Pan* subspecies genetic distances were lower than Western-Eastern gorilla genetic distances. This further suggests Western-Eastern gorilla genetic distances to be greater than 'subspecies' level.

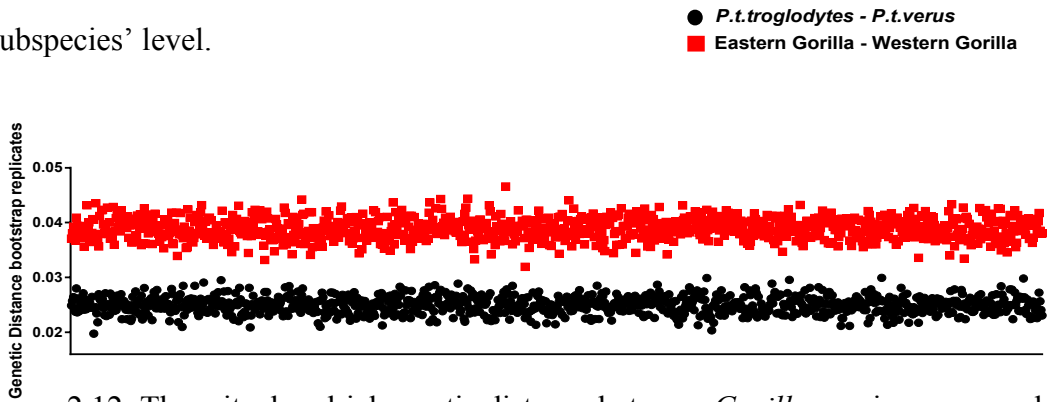


Figure 2.12: The mitochondrial genetic distance between *Gorilla* species compared to that of the *Pan* sub-species, based on 1,000 bootstrapped replicates of the twelve-gene dataset

2.3.5 Protein coding genes, tRNA and rRNA analysis

Of the 22 tRNA genes, 9 are identical between *G. gorilla* (Chipua) and *G. graueri* (M'kubwa), with 21 substitutions found among the remaining 13 tRNA genes. Individual t-RNA comparison between the *Pan* species and the *Gorilla* species is shown in Table 2.4.

Table 2.4: t-RNA comparison between the *Pan* species and the *Gorilla* species

tRNAs	**Chimpanzee-Bonobo nucleotide differences**	**Eastern-Western gorilla nucleotide differences**
Phe	2	Identical
Val	Identical	Identical
Leu	2	Identical
Ile	Identical	Identical
Gln	Identical	Identical
Met	Identical	1
Trp	2	1
Ala	1	Identical
Asn	1	1
Cys	1	Identical
Tyr	Identical	1
Ser	Identical	2
Asp	1	1
Lys	Identical	4
Gly	1	1
Arg	1	1
His	Identical	Identical

Ser2	2	4
Leu2	Identical	Identical
Glu	Identical	1
Thr	3	2
Pro	1	1
Total	**18**	**21**

There are a total of 41 nucleotide differences in the rRNAs between chimpanzee and bonobo, and 39 differences between Eastern and Western gorilla. Among 41 nucleotide differences between chimpanzee and bonobo, 12 were found in 12S rRNA and 29 were found in 16S rRNA. Eastern and western gorilla has 9 and 30 differences between each other in 12S and 16S rRNA respectively.

12 out of the 13 protein-coding genes (*ND1, ND2, ND3, ND4, ND5, COII, COIII, ATP6, ATP8, CYTB, COI,* and *ND4L*) differs by at least one predicted amino acid between our eastern lowland gorilla and our western lowland gorilla for the gene-by-gene comparisons, although only ten of these genes contain an apparent fixed difference, where all three western gorilla mitochondrial genomes code for a different amino acid than the single eastern lowland gorilla genome. With only one complete *G. beringei* sequence it is not possible to determine if such differences are polymorphic within this species. All proteins appear to be evolving under purifying selection in general, in that all gene-wide *Ka*/*Ks* values are less than one (Table 3). *ND6* is not fixed among the three western gorillas and also does not differ by any amino acid substitution between the eastern and western gorilla.

2.3.6 Rate of evolution, transitions and transversions

2.3.6.1 Transitions and transversions

The transition and transversion rates between Chipua and other taxa are shown in Table 2.5. The transition and transversion rates were then plotted against the time of divergence between the two taxa (Fig.2.13).

Table 2.5: Transition and Transversion rates between Chipua and other taxa

Taxa compared	Transition	Transversion
Western Gorilla	0.007	0.001
M'kubwa	0.034	0.006
Chimpanzee	0.119	0.021
Human	0.126	0.023
Orangutan	0.197	0.035
Gibbon	0.218	0.039
Macaque	0.334	0.06

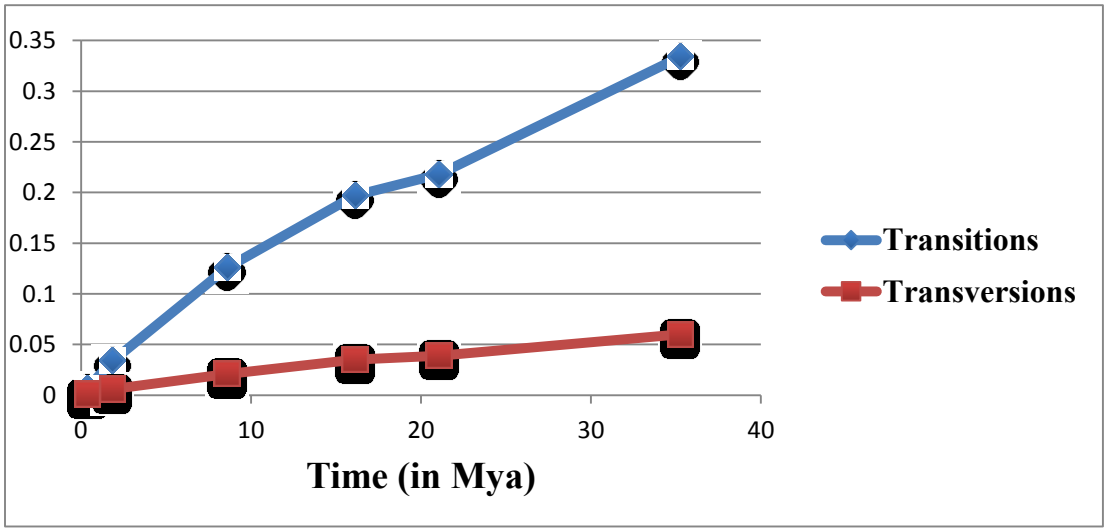


Figure 2.13: The transition and transversion rates across time in Chipua. Transversion rate, expectedly, is much lover than transition rate. Transition has not reached saturation in the given time scale

2.3.6.2 Evolutionary rate of different parts of mitochondrial genome

Consider the rate of substitution 'r'. The substitution rate varies from site to site according to a gamma distribution (Uzzell and Corbin 1971). Gamma distribution can be expressed in terms of r in the following way:

$$f(r) = \frac{(\text{mean of r/variance of r})^{a} * e^{-(\text{mean of r/variance of r})r} * r^{a-1}}{\Gamma(a)}$$

Where Γ(a) is the gamma function dependent on the shape parameter a. According to the above equation the substitution rate varies from site to site according to the shape parameter a. When a <1, the distribution of r becomes skewed and many sites have an r value close to zero and become invariable. So, the lower the value of 'a' than 1, the closer the value of r to zero (and lower the evolutionary rate).

The maximum likelihood estimates of the shape parameter 'a' were calculated in MEGA v5.2.2 for different parts of the mitochondrial genome. D-loop showed the highest rate of evolution with the estimated value of the shape parameter (a) at 0.3304. Protein coding genes came next with shape parameter (a) value of 0.1848, followed by rRNAs with 'a' value of 0.1360. The rate of evolution was the slowest for tRNAs with 'a' value of 0.0500 with virtually most sites have r = 0.

2.4 Discussion

M'kubwa, the Eastern gorilla, is the last living hominid species to have its mitochondrial genome completely sequenced and thus this study provides an important piece of information, as far as primate genome sequencing is concerned. With the exception of a stretch of cytosines in the HV2 region of D-loop as mentioned in section 2.3.1, all bases of M'kubwa mtDNA were called with complete confidence.

As mentioned in section 2.1.3, the presence of 'numt's in hominoid mtDNA causes a big problem while reconstructing hominoid mtDNA based phylogenies. The only solution to this problem is to compare longer mtDNA sequences and if possible the entire mtDNA. Since Eastern gorilla complete mtDNA sequence was unavailable before this study, avoiding 'numts' while constructing gorilla phylogeny was nearly impossible (Jensen-Seaman et al. 2004, Anthony et al. 2007, Thalmann et al. 2004). In this study the use of large amplicons successfully avoided inadvertent PCR amplification of numts. As a result I could successfully reconstruct a potential 'numt'-free phylogeny.

Our estimate of split times between the *Gorilla* species is by far the most accurate estimate of gorilla divergence times, as it is the first to be based on complete genome sequence. These estimates are sensitive to fossil calibration and the assumptions on nature of nucleotide substitutions. So, problem in either or both, like calibration from an imperfect fossil record and/or inaccurate assumptions regarding the nature of the nucleotide substitutions can generate wrong estimates of divergence times.

The discrepancy regarding the divergence timings between mitochondrial and nuclear loci can be seen in hominids (Jensen-Seaman et al. 2001, Mailund et al. 2012). Our current study is no exception. The split time between the Gorilla species as shown by the mtDNA predates nuclear DNA based phylogeny. One probable explanation of this discrepancy is that long distance male-mediated gene flow persisted much longer than female gene flow as gorilla populations became isolated. This explanation is consistent with the sex-biased described dispersal patterns of gorillas. Female gorillas tend to emigrate from their natal group to quickly join a neighboring group, whereas male gorillas may spend years traveling long distances before establishing or taking over a

reproductive group (Inoue et al. 2013). So, our estimated split time between the *Gorilla* species of 1.9 Mya, is probably an estimate of the time of cessation of female-mediated gene flow between populations that through evolutionary course of time gave rise to the Eastern and Western gorilla species. The habitat fragmentation of the early to mid-Pleistocene, created islands of forest refugia in central Africa. This along with further reduction of the available paths of migration between West Africa and the eastern populations by the formation of the Congo/Ubangui River system (approximately 1.5 Mya) probably restricted the gene flow between gorilla populations.

When the phylogeny reconstruction was repeated with GAGP data, the split date between Eastern and Western gorilla varied with the number and type of taxa used to construct the phylogeny. The phylogeny reconstructed with all primates, Eastern-Western gorilla split date was almost identical as before at 1.89 Mya. When the similar phylogeny was reconstructed with only chimpanzees and humans the split time became more ancient at ~2.5 Mya. The confidence intervals also become broader. These results show the importance of taxon sampling during broad phylogeny reconstruction. When using fewer taxa, or using more taxa from one specific group, but fewer from other groups, the phylogeny suffers from power deficiency and becomes biased.

As mentioned in section 2.1.1, gorilla speciation was always controversial. Groves (Groves 1967, 1970, 2003) revised gorilla phylogeny and considered all gorillas to be a single species with three subspecies (*Gorilla gorilla gorilla, Gorilla gorilla beringei and Gorilla gorilla graueri*). However, mtDNA based phylogenies in mid-1990s, showed an older split time between Western and Eastern gorillas (Ruvolo et al. 1994, Morell 1994, Garner and Ryder 1996, Groves 2001, 2003). Although based on

short mitochondrial DNA sequence (mostly D-loop sequences), these studies triggered the rethinking process regarding gorilla phylogeny. The consensus of the field began to change with many authors recognizing two gorilla species, *G. gorilla* and *G. beringei*. Interestingly, the average sequence divergence between Western and Eastern gorillas was reported to be larger than that of between chimpanzee and bonobo, which are universally recognized as different species, at the *COII* gene and the *HV1* (Ruvolo et al. 1994, Garner and Ryder 1996). Our results show a clear advantage in utilizing the complete mtDNA sequence, which reveal that the chimpanzee-bonobo mitochondrial genomes are actually slightly more divergent overall than the eastern-western gorilla genomes. The *COII* gene is actually the only gene where the Western-Eastern gorilla divergence is greater than chimpanzee-bonobo divergence (Table 2.3). Although we now recognize that mtDNA does not provide a complete picture of the events surrounding genetic isolation of *G. gorilla* and *G. beringei*, it is unlikely that the consensus opinion will revert to a single species taxonomy. The current emerging picture is that the ancestral gorilla populations began to separate nearly two million years ago, based on our dating of the mitochondrial divergence at 1.9 Mya as well as the dating of the average nuclear sequence divergence between eastern and western gorilla genomes at 1.75 Mya (Scally et al. 2012), or somewhat more recent (0.9 - 1.6 Mya; Thalmann et al. 2007). Following this initial split, gene flow continued among these populations until about 100 kya, perhaps predominantly via male migration.

The 1000 bootstrap replicate scatterplot of genetic distances between the chimpanzee subspecies and the same between *Gorilla* species (Fig. 2.12) clearly shows the difference in genetic divergence between the two. However, the similar plot with *Pan*

species and *Gorilla* species shows ~70% overlap in the genetic distances (Fig. 2.11). These results further reiterate the fact that Western-Eastern gorilla divergence is more equivalent to sister species pairs in primates. In fact others have noted that the overall degree of anatomical and molecular differentiation between eastern and western gorillas is clearly greater than between any chimpanzee subspecies, and equivalent to other sister species pairs in primates (Groves 2001).

In this study, eastern and western lowland gorillas were defined as different species mainly based on their split time and genetic divergence, comparing with chimpanzee and bonobo. Since eastern and western gorillas live in two geographically isolated populations, it is unknown whether they can still mate in the wild and produce fertile offspring. So, we cannot use biological species concept when addressing the 'species' question in gorillas (Mayr 1942). The two most likely species concepts that we can apply to address the 'species' question in gorillas are evolutionary species concept (Simpson 1961) and phylogenetic species concept (Cracraft 1989). Previous D-Loop based phylogeny (Jensen-Seaman and Kidd 2001) has shown that the eastern gorillas are monophyletic and they are genetically distinct from the mountain gorillas. Anthony et al. (2007) have shown that the western gorillas, also, are monophyletic. So, it can be argued that both the eastern and western gorillas are maintaining their own lineages and evolving separately from each other. Therefore, according to evolutionary species concept they can be considered as two different species (Simpson 1961). Also, since very low genetic diversity has been noted within the eastern gorilla population (Jensen-Seaman and Kidd 2001), it can be argued that this population is potentially the smallest diagnosable cluster of individual organisms, independent of other such clusters. So, eastern gorillas can also

be considered as distinct species according to phylogenetic species concept (Cracraft 1989). Finally, western gorillas are thought to be more frugivorous than eastern gorillas (Ganas and Robbins 2005) and the eastern gorillas are thought to be more folivorous but their diet may vary according to the season (Yamagiwa et al. 1994). Based on the diets of the two gorilla populations, one can argue that they have two distinct niches, which may only partially overlap during certain time of the year. So, it can be speculated that the eastern and western gorillas can also be considered as two distinct species according to ecological species concept (Ridley 1993). To overcome the biological classification dilemma, Avise and John (1999) proposed a standardized temporal scheme of classification for all living species. Although his arguments were strong, this classification system may cause several taxonomic confusions to set a time bar for each taxonomic rank. However, as mentioned before, in the current study the two gorillas are considered as different species by comparing them with chimpanzee and bonobo. Although this method of species determination was crude, it supports evolutionary and phylogenetic species concepts. Therefore, we can conclude that the two gorillas, although not found in sympatry, have either completely become two different species or are in the process of becoming distinct species.

Anatomical and molecular data from extinct hominins such as Neanderthals, Denisovans, and the hominins from Sima de los Huesos reveal a complex pattern of isolation and migration, potentially the result of hybridization between subspecies or species, sex-biased gene flow, incomplete linage sorting, mitochondrial paraphyly, and geographically structured variation (Krause et al. 2010, Reich et al. 2010, Meyer et al. 2012, Prüfer et al. 2013). Indeed, this is precisely what is observed in modern African

apes (Mailund et al. 2012). The data from our study provide the most accurate dates of mitochondrial lineage divergence in gorillas, both within the diverse western species, as well as between the western and eastern species. Combining the mitochondrial data and recent whole nuclear genome sequences with realistic estimates of migration rates and distances of both sexes in gorillas could be used to develop more complex models of ape speciation processes, which could in turn be used to inform scenarios to explain Eurasian hominin demographic evolution.

References

Anthony NM, Clifford SL, Bawe-Johnson M, Abernathy KA, Bruford MW, Wickings EJ 2007. Distinguishing gorilla mitochondrial sequences from nuclear integrations and PCR recombinants: guidelines for their diagnosis in complex sequence databases. Mol Phylogenet Evol. 43:553-566.

Archibald JK, Mort ME, Crawford DJ 2003. Bayesian inference of phylogeny: a non-technical primer. Taxon 52:187-191.

Avise J, Johns GC 1999. Proposal for a standardized temporal scheme of biological classification for extant species. Proc Natl Acad Sci. 96:7358-7363.

Ballard JWO, Whitlock MC 2004. The incomplete natural history of mitochondria. Mol Ecol. 13:729-744.

Birky WC 2001. The inheritance of genes in mitochondria and chloroplasts: Laws, Mechanisms and models. Annu Rev Genet. 35:125-148.

Bjork A, Liu W, Wertheim JO, Hahn BH, Worobey M 2011. Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. Mol Biol Evol. 28:615-623.

Chan YC, Roos C, Inoue-Murayama M, Inoue E, Shih CC, Pei KJC, Vigilant L 2010. Mitochondrial genome sequences effectively reveal the phylogeny of Hylobates gibbons. PLoS One 5:e14419.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules, networks, populations: New York:Springer Verlag. p. 207-232.

Cracraft J. 1989. Speciation and its ontology: the empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. In: Otte D, Endler JA, editors. Speciation and its consequences: Sinauer, Sunderland, MA, USA. p. 28-59.

Darriba D, Taboada GL, Doallo R, Posada D 2012. jModelTest 2:more models, new heuristics and parallel computing Nat Methods 9:772.

Delisle I, Strobeck C 2005. A phylogeny of the Caniformia (order Carnivora) based on 12 complete protein-coding mitochondrial genes. Mol Phylogenet Evol. 37:192-201.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond AJ, Rambaut A 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol.:7-214.

Faith DP 1985. Distance methods and the approximation of most-parsimonious trees. Syst Zool. 34:312-325.

Farris JS 1999. Likelihood and inconsistency. Cladistics 15:199-204.

Farris JS, Albert VA, Källersjö M LDKAG 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12:99-124.

Felsenstein J 1968. Statistical inference and the estimation of phylogenies. [Ph.D. thesis]. [Chicago, IL, USA]: University of Chicago, Chicago.

Felsenstein J 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27:401-410.

PHYLIP (Phylogeny Inference Package) version 3.5c [Internet]. Available from; 1993. Available from:http://evolution.genetics.washington.edu/phylip.html

Ganas J, Robbins MM 2005. Ranging behavior of the mountain gorillas (Gorilla beringei beringei) in Bwindi Impenetrable National Park, Uganda: a test of the ecological constraints model. Behav Ecol Sociobiol. 58:277-288.

Garner KJ, Ryder OA 1996. Mitochondrial DNA diversity in gorillas. Mol Phylogenet Evol. 6:39-48.

Gillham NW. 1994. Organelle genes and genomes: Oxford University Press.

Green PJ 1995. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711-732.

Groves C 1996. Do we need to update the taxonomy of gorillas? Gorilla J.:3-4.

Groves CP 1967. Ecology and taxonomy of the gorilla. Nature 213:890-893.

Groves CP 1971. Distribution and place of origin of the gorilla. Man 6:44-51.

Groves CP. 2001. Primate taxonomy. Washington and London: Smithsonian Institution Press.

Groves CP. 2003. A history of gorilla taxonomy. In: ML TAaG, editor. Gorilla Biology: a multidisciplinary perspective:Cambridge University Press.

Guindon S, Gascuel O 2003. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. Syst Biol. 52:696-704.

Ho SYW 2007. Calibrating molecular estimates of substitution rates and divergence times in birds. J Avian Biol. 38:409-414.

Huelsenbeck J, Larget B, Miller RE, Ronquist F 2002. Potential application and pitfalls of Bayesian inference of phylogeny. Syst Biol. 51:673-688.

Huelsenbeck J, Ronquist F, Nielsen R, Bollback JP 2001. Bayesian influence of phylogeny and its impacton evolutionary biology. Science 294:2310-2314.

Huelsenbeck JP 1995. Performance of phylogenetic methods in simulation. Syst Biol. 44:17-48.

Huelsenbeck JP, Hillis DM 1993. Success of phylogenetic methods in the four-taxon case. Syst Biol. 42:247-264.

Huelsenbeck JP, Ronquist F 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-775.

Hurst GDD, Jiggins FM 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc Roy Soc Biol. 272:1525-1534.

Ingman M, Kaessmann H, Paabo S, Gyllensten U 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708-713.

Inoue E, Akomo-Okoue EF, Ando C, Iwata Y, Judai M, Fujita S, Hongo S, Nze-Nkogue C, Inoue-Murayama M, Yamagiwa J 2013. Male genetic structure and paternity in western lowland gorillas (Gorilla gorilla gorilla). Am J Phys Anthropol. 151:583-588.

AWTY [Internet]. 2014. Available from:http://ceb.csit.fsu.edu/awty/

Jensen-Seaman MI, Kidd KK 2001. Mitochondrial DNA variation and biogeography of East African gorillas. Mol Ecol. 10:2241-2247.

Jensen-Seaman MI, Sarmiento EE, Deinard AS, Kidd KK 2004. Nuclear integrations of mitochondrial DNA in gorillas. Am J Primatol. 63:139-147.

Kimball RT, Crawford DJ, Smith EB 2003. Evolutionary processes in the genus Coreocarpus: insights from molecular phylogenetics. Evolution 57:52-61.

Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, Pääbo S 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 464:894-897.

Krause J, Unger T, Nocon A, Malaspinas AS, Kolokotronis SO, et al. 2008. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evol Biol. 8:220.

Kunhner M K FJ 1994. A simulation comparison of phylogeny-algorithms under equal and unequal evolutionary rates. Mol Biol Evol. 11:459-468.

Larizza A, Pesole G, Reyes A, Sbisà E SC 2002. Lineage specificity of the evolutionary dynamics of the mtDNA D-loop region in rodents. J Mol Evol. 54:145-155.

Larkin MA, Blackshields G, Brown NP, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948.

Lewis PO. 1998. Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. In: Solits DE SP, Doyle JJ editor. Molecular Systematics of Plants II: Kluwer, Boston, USA. p. 132-163.

Lewis PO 2001. Phylogenetic systematics turns over a new leaf. Trends Ecol Evol. 16:30-37.

Li S 1996. Phylogenetic tree construction using Markov Chain Monte Carlo. [Ph.D. thesis, Ohio State Univ.,Columbus]:Ohio State University, Columbus, OH, USA.

Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, Lunter G, Prüfer K SAHASMH 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet. 8:e1003125.

Matsui A, Rakotondraparany F, Munechika I, Hasegawa M, Horai S 2009. Molecular phylogeny and evolution of prosimians based on complete sequences of mitochondrial DNAs. Gene 441:53-66.

Mau B 1996. Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods. [Ph.D. thesis, Univ.Wisconsin, Madison]: University of Wisconsin, Madison, USA.

Mayr E. 1942. Systematics and the origin of species from the viewpoint of a zoologist. New York: Columbia University Press.

Metropolis N, Rosenbluth AW, Rosenbluth N, Teller AH, Teller E 1953. Equations of state calculations by fast computing machines. J Chem Phys. 21:1087-1091.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K dFC, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222-226.

Miya M, Kawaguchi A, Nishida M 2001. Mitogenomic exploration of higher teleostean phylogenies: a case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial DNA sequences. Mol Biol Evol. 18:1993-2009.

Morin PA, Archer FI, Foote AD, Vilstrup J, Allen EE, et al. 2010. Complete mitochondrial genome phylogeographic analysis of killer whales (Orcinus orca) indicates multiple species. Genome Res. 20:908-916.

Moritz C. 1996. Molecular systematics. Massachusetts: Sinauer Associates, Sunderland, MA, USA.

Paradis E, Claude J, Strimmer K 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289-290.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, et al. 2013. Great ape genetic diversity and population history. Nature 499:471-475.

Prüfer K RFPNJFSSSSHARGSPHdFC, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43-49.

Raaum RL, Sterner KN, Noviello CM, Stewart CB, Disotell TR 2005. Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. J Hum Evol. 48:237-257.

FigTree v1.4.2 [Internet]. 2014. Available from:http://tree.bio.ed.ac.uk/software/figtree

Tracer v1.5 [Internet]. 2009. Available from:http://beast.bio.ed.ac.uk/Tracer

Rannala B, Yang Z 1996. Probability distribution of molecular evolutionary trees:a new method of phylogenetic inference. J Mol Evol. 43:304-311.

Reed DL, Carpenter KE, deGravelle MJ 2002. Molecular systematics of the jacks (Perciformes: Carangidae) based on mitochondrial cytochrome b sequences using parsimony, likelihood, and Bayesian approaches. . Mol Phylogenet Evol. 23:513-524.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053-1060.

Ridley M. 1993. Evolution. Oxford: Blackwell Scientific Publications.

Rohland N, Malaspinas AS, Pollack JL, Slatkin M, Matheus P, et al. 2007. Proboscidean mitogenomics: chronology and mode of elephant evolution using mastodon as outgroup. PLoS Biol. 5:1663-1671.

Rokas A, Ladoukakis E, Zouros E 2003. Animal mitochondrial DNA recombination revisited. Trends Ecol Evol. 18:411-417.

Ruvolo M 1996. A new approach to studying modern human origins: hypothesis testing with coalescence time distributions. Mol Phylogenet Evol. 5:202-219.

San Mauro D, Gower DG, Zardoya R, Wilkinson M 2006. A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. Mol Biol Evol. 23:227-234.

Sarmiento E, Butynski T 1996. Present problems in gorilla taxonomy. Gorilla J.:5-7.

Satoh M, Kuroiwa T 1991. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Exp Cell Res. 196:137-140.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483:169-175.

Schrago CG, Russo CAM 2003. Timing the origin of New World monkeys. Mol Biol Evol. 20:1620-1625.

Siddall ME 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris Zone. Cladistics 14:209-220.

Simpson GG. 1961. Principles of animal taxonomy. New York: Columbia University Press.

Steel M, Penny D 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol. 17:839-850.

Swofford D, Olsen G. 1990. Phylogeny reconstruction. In: Hills D, editor. Molecular Systematics:Sinauer Associates, Sunderland, MA, USA. p. 411-501.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol. 28:2731-2739.

Thalmann O, Hebler J, Poinar HN, Pääbo S VL 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of human and other great apes. Mol Ecol. 13:321-335.

Uzzell T, Corboin KW 1971. Fitting discrete probability distribution to evolutionary events. Science 172:1089-1096.

Xu X, Arnason U 1996. The mitochondrial DNA molecule of Sumatran orangutan and a molecular proposal for two (Bornean and Sumatran) species of orangutan. J Mol Evol. 43:431-437.

Yamagiwa J, Mwanza N, Yumoto T, Maruhashi T 1994. Seasonal change in the composition of the diet of eastern lowland gorillas. Primates 35:1.

Yu L, Li YW, Ryder OA, Zhang YP 2007. Analysis of complete mitochondrial genome sequences increases phylogenetic resolution of bears (Ursidae), a mammalian family that experienced rapid speciation. BMC Evol Biol. 7:198.

Zhang P, Papenfuss TJ, Wake MH, Qu L, Wake DB 2008. Phylogeny and biogeography of the family Salamandridae (Amphibia:Caudata) inferred from complete mitochondrial genomes. Mol Phylogenet Evol. 49:586-597.

Chapter 3: Evolution of Cartilage Acidic Protein 1 (CRTAC1) in response to sexual selection among hominoid primates

3.1 Introduction

3.1.1 Hominoid primate society and sexuality

All hominoids share a great degree of genetic similarity. Although all apes are genetically similar, the social lives of the great apes vary from each other. The hominoid society can range from simple monogamy to complex multi male-multi female groups. Humans are thought to be monogamous (Fleagle 1999) but polygyny seems to be the commonest mating system in human society (Low 2007). On the contrary chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*) live in multimale-multifemale groups, where females mate with multiple males during ovulation (Hasegawa and Hiraiwa-Hasegawa 1990, Kano 1992). The chimpanzee mating system is also thought to be opportunistic, where females mate promiscuously but in a nonrandom manner (Oda 1999). Gorilla (*Gorilla gorilla* and *Gorilla beringei*) society has a distinct dominance hierarchy, where only the dominant alpha male typically mates with the females of that group (Watts 1990). The males of the Asian orangutan (*Pongo pygmaeus*) have two distinct morphs: flanged and unflanged. These two male morphs have two different mating strategies. The larger flanged males are chosen by females when the females are most fertile and participate in cooperative mating (Utami et al. 2002) while the mating involving the unflanged males are forced and less cooperative (Mitani 1985). This type of mating system is called "dispersed harem polygyny" (Maggioncalda et al. 2002), where selection pressure created an alternative mating strategy for the subordinate males. The

gibbons (*Hylobates sensu lato*) are supposed to be obligatory monogamous and maintain lifelong pair-bonds (Tilson 1981). The above-mentioned mating systems are generalized summaries of different social structures observed among hominoid primates. Primate mating systems, in reality, show great degree of variability and plasticity. For example, polygynous mating system coexists with monogamy in Black-crested gibbon (*Hylobates concolor*) (Wang and Wang 1999). Unlike lowland gorillas, in mountain gorillas (*Gorilla beringei beringei*), where ~40% of social units are multimale groups, females have been observed mating with multiple males (Robbins 1999). Although rare, 0.05% human societies are polyandrous (Low 2007).

3.1.2 Sperm competition and sexual selection

Sexual selection can be defined as "a struggle between the males for the possession of the females" (Darwin 1859). According to Darwin, sexual selection can operate mainly in two ways: male-male competition or intrasexual selection, and female choice or intersexual selection. In male competition the females do not participate actively. The males fight among each other to defeat their rivals, directly or indirectly. In female choice, the females participate actively and choose the most desirable male (Darwin 1871). The male-male competition can take place both before and after copulation.

The male-male combat after copulation does not involve aggression between the males but involve sperm competition. Sperm competition takes place when gametes from two or more males compete to fertilize the same ova (Parker 1970). The sperm competition may lead to larger testes, as they are required to accumulate a larger mass of seminiferous tubules-the sperm producing tissues, which adds to the larger amount of

sperm in the ejaculates (Dixson 1993). In hominoids the size of the testes depends on the social structure of that species (Dixson 1993). In a multimale-multifemale group like in chimpanzees and bonobos, where the females generally mate with multiple males, the sperm competition is high. As a result of this chimpanzees and bonobos possess the largest testes among the hominoids (Harcourt et al. 1981). In the monogamous hominoids like gibbons, the testes size is smaller in relation to their body weight. A similar situation is observed in the case of the polygynous hominoids like gorilla, who have small testes in relation to their body (Harcourt et al. 1981). This may be due to the fact that there is essentially no sperm competition among the gorillas for fertilization.

Female promiscuity in chimpanzees has led to several different modifications in the sperm of this animal. For example the chimpanzee has a higher sperm motility (Møller 1988), higher sperm concentration in the ejaculates (Møller and Brickhead 1989), and higher ratio of seminiferous tubule to connective tissue (Harvey and Harcourt 1984). Moreover, the chimpanzee sperm swim faster than its human counterpart (Nascimento et al. 2008), has significantly larger mid-piece volume that contains energy producing mitochondria (Anderson and Dixson 2002), and has higher mitochondrial membrane potential (Anderson et al. 2007).

Sperm competition may have also caused several chemical changes in the ejaculates. The chemical species derived from the seminal vesicles and prostate help in seminal coagulation soon after the ejaculation (Dixson and Anderson 2002). The coagulum can be either soft or more compact like a copulatory plug. A copulatory plug is formed due to semen coagulation and it may help in sperm positioning, prevention of sperm loss, and generation of a physical barrier (Dixson 2012). Firstly by generating a

physical barrier, it prevents the entry of other sperm to the female genital tract and secondly, it minimizes the sperm loss and protects the sperm till they reach the uterus (Dixson and Anderson 2002). Chimpanzees and bonobos, with high degree of female promiscuity, produce rigid and compact copulatory plugs to avoid sperm competition. Dixson and Anderson (2002) have shown that there is a direct correlation between high degree of sperm competition and copulatory plug formation, as copulatory plug formation is more common in animals where females generally mate with multiple males.

The semen of hominoid primates shows a great degree of variation in respect to coagulation depending on the degree of sperm competition prevailing in the species concerned. Human semen coagulates into a semisolid mass but liquefies soon at 37° C (de Lamirande 2007), gorilla semen never coagulates and remains liquid (Martin and Gould 1977), while the chimpanzee semen coagulates into a rigid copulatory plug (Dixson 1998) as mentioned before.

3.1.3 Proteins found in the seminal fluid

The seminal proteins that are secreted from seminal vesicles, prostate gland and bulbourethral glands play various different roles including semen coagulation, seminal liquifaction, nutrient transport, and immunological roles. Some seminal proteins like prolactin-induced protein (PIP) (Gaubin et al. 1999), and cathelicidin (CAMP) (Zelezetsky et al. 2006) show antibacterial activity. Transferrin inhibits bacterial growth (Ford 2001). The proteins that are found in the coagulum in high abundance include semenogelins 1 and 2 (SEMG1 and SEMG2) and fibronectin 1 (FN1) (Lilja et al. 1987, Malm et al. 1996). SEMG2 helps in seminal coagulation when it is cross- linked by prostate- derived transglutaminase (Lin et al. 2002, Lundwall et al. 1997). In humans the

prostate spectfic transgultaminase is transglutaminase 4 (TGM4) (Dubbink et al. 1998). SEMG monomers are polymerized by TGM4 by crosslinking. High degree of crosslinking is expected in chimpanzees and bonobos with high degree of polyandry.

Seminal proteins also include various proteases like kallikrein 2 and 3 (KLK2 and KLK3) and prostatic acid phosphatases (ACPP) that help in liquifaction of coagulated semen (Lilja 1985, Lövgren et al. 1999, Brillard- Bourdet et al. 2002).

Proteomic analysis of human seminal plasma by Fung et al. (2004), and Pilch and Mann (2006) have found several other abundant proteins in semen including lactotransferrin (LTF), transferrin (TF), albumin (ALB), clusterin (CLU), the laminins (LAMA, LAMB, LAMC), and Zn- binding α- 2- glycoprotein (AZGP1).

3.1.4 Molecular evolution of seminal proteins

Genes or proteins may undergo changes under the evolutionary forces of mutation, migration, random genetic drift, and selection. The term selection includes all kind of selection processes including sexual selection. In late 1960s Motoo Kimura proposed the neutral model of molecular evolution, according to which most of the changes taking place in the genome are selectively neutral and caused by random genetic drift (Kimura 1983). Various authors have criticized the neutralist model of Motoo Kimura (reviewed in Hahn 2008). But it still remains as the "null model" for molecular evolution providing the basis for many statistical tools that determine the strength of natural selection (Kreitman 2000, Nielsen 2001, Hahn 2007). Various methods have been formulated to test whether selection is operating in the genome including Tajima's D, (Tajima 1989), Hudson–Kreitman–Aquade test or HKA test (Hudson et al. 1987), McDonald- Kreitman test or MK test (McDonald and Kreitman 1991).

A more simplified codon based method is just to calculate the ratio of non-synonymous nucleotide substitution rate (d_N) to the synonymous substitution rate (d_S) in protein coding DNA sequences (Yang and Bielawski 2000). Synonymous nucleotide substitutions are those that do not change the amino acid sequence and non-synonymous substitutions change the amino acid sequence. This ratio is denoted by ω. Under neutrality the non-synonymous amino acid changes will be fixed at the same ratio as a synonymous substitution and ω will be 1. When the amino acid substitution is deleterious, it will be fixed at a lower rate compared to a given synonymous substitution due to the action of purifying selection and ω will be < 1. Finally, if the amino acid substitution is advantageous for the organism, it will be fixed at a higher rate compared to the synonymous substitution rate due to positive selection generating a positive value for ω ($\omega > 1$) (Yang and Bielawski 2000). The first step of this process is to count the number of synonymous and non-synonymous changes (M_S and M_A). Then M_S and M_A are normalized by dividing with the number of synonymous sites and non-synonymous sites (N_S and N_A) respectively, calculated using the codon table. Finally, a suitable nucleotide substitution model is chosen to calculate genetic distances d_N and d_S. Choosing the right substitution model is very crucial to get accurate values of d_N and d_S (Yang and Bielawski 2000).

Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang 2007) software incorporates various different models for calculating ω, taking various nucleotide substitution models into consideration. The basic model or the uniform model assumes a single ω for all branches (Yang and Neilson 1998). The branch models or the free-ratio models allow ω to vary among branches in the phylogeny (Yang and Nielson 1998). A

third type, known as site models, allows ω to vary among codons or amino acids in a protein (Nielsen and Yang 1998, Yang 2000b). The branch-site model incorporates both a branch and a site model and allows ω to vary both among amino acids in a protein and among branches in the phylogeny (Yang and Nielson 2002).

Selection operates on various genes in the human genome that directly or indirectly participate in reproduction. Signs of recent selective sweeps have been observed in genes like *SPAG4* (Spaghetti 4), *ODF2* (Outer Dense Fiber Of Sperm Tails 2) (Voight et al. 2006) and *SPAG6* (Williamson et al. 2007) that aid in sperm motility. *CATSPER1* (Cation Channel, Sperm Associated 1) that facilitates sperm hyperactivation during egg penetration also shows sign of positive selection (Podlaha and Zhang 2003). *ZP3* (Zona protein 3, found on zona pellucida on egg) that helps in sperm recognition and acrosome reaction is under strong positive selection (Swanson et al. 2001). ADAM (A Disintegrin And Metalloproteas) family proteins like ADAMs 1, 2 and 32 that help in sperm-egg binding are also under strong positive selection (Swanson et al. 2003). Many proteins found in seminal plasma show sign of positive selection. For example, proteins like prolactin-induced protein (PIP), beta-microseminoprotein (MSMB), and cathelicidin (CAMP), with antibacterial properties, are under positive selection in primates (Clark and Swanson 2005, Zelezetsky et al. 2006). Hominoid primates show variation among each other in terms of the operation of selective forces on reproductive genes. *TGM4*, for example, shows a positive selection in chimpanzees and bonobos, while it has probably become nonfunctional in gorilla with deletions in the coding region (Clark and Swanson 2005, Carnahan and Jensen-Seaman 2008). *SEMG1* and *SEMG2* also show indications of positive selection in chimpanzees and bonobos (Jensen-Seaman and Li 2003, Dorus et al.

2004). In gorillas both *SEMG1* and *SEMG2* exist with premature stop codons (Jensen-Seaman and Li 2003, Kingan et al. 2003). *KLK2* also shows the sign of positive selection in chimpanzees but has become nonfunctional in gorilla (Clark and Swanson 2005). Two testes specific gene families *PRAME* (Preferentially Expressed Antigen In Melanoma) and *SPANX* (sperm protein associated with the nucleus on the X chromosome) show positive selection during human evolution (Kouprina et al. 2004, Birtle et al. 2005, Gibbs et al. 2007) probably indicating modifications in spermatogenesis in humans.

3.1.5 *In vitro* promoter expression assay to identify regulatory differences among hominoids

Eukaryotic gene regulation is divided into several different steps: transcription initiation, elongation and termination, mRNA processing and splicing, translation, and post translation modifications. One of the most important regulatory steps among them is transcription initiation that involves both accession of the gene and proper placement and function of transcription machinery. In other words, this process includes regulation of transcriptional initiation, chromatin condensation and decondensation, DNA acetylation or methylation (reviewed by Berger 2000, Li et al. 2007a, Orphanides and Reinberg, 2002, Pugh 2000, Venters and Pugh 2009). Transcription initiation is aided by various *cis* regulatory elements (non-coding DNA sequences) including promoters, enhancers, silencers and insulators (Venters and Pugh 2009). Promoter regions are generally located just 1 -2kb upstream of the transcriptional start site (TSS) and so are the easiest to identify and characterize (Maston et al. 2006). The promoter regions can be imagined as a collection of transcription factor (TF) binding sites, where TFs bind differentially and

regulate transcription. Transcription machinery (*trans* regulatory elements) binds to the regulatory non-coding DNA sequences and carry out transcription.

It is thought that most morphological adaptations take place through changes in non-coding DNA sequences (Haygood et al. 2010). Since natural selection can only operate on phenotype (not on genotype), which is the outcome of gene expression, all stages of gene expression are under natural selection (Wray et al. 2003). A great deal of gene regulation takes place at the transcriptional level that makes transcription ideal target of natural selection (Wray et al. 2003). Although not completely understood, it is speculated that the changes in *cis* elements cause changes in mRNA expression, which may lead to adaptive evolution (Chabot et al. 2007). Abzhanov et al. (2004) have shown that the *cis*-regulatory mutations and the changes in gene expression are behind the differential beak morphology of various Darwin's finches. Similar events can be observed in case of wing pigmentation in fruit flies (Stern 1998, Gompel et al. 2005), maize branching pattern (Clark et al. 2006), pelvic reduction in sticklebacks (Cresko et al. 2004, Shapiro et al. 2004), and parental care in rodents (Hammock and Young 2005). Since most transcription factors that help in transcriptional initiation bind within ~1kb of TSS, this area is often subjected to the forces of natural selection. Although in some cases a few nucleotide substitutions in this region can cause substantial change in gene expression (Storgaard et al. 1993, Haudek 1998), it is not universally true (Takahashi et al. 1999, Wolff et al. 1999). In humans only ~20% of polymorphic sites within the putative promoter region estimated to have an effect on gene regulation (Buckland et al. 2004a, b).

One of the best approaches to assess *cis* regulatory variation is to investigate differential transcriptional activity by developing reporter gene assays (Wray 2007). *In vitro* cell culture and luciferase assay, which has successfully been used by many authors for the reporter gene assay, (Huby et al. 2001, Rockman et al. 2005, Inoue-Murayama et al. 2006, Loisel et al. 2006, Chabot et al. 2007), can provide convincing results and may be the most appropriate technique for studying transcriptional regulation. In this method, the putative promoter region (1-2 kb upstream of TSS) is first amplified from the genomic DNA, and subsequently cloned into a firefly luciferase reporter vector and transfected into specific cell lines with or without a control plasmid. The expression level is then measured by taking the ratio of signal (firefly luciferase) to control (*Renilla* luciferase) or just by taking the signal from the firefly luciferase.

3.1.6 Introduction to Cartilage Acidic Protein 1 (CRTAC1)

CRTAC1 is often used as a marker to distinguish chondrocytes from osteoblasts (Steck et al. 2007). It is a glycosylated extracellular matrix protein (Steck et al. 2007). Homology modeling suggests that CRTAC1 has a β-propeller structure similar to integrin. It has an EGF-like calcium-binding domain in the C- terminal end (Redruello et al. 2010) (Fig. 3.1). Beside this functional domain, it has two additional domains: FG-GAP that folds into a β-propeller structure, and UnbV_ASPIC conserved protein domain. Both domains are found in integrin-like proteins.

Recent phylogenetic analysis showed that CRTAC1 is conserved from cyanobacteria to humans (Redruello et al. 2010). There are two principal cartilage acidic proteins found in the vertebrates. Teleosts uniquely have CRTAC2 in addition to

CRTAC1. Tetrapods only have CRTAC1 and do not possess CRTAC2 (Redruello et al. 2010).

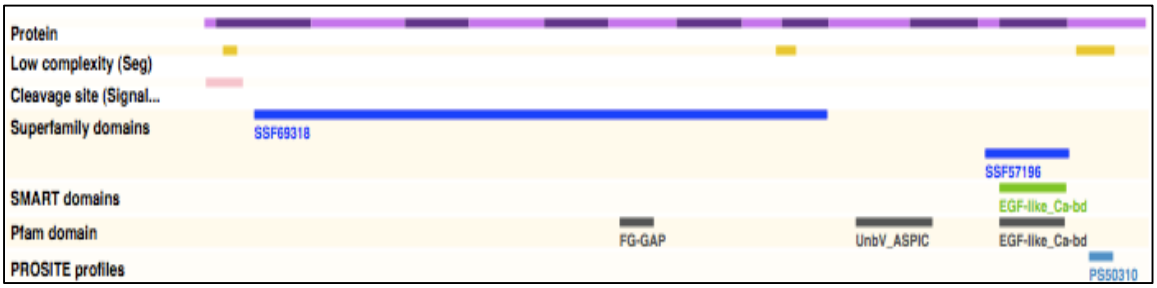


Figure 3.1: Screenshot of CRTAC1 protein from Ensemble genome browser (release 74) showing the location of different domains and cleavage signals

CRTAC1 is located in Chromosome 10 in humans in between coordinates 99,624,757 and 99,790,585 (0 based coordinate, hg19). It is a large gene with 15 exons and it shares the last exon with tail-to-tail oriented gene *GOLGA7B* (Fig. 3.2). In all metazoans the start codon (ATG) is located in the first exon and the stop codon is in the 15th exon. An additional start codon is present in exon 2. There are three poly (A) sites: one each at the end of exon 14 and exon 15, and a third one at intron 14. Transcription ends in one of these three locations.

There are two miRNA target sites near exon 15. Two Mir-5441genes are found in *CRTAC1*: one near exon 2 and the other near exon 6. H3K27Ac signal is associated with active regulatory elements outside the promoter region. ENCODE database have marked H3K27Ac signals for human genes in 7 cell lines (Bernstein Lab data at the Broad Institute). Strong H3K27Ac mark can be observed in intron 11 of *CRTAC1* for all 7 cell lines. The 2.2kb region in between *CRTAC1* exons 11 and 12 with strong H3K27Ac mark

is shown in Fig. 3.2. This region is also one of the strongest DNase hypersensitive regions (Dunham et al. 2012, Thurman et al. 2012) in *CRTAC1*. The DNase hypersensitivity signal is visible in LNCaP cells (Fig. 3.3).

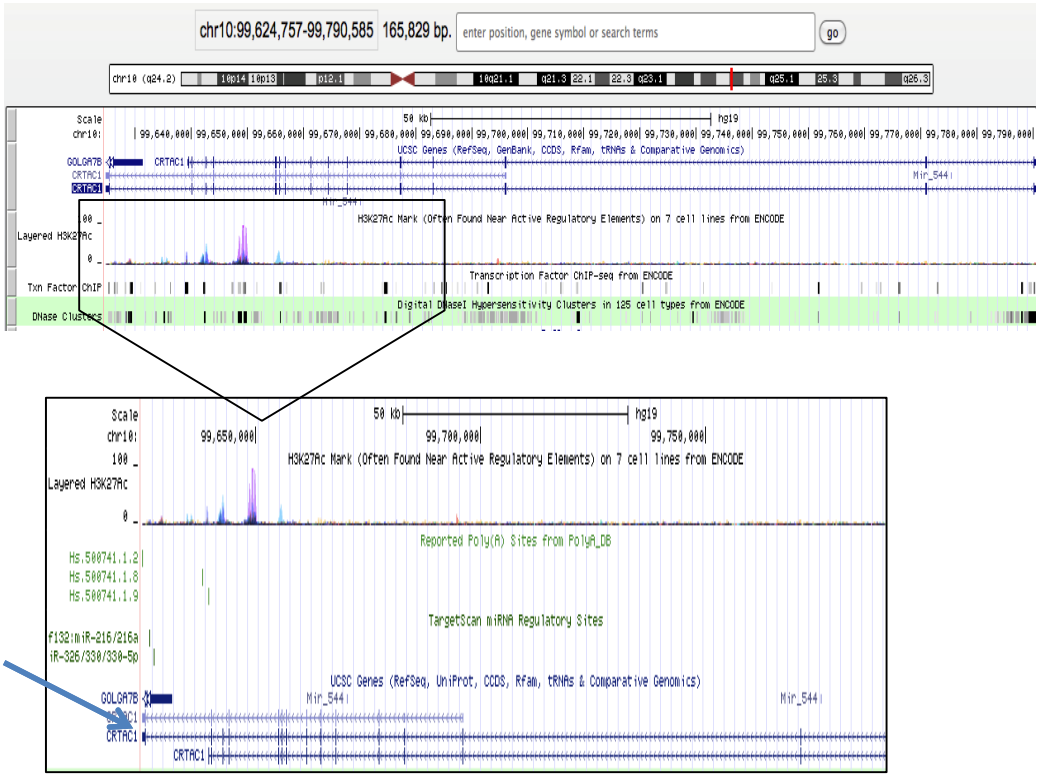


Figure 3.2: Screen shot of UCSC genome browser, showing the genomic location of *CRTAC1* along with miRNA regulatory sites, Poly (A) site, H3K27Ac site. *CRTAC1* and *GOLGA7B* overlapping exon in shown by the arrow. *CRTAC1* is transcribed right to left in this representation

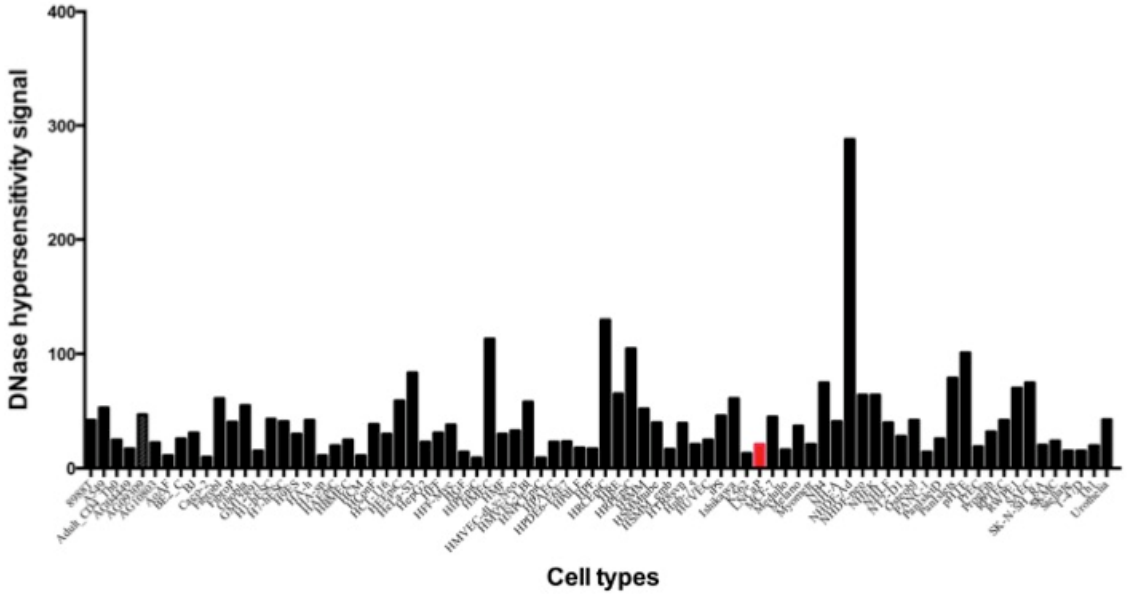
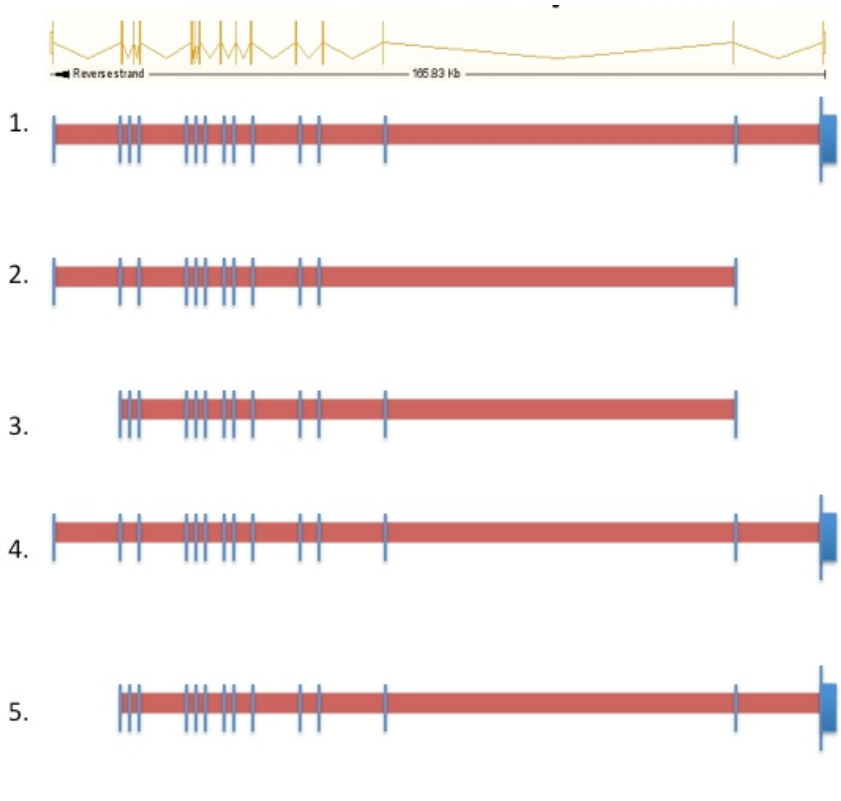


Figure 3.3: DNase hypersensitive signals visible in different human tissues for the regulatory region in intron 11. The red bar shows the DNase signal in LNCaP prostate cell line. The graph was constructed in Graphpad Prism statistical package using ENCODE data (Dunham et al. 2012, Thurman et al. 2012). Y-axis shows normalized signal value. Highest signal was observed in NHDF-Ad (an adult dermal fibroblast) cell line.

CRTAC1 has seven predicted splice variants (five of them are shown in Fig. 3.4) in humans and potentially three naturally found protein isoforms (data from Ensembl genome browser 2013, release 74). Splice variant ENST00000370597 (No. 1 in Fig. 3.4) and ENST00000298819 (No. 4 in Fig. 3.4) start at exon 1 and end at exon 15, but ENST00000298819 lacks the potentially functional EGF-like Ca^{+2} binding domain. ENST00000413387 starts at exon 1 but ends at exon 12. It too lacks the EGF like Ca^{+2} binding domain (not shown in figure). ENST00000370591 starts at exon 1 but creates at stop codon at intron 14 by alternative splicing (No. 5 in Fig. 3.4). ENST00000309155 starts at exon 2 and ends at exon 15 (No. 2 in Fig. 3.4).



Source	Protein size
Eye, Brain, Lung, Cornea etc.	71 kD
Most Tissues	56 kD (EGF like Ca^{+2} binding domain missing)
Eye, Brain, Lung etc.	69 kD
Brain, synovial membrane of Reheumatoid arthritis	68 kD (EGF like Ca^{+2} binding domain missing)
Synovial membrane of Rheumatioid arthritis	70 kD

Figure 3.4: Cartoon representation of *CRTAC1* predicted splice variants in human, along with the tissues they are expressed and the different size proteins coded by them. Data modified from Ensemble genome browser (release 74). Gene is transcribed right to left in this representation

According to NCBI AceView database, CRTAC1 is a highly expressed gene, expressed in 56 tissues. 214 cDNA clones, reported in GenBank/dbEST show that the gene is expressed in synovial membrane tissue from rheumatioid arthritis (seen 50 times), eye (38), brain (23), lung (13), cornea (12), hypothalamus (11), cartilage (7), lens (6), hippocampus (5), knee (5), whole brain (5), and 45 other tissues.

3.1.7 Basic description of CRTAC1 putative promoter region

CRTAC1 promoter region has never been characterized in the published literature. ~1.9kb area around the transcriptional start site (TSS) of *CRTAC1* was selected as the putative promoter region (see Methods). All coordinates are calculated based on the location of human *CRTAC1* putative TATA box, considering the 'T' in the consensus sequence 'TATAAT' as +1. The forward primer is located at -1731bp upstream of putative TATA box and the reverse primer is located +154bp downstream of putative TATA box (Fig. 3.5). The putative promoter region is highly GC rich (> 70%). There are three different classes of GC repeats found in this region. CGG family simple repeats, GC family low complexity repeats, and CCG family simple repeats (Fig. 3.5). CGG family repeats cover ~200-250bp and are the longest among the three. CCG family repeats are found immediate downstream of the putative TATA box region. A GT microsatellite repeat is present in the putative promoter region, which starts from -1641bp upstream of the putative TATA box. When compared to human *CRTAC1* putative promoter region, chimpanzee putative promoter has a ~475bp gap in the current version of the genome sequence assembly (panTro4). This area starts immediately before the putative TATA box region and ends in intron 1. Gorilla has a longer (~1000bp) gap in the current assembly (gorGor3) that covers the same area missing in chimpanzee but misses more of intron 1. Orangutan and gibbon do not have any gap in the current genome assemblies for this region.

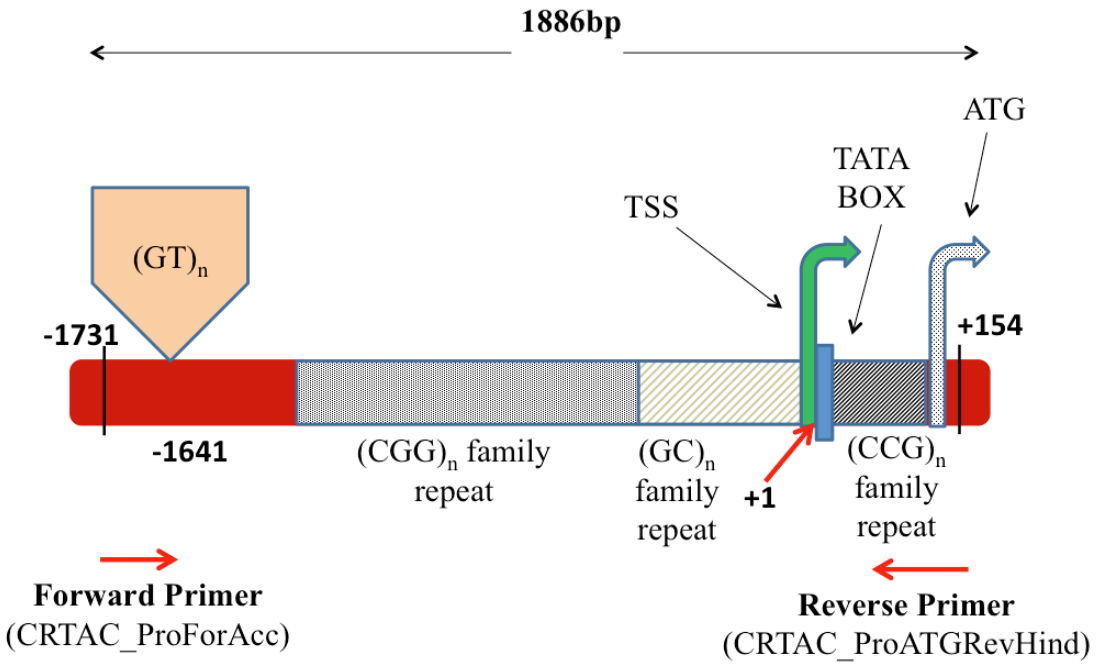


Figure 3.5: The putative promoter region of *CRTAC1* with locations of TATA box (blue box), GC rich regions (grey boxes), transcription start site (TSS) (green arrow), start codon (grey arrow) and GT microsatellite repeats (orange box).

3.1.8 Cartilage Acidic Protein 1 (CRTAC1) is potentially under sexual selection

CRTAC1 has never been described to have any role in sperm competition, sexual selection or reproductive biology. But recent 2D gel electrophoresis in our lab (Chovanec *et al.* 2011) detected the presence of CRTAC1 in chimpanzee seminal plasma as well as copulatory plug. Subsequent shotgun mass spectrometry data showed that CRTAC1 exists in 142 fold more concentration in chimpanzee semen compared to human semen. More interestingly, CRTAC1 was found to be 179 fold excess in the chimpanzee copulatory plug compared to human semen (Chovanec et al. 2011). This study suggested a potential role of CRTAC1 in sperm competition.

3.2 Methods

3.2.1 Samples used in the study and their sources

Six hominoid samples were used in this study for cloning and construction of reporter vectors. Additionally, 10 human, 10 chimpanzee, 10 gorilla, and 8 bonobo samples were used for genotyping the 'GT' microsatellite, present in the putative promoter region. The samples and their sources are summarized in Table 3.1.

Table 3.1 Samples used in the study with their sources

Species	**Samples with Sources**	**Purpose**
Human	NA15283	Genotyping
	NA15047	Genotyping
	NA15504	Genotyping
	NA15230	Genotyping
	NA15242	Genotyping
	NA15216	Genotyping
	NA15221	Genotyping
	NA15245	Genotyping
	NA15215	Genotyping
	NA15341	Genotyping
	MJS	Cloning and sequencing
Chimpanzee	PR496 (Coriell Institute)	Genotyping, Cloning and sequencing
	P. t. sweinfurthii - Kobi	Genotyping
	P. t. sweinfurthii - Harriet	Genotyping
	P. t. verus - Lottie	Genotyping
	P. t. verus - Lowie	Genotyping
	P. t. verus - Colin	Genotyping
	P. t. troglodytes - Dodo	Genotyping
	P. t. troglodytes - Cheetah	Genotyping
	P. t. troglodytes - Julie	Genotyping
	P. t. troglodytes - Noemie	Genotyping
		Genotyping
Bonobo (*Pan paniscus*)	Lomoko	Genotyping
	Lenore	Genotyping
	Matata	Genotyping
	Kevin	Genotyping
	Lody	Genotyping

	Maringa Bosonjo PR251 (Coriell institute)	Genotyping Genotyping Genotyping, Cloning and sequencing
Gorilla (*Gorilla gorilla gorilla*)	F'rika H G F E D C B A J'phine	Genotyping Genotyping, Cloning and sequencing Genotyping Genotyping Genotyping Genotyping Genotyping Genotyping Genotyping Genotyping
Orangutan	*Pongo pygmaeus* –PR253 (Coriell Institute)	Cloning and sequencing

3.2.2 Sequencing and 'GT' microsatellite genotyping of the putative promoter region from hominoids

The putative promoter region was first PCR amplified from the genomic DNA. The forward primer (CRTAC_ProForAcc) is located -1745 downstream of TSS and the reverse primer (CRTAC_ProATGRevHind) is located +184 upstream of TSS (Fig. 3.5). Since the *CRTAC1* putative promoter region is highly GC rich, it could not be amplified by regular polymerase chain reaction (PCR). I went through several modification and optimization processes to finally PCR amplify this region from all hominoids (see Results). PCR was carried out in 20 µl reaction, containing 1X ThermoPol B9004S PCR buffer (New England Biolab Inc) containing Mg^{2+}, 0.5U *Taq* DNA Polymerase, 250µM of dNTP (dATP, dCTP, dGTP, dTTP), 5% DMSO (v/v), 5% Betaine (v/v), 5% Glycerol (v/v), and 0.25µM of each primer. Thermal cycling started with 5 min denaturation at 95°C, followed by 6 cycles of denaturation (95°C, 15 sec), annealing (68°C, 15 sec with a Touchdown of 1°C per cycle), and primer extension (72°C, 3 min), it was followed by 6

cycles of denaturation (95°C, 15 sec), annealing (62°C, 15 sec with a Touchdown of 1°C per cycle), and primer extension (72°C, 3 min), then final 25 cycles of denaturation (95°C, 15 sec), annealing (56°C, 15 sec), and primer extension (72°C, 3 min), PCR concluded with a terminal extension at 72°C (10 min), and final holding at 4°C. A lower ramp speed (90%) was maintained throughout PCR. The ramp speed was decreased to 60% during annealing. The amplified products were gel purified using a 1% crystal violet agarose gel and Wizard® SV Gel and PCR Clean-Up System (Promega). Ethidium bromide gel was avoided to prevent DNA damage by UV light exposure. Entire putative promoter regions were sequenced from all hominoids using the two PCR primers and additional sequencing primers (Table 3.2). Sanger-sequencing was used on the Applied Biosystem platform (BigDye® Terminator v3.1 Cycle sequencing kit, Applied Biosystems 3100 and 3130 Genetic Analyzer, Life Technologies).

Additionally, the GT repeats in the putative promoter region were separately PCR amplified and genotyped from human, chimpanzee, gorilla and bonobos. For genotyping the 'GT' microsatellite repeat in the putative promoter region of *CRTAC1,* ~227bp was amplified using the primer set CRTAC_Msat_F_FAM and CRTAC1_REV1 (Table 3.1) from 10 humans, 10 chimpanzees, 10 gorillas, and 8 bonobos. The PCR products were then genotyped in Applied Biosystems 3130 Genetic Analyzer. The output files were analyzed using Peak Scanner Software v1.0 (Applied Biosystems). Population genetic analysis such as Allele counts, Observed and Expected Heterozygosity Calculation, and Test for Hardy-Weinberg Equilibrium (HWE) were carried out using this genotype data in Arlequin v3.5 (Excoffier et al. 2010).

3.2.3 Construction of reporter vectors containing putative promoters

CRTAC1 putative promoter region from the hominoid primates were cloned into luciferase reporter constructs. Once the PCR conditions were optimized with standard primers, the putative promoter regions were amplified using primers with restriction enzyme sites (5'-restriction enzyme sequence-primer sequence-3'). To avoid any random mutations that can be generated during PCR, a high-fidelity *Taq*-polymerase was used (iProof ™ High-Fidelity DNA Polymerase, BioRad). After PCR, the amplified products were gel purified using a 1% crystal violet agarose gel and Wizard® SV Gel and PCR Clean-Up System (Promega). The cleaned up products were sequence verified and compared to the published genome assemblies for integrity.

Table 3.2 List of primers used for cloning, sequencing and genotyping the promoter region

Primer name	**Purpose**	**Sequence (5'-3')**
CRTAC_Msat_F_FAM	Genotyping GT repeat	6-FAMN/TACTGTCCTAGACCCCTGAA
CRTAC_ProATGRevHind[1]	PCR and cloning	G**AAGCTT**AGCCGTCCTCCCGCTCTC
CRTAC_ProForAcc[1]	PCR and cloning	**GGTACC**TACTGTCCTAGACCCCTGAA
CRTAC1_FOR1	Sequencing and PCR standardization	TACTGTCCTAGACCCCTGAA
CRTAC1_FOR2	Sequencing	CACAGAGACCTGAAAACAGA
CRTAC1_FOR3	Sequencing	CCTACTATGTGCCAGGCTC
CRTAC1_FOR4	Sequencing	CTTAGCACCCCCATTCCC
CRTAC1_FOR5	Sequencing	GGACCCTCGCTTCCCTCC
CRTAC1_REV1	Sequencing and genotyping GT repeat	CCCCATCAAGCCTGTAAGGT
CRTAC1_REV2	Sequencing	GCTTTGATCACAGGTACTGCC
CRTAC1_REV3	Sequencing	CTTTATCCAGCCTGGGGA

CRTAC1_REV4	Sequencing	GTAACCTTCAGGCGGCAG
CRTAC1_REV5	Sequencing and PCR standardization	CACCGGTGCAGATACTCA

[1]The location of the restriction site is shown in bold

The purified and sequence-verified PCR products were then cloned into TOPO® TA vector (Life Technologies, Fig. 3.6).

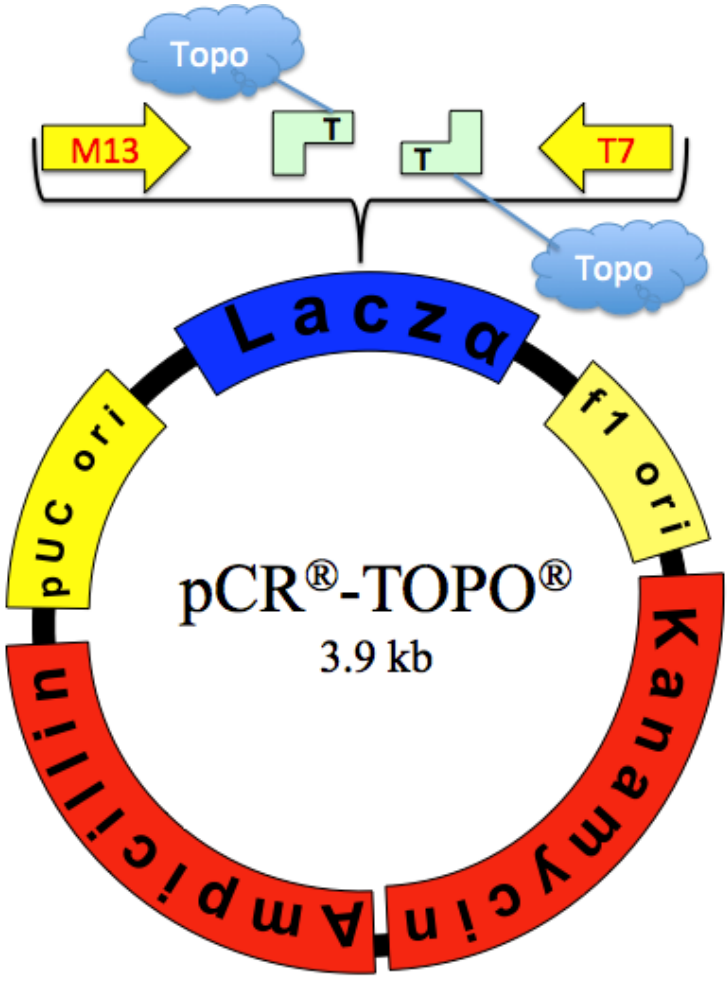


Figure 3.6: TOPO® TA vector (Life Technologies) showing 'T' overhangs, primer and multiple cloning site on lacZα and antibiotic resistant genes

Since TOPO® TA vector has a 'T' overhang, the PCR product needs to have multiple 'A's added at the 3' end to become compatible with the vector. The tendency of all standard *Taq* polymerases to add multiple 'A's at the 3' end of the

product was used to make the PCR products compatible with the TOPO® TA vector. The purified PCR product was incubated with a standard *Taq* polymerase, *Taq* polymerase buffer, and dNTPs at 72°C for 15 minutes. 4µl of the incubated PCR product was then incubated with 1µl TOPO® TA vector and 1µl salt solution (total 6µl reaction mix) at room temperature overnight (~ 16 hours). The ligation mixes from all hominoids were then transformed into One Shot® TOP 10 chemically competent cells following the manufacturer's protocol (http: //tools.lifetechnologies.com/content/sfs/manuals/topotaseq_man.pdf), plated onto LB agar plates with kanamycin and X-Gal, and incubated at 37°C overnight. Next morning blue/white screening was performed. Since the inclusion of the PCR product makes LacZα non-functional, it cannot participate in α complementation procedure to generate β-galactosidase that digests X-gal. As a result white bacterial colonies are generated. If PCR product is not inserted within the vector, the functional LacZα will complement with its cellular counterpart and functional β-galactosidase enzyme will be generated that will digest X-gal. As a result blue color colonies will be generated. So, white colonies were picked as potential candidates to have the putative promoter regions. Two blue colonies were also picked as the control. Colony PCR was performed with these colonies. Colonies were added directly to a 19µl PCR reaction mix containing M13 vector primers (Table 3.3), standard *Taq* polymerase buffer and standard *Taq* polymerase, and simultaneously streaked onto a replicate plate (LB-agar with kanamycin, incubated at 37°C). An additional five minutes of denaturing (95°C) was added prior to the beginning of the PCR cycling to facilitate bacterial cell lysis. The PCR products were visualized on an ethidium bromide stained 1% agarose gel and true positive candidates were identified. Two true positive candidates from the replicate plate

of each hominoid were inoculated into a 5ml overnight culture of LB-kanamycin media. A freezer stock was made for TOPO clones of all hominoids containing 750µl culture and 250µl 60% glycerol. The stock was stored at -80°C. The remaining culture was used for isolating the plasmid DNAs containing the putative promoter region, using QIAprep spin miniprep kit (Qiagen). The isolated plasmid DNAs were then subjected to a second sequence verification. After sequence verification, the TOPO vector was restriction digested and the inserted promoter sequence, containing the restriction site overhang, was isolated. The restriction enzymes used in this study were *Acc*65I, and *Hin*dIII. The digested products were gel purified using a 1% crystal violet gel with the Wizard® SV Gel and PCR Clean-Up System (Promega) and the plasmid DNA concentration was determined using a Qubit ® fluorometer (Life Technologies).

The final step of generating the reporter construct is inserting the digested and purified products (mentioned-above) into luciferase containing reporter vectors. pGL4.10 vector was used for this study (Fig. 3.7). It is a promoter-less vector, containing the firefly luciferase enzyme coding sequence.

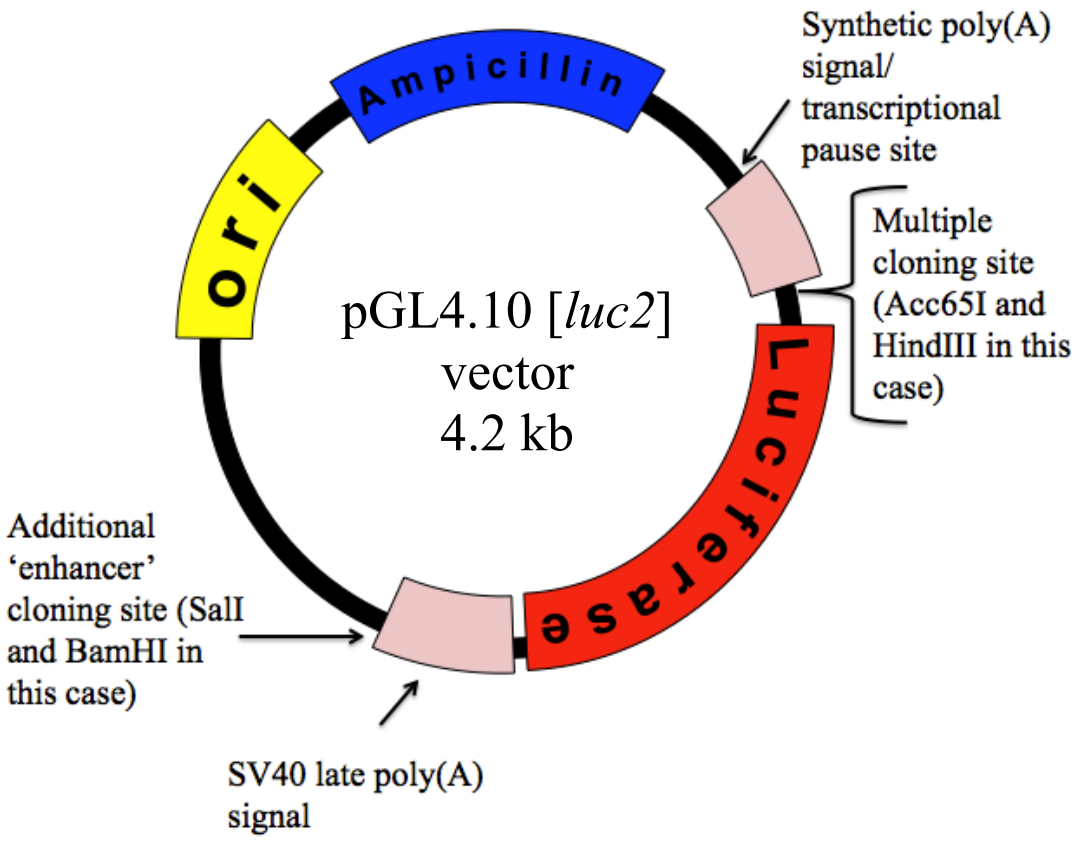


Figure 3.7: pGL4.10 vector (Promega) showing multiple cloning sites, additional cloning site, luciferase gene and antibiotic resistant gene

Purified digested products were ligated into pGL4.10 vector using T4 DNA ligase (Promega). I used 1: 1 molar ratio of vector to DNA product for this ligation reaction. The reaction mix was incubated overnight (~16 hours) at 16°C in a water bath. Ligation mixtures were transformed into chemically competent T1 *E.coli*, plated on LB-agar carbenicillin (ampicillin substitute) plates, and incubated at 37°C overnight. Similar (as performed during TOPO cloning) colony PCR-based screening process was implemented to identify candidate colonies and replicate plates (LB agar with ampicillin) were created. True positive candidates were identified by colony PCR screening and were transferred

from the replicate plate into a 5ml starter culture (LB broth with ampicillin at 37°C for 6 hours). After 6 hours 500µl of the starter culture was used to inoculate a 50ml LB-ampicillin overnight cultures (37°C, ~16 hours). Next day freezer stocks were prepared from the overnight cultures (750µl culture + 250µl 60% glycerol, stored at -80°C). The rest of the cultures (~49ml for each) were used for isolating plasmid DNA using Qiagen Plasmid Midi Kit (Qiagen) and resulting plasmid DNA concentration was determined using a Qubit ® fluorometer (Life Technologies). The isolated plasmid DNAs was then subjected to vector end sequencing to verify the presence of the product, using pGL4 vector primers (Table 3.3).

Table 3.3: Primers used in screening of TOPO or pGL4 constructs

Primer name	Purpose	Sequence (5'-3')
M13_For(-21)mod	TOPO screen	GTTGTAAAACGACGGCCAGT
M13_Rev_mod	TOPO screen	CACAGGAAACAGCTATGACC
pGL4_92_R1	pGL4.10 screen	TTACCAACAGTACCGGATTG
pGL4_4223_F1	pGL4.10 screen	AGGTGCCAGAACATTTCTCT

3.2.4 Construction of reporter vectors containing putative promoters and additional cis-regulatory region

A 2.2kb region in between *CRTAC1* exons 11 and 12 with a strong H3K27Ac mark and DNase hypersensitive cluster was identified as a possible additional cis-regulatory region. This region shows DNase hypersensitivity signal in LNCaP cells as well (see Introduction). This region from human and chimp was inserted inside human and chimp pGL4.10 constructs respectively at the additional 'enhancer' cloning site (Fig. 3.7), already containing the putative promoter regions (Fig. 3.8). The same method was

used to clone human and chimp additional cis-regulatory element into pGL4.10 vectors as described in section 3.2.3. The primers used for cloning and sequencing the additional cis-regulatory region is shown in Table 3.4.

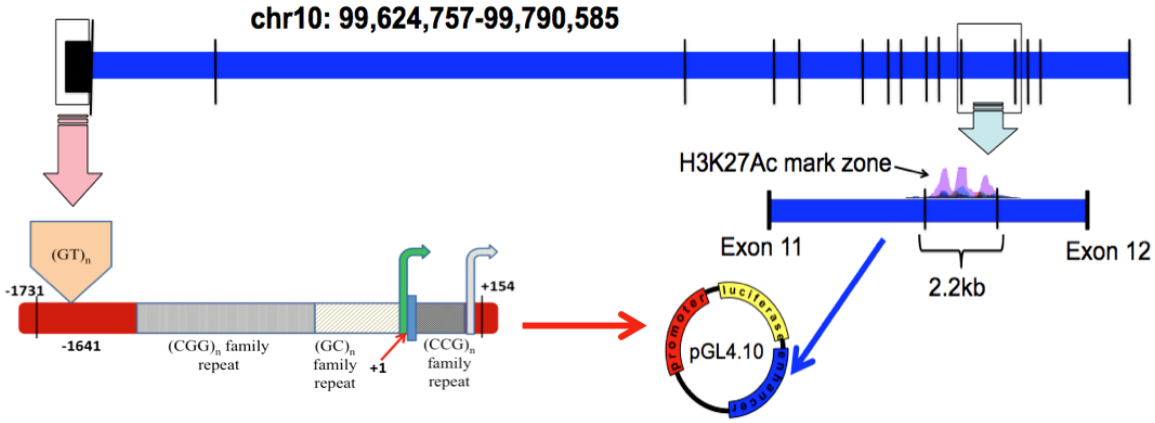


Figure 3.8: Cartoon showing the generation of 'promoter + additional *cis*-regulatory element' constructs in human and chimpanzee.

Table 3.4 List of primers used for cloning and sequencing the additional cis-regulatory region

Primer Name	**Purpose**	**Sequence (5'-3')**
Enhancer_BamHI_Forwar[1]	PCR and cloning	GTATTGCAGG**GGATCC**AAG AGTTTGTG
Enhancer_SalI_Reverse[1]	PCR and cloning	GATTTCCCAG**GTCGAC**CCTA TGTCCAA
CRTAC1_EnSeq_1	Sequencing	GTGTCAGAATGTGTATCAGG
CRTAC1_EnSeq_2	Sequencing	TCCTGCCAGGCCTGTGTATA G
CRTAC1_EnSeq_3	Sequencing	ATGAGCATCAGCCGGCTCGG
CRTAC1_EnSeq_4	Sequencing	AGGGCCATAAGCAGAAATGC

		T
CRTAC1_EnSeq_5	Sequencing	ATGTCCGTGTACTCACACATG
CRTAC1_EnSeq_6	Sequencing	ACAATCTAACTGTCCTTCCAGA
CRTAC1_ENHANCER_7	Sequencing	TGAGTCACCTCTGGCAGCTT
CRTAC1_ENHANCER_8	Sequencing	AGTGGAGCTGGCGGAGGCAA
En_Seq9	Sequencing	TGAAAGGCGGTGGCATGTGT
En_Seq10	Sequencing	AGGCAGCCACCCCAACCATT
En_Seq11	Sequencing	GCTGGCAGCGGGCTGAGGCA
CRTAC1_EnSeq12	Sequencing	GTGGACCCTGCCTTGCTCAG
CRTAC1_EnSeq13	Sequencing	AGCATTTCTGCTTATGGCCCT

[1]The location of the restriction site is shown in bold

3.2.5 Luciferase expression assays

3.2.5.1 Maintaining and subculturing of LNCaP cells

A human prostate cell line (LNCaP clone FGC, ATCC® CRL-1740™) was used in this study. One week before the experiment, a vial containing frozen LNCaP cells was removed from a liquid nitrogen incubator and rapidly thawed in a 37°C water bath. Then the contents of the vial was transferred into a tissue culture flask containing RPMI-1640 medium (ATCC) supplemented with 5% fetal bovine serum (FBS), and antibiotic (penicillin and streptomycin, (P/S)) and kept in a humidified, 37°C, 5% carbon dioxide incubator.

3.2.5.2 Transfection of luciferase constructs

Once the cultures reach >75% confluency, cells were trypsinized, counted using a hemocytometer, and plated in a 12-well cell culture plate at a density of 200,000 cells per well in 1ml of complete growth media (RPMI+FBS+P/S). Twenty-four hours after plating, cells were transfected with the luciferase constructs containing only promoters or both promoter and additional cis-regulatory region using 3µl Fugene® HD transfection reagent (Promega) per 1µg DNA. This transfection reagent is a cationic lipid reagent, which forms complex with DNA. Because of its positive charge and lipid makeup, it forms micelles around the DNA and neutralizes its charge. This process helps the passage of DNAs inside the mammalian cell through the cell membrane.

All transfections were performed in triplicate. For some experiments, 24 hours post transfection, cells were stimulated with 10 nM synthetic androgen (R1881) (dissolved in 100% ethanol) and the rest were supplemented with same quantity100% ethanol (used as the 'vehicle control'). One of the experimental designs is shown in Fig. 3.9. Fourty-eight hours post transfection, the cells were harvested, lysed, and the luciferase activity was quantified using the Luciferase Assay System (Promega) using a luminometer (Turner Designs). During the transfection standardization stage (see Results) various different parameters (Fugene amount, time duration between transfection and cell lysis, whether to use of *Renilla* luciferase as control) were tested. When *Renilla* was used as control Dual-Luciferase® Reporter Assay System (DLR™ Assay, Promega) was used instead of the Luciferase Assay System. Statistical analyses were performed in GraphPad Prism v6 statistical software. Additional statistical analyses were performed in in R v3.0.2 statistical software package.

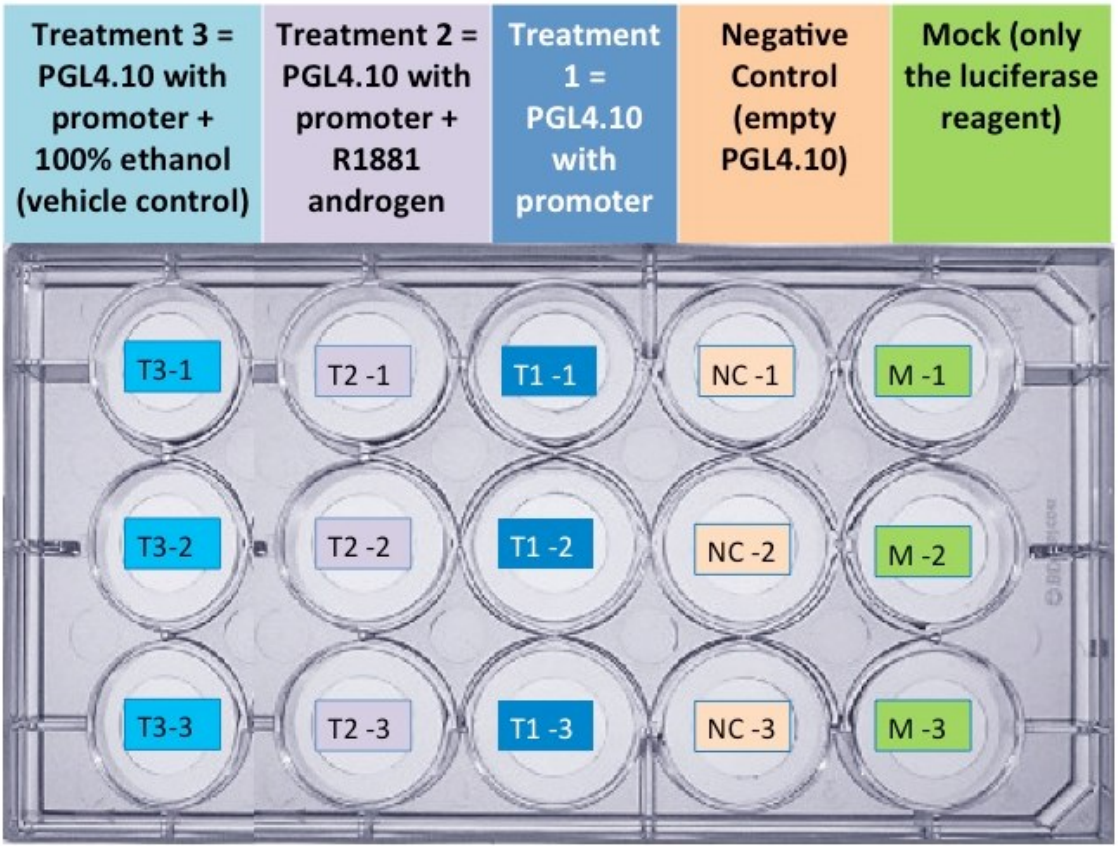


Figure 3.9: An example of the experiment designs used in transfection experiments

3.2.5.3 Bradford Assay

To determine whether I was plating equal number of cells in each well on a given day, I quantified the amount of protein in each well of a plate. The amount of protein is thought to be proportional to the total number of cells in each well. The Quick Start™ Bradford Protein Assay (BioRad) was used in accordance with manufacturer's protocol for the 1ml assay including making a standard curve with known concentrations. Readings for the assay were taken using ThermoSpectronic Genesys 10 UV spectrophotometer. This experiment was repeated for two consecutive weeks. Variation among wells was assessed with a Kruskal-Wallis test, as implemented in Graphpad v6.

3.2.6 Characterization of Gorilla putative promoter region of CRTAC1 using a Gorilla BAC library

After several unsuccessful trials to amplify the gorilla putative promoter region from genomic DNA, we decided to use gorilla Bacterial Artificial Chromosomes (BACs) to isolate the gorilla *CRTAC1* putative promoter region. To do so, we first designed two 24bp oligos for the 'promoter' region around the 'gap' area (See Introduction section 3.1.7) shown in UCSC genome browser with 8bp overlap using human genome sequence.

Gor_CRTAC1_dngap_OV (F): 5' ATTCTTGAGTTGCTTTCTGCAGAA 3'

Gor_CRTAC1_dngap_OV (R): 5' CCCTGGTTAGTCCCGCTTCTGCAG 3'

The BAC overgo hybridization method was employed using the protocol by Ross et al. (1999). The oligos were then mixed well, denatured at 80°C for 10 min, and annealed at 37°C for 10 min. The annealed oligos were then radio-labeled with radioactive dATP and dCTPs in presence of 5% Klenow enzyme (v/v), and 10% BSA (v/v). The mixture was incubated at room temperature for an hour. The 1X TE buffer was added to the labeled probes (3: 1 v/v) to stop reaction, mixed well, and applied at the center of an Illustra G-50 sephadex Column (GE Healthcare Life Sciences) for purification. The radio-labeled probes were stored at 4°C.

The radio-labeled oligos were then hybridized with a gorilla genomic BAC library arrayed on nylon membranes (Library CH255, obtained from BACPAC resources, Oakland, CA). The first step was to pre-warm 1X wash buffer and Express Hybridization solution (Clonetech) to 61°C in a water bath for 20 min. The hybridization membranes

were washed with 1X SSC (0.15M NaCl and 0.015M Sodium Citrate), 0.1% SDS (v/v) at 90°C for 30 min with gentle agitation. Then the membranes were placed in a container with 2X SSC (0.3M NaCl and 0.03M Sodium Citrate), 0.1% SDS (v/v) at room temperature and washed. The membranes were then sandwiched with Flow Mesh (Diversified Biotech) and rolled up into a hybridization bottle. Pre-warmed Express Hybridization solution was then added into the hybridization bottle. The bottles were pre-hybridized at 60.7°C for 30 min. The radio-labeled probes were then denatured at 90°C for 5 min and placed on ice. The denatured probes were then added to pre-warmed 5ml Express Hybridization buffer, mixed well, and added to the hybridization bottle, mixed well, and incubated overnight at 60.7°C. On the next morning, the hybridization solution was emptied from the bottle, and the membranes and the Flow Mesh were washed 3 times with 1X SSC and 0.1% SDS. The membranes in the tube were then washed with 1X SSC and 0.1% SDS at 61°C in the hybridization oven for 30 min and this step was repeated. The membranes were removed from the tube and the Flow Mesh was discarded. The membranes were then wrapped in face up position with plastic wraps, and attached to a paper with tape. The papers were placed inside a closed cassette with membrane number, date, and time.

In the dark room, X-ray films were placed on top of the papers, inside the cassette, and the film and the paper were flipped together. The cassettes were then placed in -80°C freezer. Films were then developed after 96 hours, and scored for positive hits in the genomic library. We found six hits in the entire BAC library, two of which were ordered for further analysis (25B23 and 36G6).

BAC clones CH25B23 and CH36G6 were grown up and purified. The clones were end-sequenced using T7 and SP6 primers to map to the gorilla genome assembly (gorGor3).

3.2.7 Characterization of the coding region of CRTAC1 from four hominoid primates

3.2.7.1 Polymerase Chain Reaction (PCR)

PCR was carried out in 20 µl reaction, containing 1X ThermoPol B9004S PCR buffer (New England Biolab Inc) containing Mg^{2+}, 0.5U *Taq* DNA Polymerase, 250µM of dNTP (dATP, dCTP, dGTP, dTTP), and 0.25µM of each primer. Thermal cycling started with 5 min denaturation at 94°C, followed by 36 cycles of denaturation (94°C, 15 sec), annealing (55°C, 15 sec), and primer extension (72°C, 2 min); PCR concluded with a terminal extension at 72°C (7 min), and final holding at 4°C. The primers used in PCR are listed in Table 3.5.

Table 3.5: The PCR primers used to sequence the coding region of *CRTAC1*

Primer name	Sequence (5'-3')
CRTAC1_exon1F1	AGTCGAAATCTCGCCATCAG
CRTAC1_exon1R1	TCTCAGCCTGGCTGACTG
CRTAC1_exon2F1	TAGGCACTGTCAAGCTCTC
CRTAC1_exon2R1	GAAAAGCCTGTGGATCAAAC
CRTAC1_exon3F1	GCAAAATCTGATTCAGGGAC
CRTAC1_exon3R1	GACAATGACAGTTGCCTGAG
CRTAC1_exon4F1	CCATGAGACATACCCAGAG
CRTAC1_exon4R1	TGTCAGGGGACGTCTAC
CRTAC1_exon5F1	ATGTGAAAGAGCTATGGTC
CRTAC1_exon5R1	GGAATGGAGCCTTCATCTC
CRTAC1_exon6F1	TTACCTTCCACACAGATTCA
CRTAC1_exon6R1	GTGTGGTGCAGACGATGA
CRTAC1_exon7F1	AGTGTTTGTGGAGTGTGCA
CRTAC1_exon7R1	CCTGCAGTGGTGCTGTAG
CRTAC1_exon8F1	CCCCACACTCCATAGAG

CRTAC1_exon8R1	ACCCTCCCCTTCTGATTC
CRTAC1_exon9F1	AGAGTGGCTCCGTGGGCA
CRTAC1_exon9R1	AGGAGTGGCTCTGCTGTG
CRTAC1_exon10F1	TATCATGAGGCTGCTGTTAG
CRTAC1_exon10R1	GGCCCTTCCTGAGCTTC
CRTAC1_exon11F1	ACCCACCATGCTGATGCC
CRTAC1_exon11R1	CAGGCTCATCTCAGAGTAG
CRTAC1_exon12F1	ATGCTGATGCTTCCGCTG
CRTAC1_exon12R1	ATCAGTATGAAGCCTTCGC
CRTAC1_exon13F1	GCCCAGAGCTCAGAGCA
CRTAC1_exon13R1	ACCCTGGTGAGTCATGT
CRTAC1_exon14F1	TTGCTGCCCCACACCTTC
CRTAC1_exon14R1	CTGGCCCTTCAGGTGATG
CRTAC1_exon15F1	CCCAAAGAATGACTCAGAAG
CRTAC1_exon15R1	TAGTGTGATCTGGGTGTG
Orang CRTAC1 exon6F	CACTCACTGTGGGTCGATG
Orang CRTAC1 exon6R	GCCAGAAGAGACCATCCTG
Orang CRTAC1 exon8F	CACACTCCATAGAGGAGAG
Orang CRTAC1 exon8R	AGCTGTCAAGGGTGAAGAG
Orang CRTAC1 exon9F	ATGCTCAGGGGACAGAATG
Orang CRTAC1 exon9R	CACCCCAGTGTATGAACAG
Orang CRTAC1 exon12F	CTGATGCTTCTGCTGAGAG
Orang CRTAC1 exon12R	CGCTTTTCTGGCTATGAGG
Gorilla CRTAC1 exon7F	TTTCCCCCAGTCTCCCTC
Gorilla CRTAC1 exon7R	TGAGACGGAGTCTCGCTC
Gorilla CRTAC1 exon8F	TGTACTGCCTCAAGGGATG
Gorilla CRTAC1 exon8R	GTATCTTGTGGTGCTTGGG
Gorilla CRTAC1 exon12F	CCCACTTCAAGTGCTCAAC
Gorilla CRTAC1 exon12R	CCTTGGTGGATTTCCTCTC

The samples that could not be amplified by the standard PCR were amplified by adding 0.25M Betaine, 5% DMSO (v/v), with the same thermal cycling as above.

3.2.7.2 PCR Purification and DNA sequencing

PCR products were purified using Wizard® SV Gel and PCR Clean-Up System (Promega Corporation) following the manufacturer's protocol.

The purified PCR products were sequenced using BigDye cycle sequencing chemistry on capillary ABI-3100 auto sequencer.

3.2.7.3 Interspecific analysis of protein coding region of CRTAC1

The consensus sequences of each exon for every species were generated using SeqMan software package (DNASTAR Inc., Madison, WI, USA), by aligning the sequences obtained from the forward and reverse primers. The consensus sequences from each exon were subsequently joined together to generate the complete virtual cDNA for each species. The cDNA sequences from different species were aligned using ClustalW. The pair-wise ω (d_N/d_S) was calculated using MEGA 5.2 (Tamura et al. 2007).

The maximum likelihood estimates of ω were calculated using PAML4 (Yang 2007) considering uniform, branch, site and branch–site models (see Introduction). The control file (codeml.ctl) of 'codeml' in PAML4 was modified according to the model being tested. The 'model' parameter in codeml.ctl designates whether ω is the same for all lineages or each lineage has its own ω. 'NSsites' parameter designates parameter set variation among sites. For branch models (uniform vs. free ratio) the 'model' parameter in the codeml.ctl file was changed either to 0 (for uniform ratio, Model M0) or to 1 (for free ratio, model M1), but the 'NSsites' parameter was kept at 0 (sites are not variable) in both cases. For site models, the 'model' parameter was fixed at 0 (ω does not vary across branches) and the 'NSsites' parameter was changed to either 1 (codons evolving neutrally, Model 1a) or 2 (codons are under positive selection, Model 2a). The 'NSsites' parameter was changed to 3 to suggest discrete selective pressure among sites (Model 3). Beta distributed neutral model was generated by keeping the 'model' parameter fixed at 0 and changing 'NSsites' parameter to 7 (Model 7). Beta distributed positive selection models were generated by changing the 'NSsites' parameter to 8 and either fixing ω at 1 (Model 8) or considering $\omega > 1$ (Model 8a). Finally, maximum likelihood estimates of ω

were also calculated from the branch-site model (allowing ω to vary among branches as well as codons) by setting the 'model' parameter to 2 (multiple ωs for all branches) and 'NSsites' parameter to 2 (codons are under positive selection) and considering $\omega > 1$ (Model A).

Likelihood ratio test (LRT) was performed to find out the best fitting model. The degrees of freedom were calculated by subtracting the number of parameters of the simpler model from the more complex model. Statistical significance of the difference in likelihoods between two models was assessed by comparing twice the difference ($2\Delta l$) to a chi-square distribution.

3.2.7.4 Intraspecific analysis of protein coding region of CRTAC1

I analyzed multiple genome data from Great Ape Genome Project (Prado-Martinez et al. 2013, http: //biologiaevolutiva.org/greatape/). I downloaded the variant calling files (VCF) from the database and converted them into usable formats using UNIX command lines. The modified files were used to analyze the Single Nucleotide Polymorphisms (SNPs) existing in the protein-coding region of *CRTAC1* within five hominids: human, chimpanzee, bonobo, gorilla, and orangutan. Population genetic analyses were performed using Online Encyclopedia for Genetic Epidemiology studies (OEGE http: //www.oege.org/software/hwe-mr-calc.shtml) web server. Additional population genetic analysis including Analysis of Molecular Variance (AMOVA), Pair-wise AMOVA, Neutrality tests (Tajima's D) were performed using Arlequin v3.5 population genetic software.

Inbreeding-coefficient (F) was calculated using the formula:

1- (observed heterozygosity/expected heterozygosity)

3.3 Results

3.3.1 Standardization and optimization of Polymerase Chain Reaction for amplifying CRTAC1 putative promoter region

As mentioned before, the highly GC rich areas (> 75%) in the putative promoter region of *CRTAC1* make it very difficult to amplify. I tried 5% DMSO, 5% betaine, or a combination of both, which have shown to be efficient for amplifying the GC rich regions. Also the PCR conditions like the annealing temperature and extension time had also been changed to amplify the putative promoter region of ~2kb. I also employed Hot start, and Touch down PCR. But none worked (Fig. 3.10). I finally, successfully adapted and modified "SAFE (satisfactory, adaptable, fast, efficient) PCR" (Fig. 3.11) as described in Wei *et al.* 2010, for the amplification of *CRTAC1* putative promoter region.

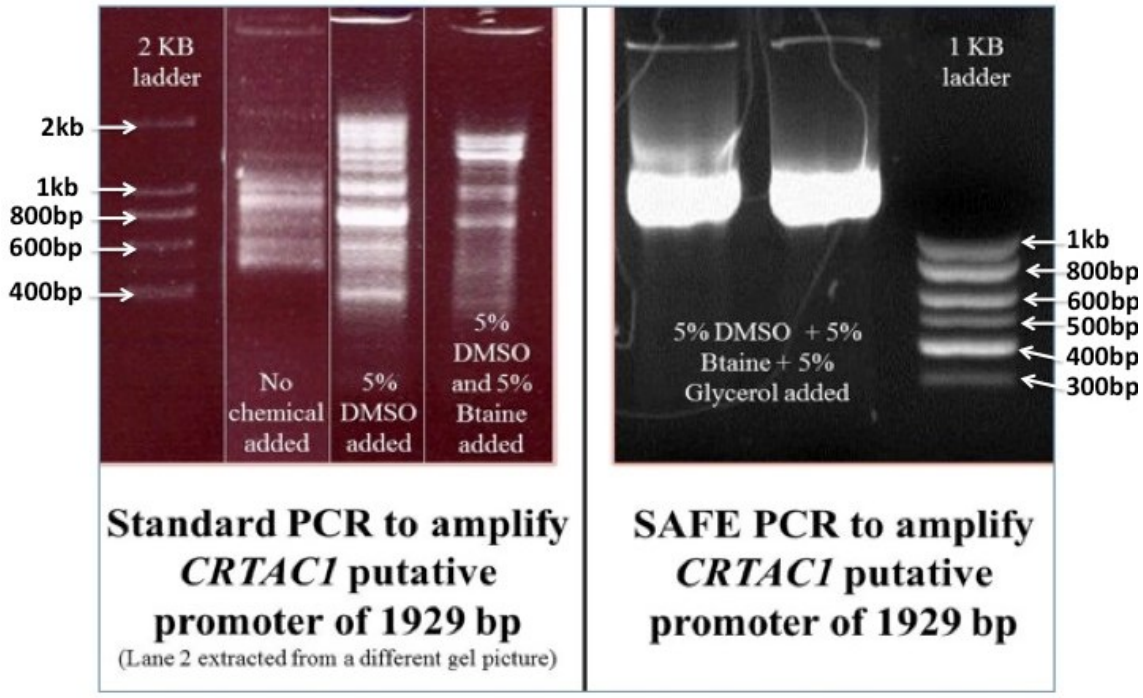


Figure 3.10: Comparison of Standard and SAFE PCR techniques for the amplification of *CRTAC1* putative promoter region

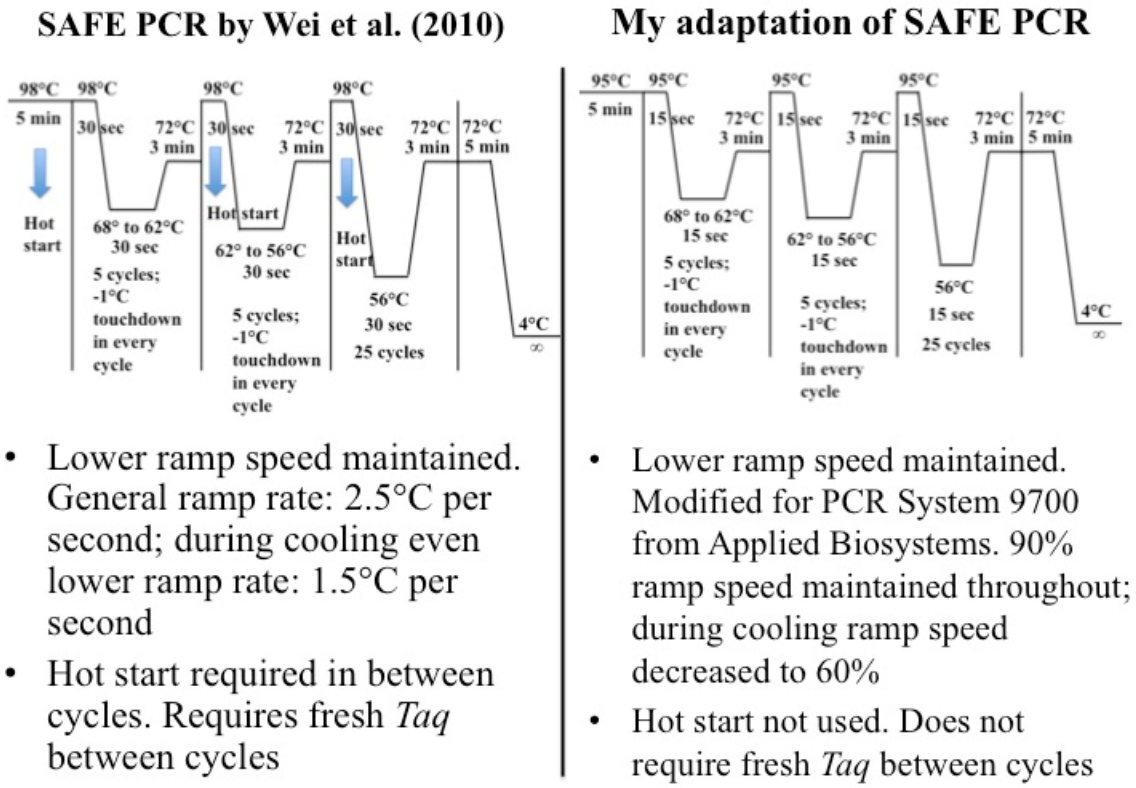


Figure 3.11: Comparison of SAFE PCR and my adaptation of the technique. The initial denaturation temperature was set at 95°C in my adaptation of SAFE PCR

The SAFE PCR protocol required the addition of 5% DMSO, 5% betaine, and 5% glycerol to the PCR mix. A comparison of SAFE PCR by Wei et al. and my adaptation of the method are shown in Fig. 3.11.

Following the amplification of human *CRTAC1* promoter, I successfully used the same SAFE PCR technique to amplify the putative promoter regions from all hominoid species.

3.3.2. Sequencing analysis of the putative promoter region of *CRTAC1*

The entire putative promoter regions from the hominoids were sequenced twice: once after PCR amplification and then after TOPO cloning. The human promoter

sequence has three SNPs (Fig. 3.12) in the putative promoter region: rs514554 (A>C transversion), rs544022 (G>A transition), and rs61873668 (C>G transversion), when compared to the human reference sequence (hg19) in the UCSC genome browser (http://www.genome.ucsc.edu). In rs514554 SNP, the C allele is observed in 24.27% humans. Chimpanzee and Neanderthal both are fixed with C. In rs544022 SNP, the A allele is observed in 24.18% humans. Chimpanzees are fixed with G and Neanderthals had A here. In rs61873668 SNP, G allele is observed in 12% of humans. Both Chimpanzee and Neanderthals are fixed with C. According to RegulomeDB (http://www.regulomedb.org/) rs514554 may affect binding of some transcription factors including GATA2. rs544022 and rs61873668 probably do not affect binding of transcription factors. The human in this study also has three 'GT' microsatellite repeats less compared to the reference human sequence (hg19).

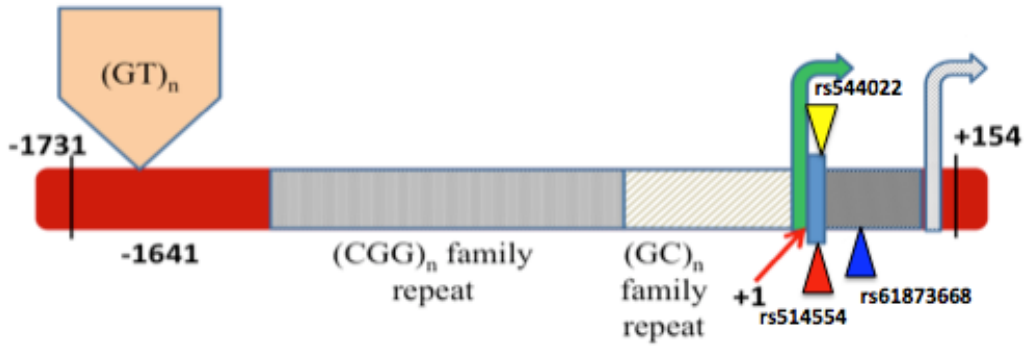


Figure 3.12: The location of human SNPS at the putative promoter region of *CRTAC1*

The chimpanzee sample used in this study has three nucleotide differences compared to the chimpanzee reference sequence (panTro4) along with one 1bp insertion and one 1bp deletion. All of these nucleotide differences correspond to the 18 SNPs present in this area in chimpanzee (Great Ape Genome Project http:

//biologiaevolutiva.org/greatape/). Orangutan has three nucleotide differences compared to the orangutan reference sequence (ponAbe2), which also correspond to the 21 SNPs present in this area.

As mentioned in section 3.2.6, the gorilla promoter could not be amplified from the genomic DNA. After several trial and errors, the entire gorilla putative promoter was amplified and sequenced from BAC CH255-36G6. 36G6 starts ~90kb upstream of the TSS and ends in intron 8 (Gorilla Chr10: 111,279,126-111,507,783). The gorilla putative promoter was not used for making expression constructs. There are five nucleotide differences in gorilla compared to the gorilla reference sequence (gorGor3). According to the Great Ape Genome Project (GAGP) database (Prado-Martinez et al. 2013), gorilla has maximum number of SNPs (47) in the putative promoter region. All nucleotide differences observed in the gorilla under study correspond to these SNPs. As mentioned before, the UCSC genome browser has gaps in the putative promoter region of both chimpanzee and gorilla. I found that the gaps are not real deletions, but rather missing data. I amplified and sequenced the entire putative promoters of both species and have sequenced the entire 'gap' in both species (Fig. 3.13).

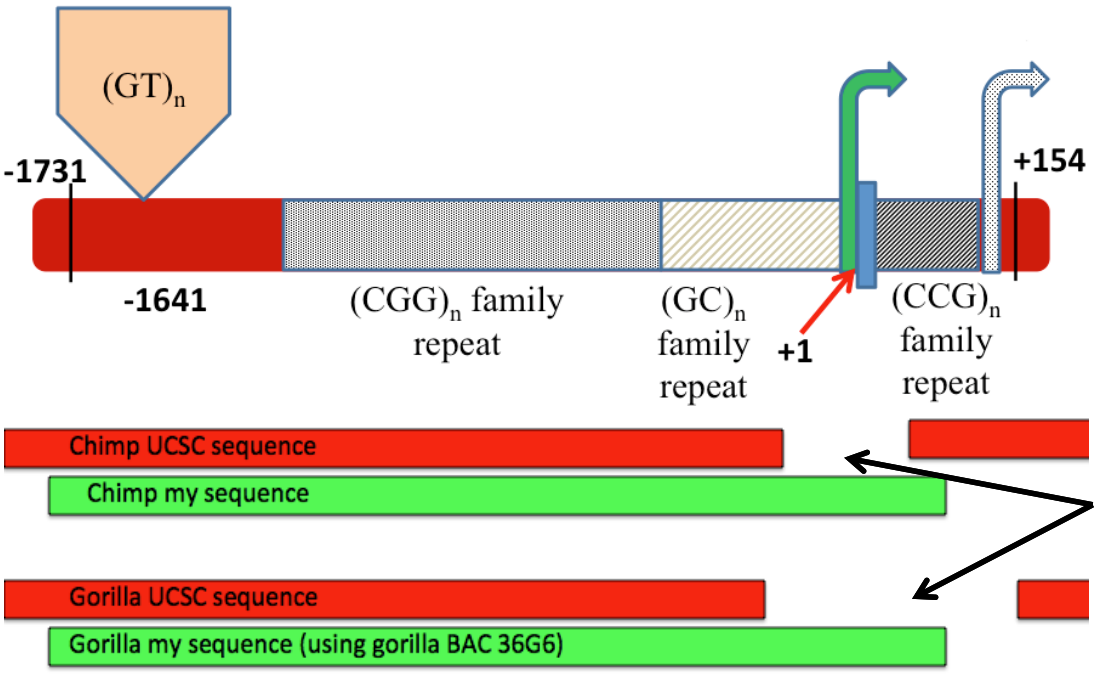


Figure 3.13: Cartoon showing the location of the gaps in UCSC browser. I could sequence through the gaps in both chimpanzee and gorilla

Human putative promoter has sixteen nucleotide differences (0.84%), and a two bp (AG) deletion compared to chimpanzee at -787 within $(CGG)_n$ repeats. This deletion disrupts the consensus sequence of CACCC-binding protein, which aids in transcriptional repression of several genes (vanVliet et al. 2000, Funnell et al. 2012). Out of the 16 nucleotide differences seven were found to be human specific and nine were chimpanzee specific, when compared to a multispecies alignment for the orthologous region in the genome (See Appendix 2.2.1). Chimpanzee and bonobo share two 1bp deletions in the GC rich area. The two 1bp deletions in the chimpanzee and bonobo putative promoter regions results in the loss of potential binding sites for Ncx, FACB, STAT5A, and Elk-1 transcription factors. Ncx and STAT5A can potentially act as transcriptional activators (Shimizu et al. 2000). Moreover, six out of 16 nucleotide differences between human and

Pan species putative promoter region concentrate within ~150bp downstream of putative TATA box region. The nucleotide differences in this region can potentially affect the binding of MF3, Msx-1, and MZF-1 in *Pan* species (Fig. 3.14a). There are 19 nucleotide differences between human and bonobo, six of which are uniquely gained by bonobo (Fig. 3.14a).

Orangutan putative promoter looked most different among all hominoids. Orangutan promoter has a 'CCCCACCACCAC' insertion in the putative promoter region (Fig. 3.14b). 'CACCACCAC' is the consensus sequence for binding of the transcription factor Msx-1. Orangutan also has a 'TAGGA' insertion (Fig. 3.14b). 'TAGGA' matches the consensus sequence of Ebf1 binding site. Orangutan also has a 13bp deletion (Fig. 3.14b) in the putative promoter region. In other hominoids this region has the sequence: 'CGCCCGCCCTCGC'. This sequence is a potential Activating enhancer Protein-2 alpha (AP-2 alpha) binding sequence (GCCXXXGGC).

Gorilla has a 12bp deletion (Fig. 3.14b) in the putative promoter region. In other Hominoids this area has the sequence: 'TCGCCGCCGCCC'. Transcription factors Sp1 (Raiber et al. 2012) and ZF5 (Orlov et al. 2006) potentially bind to this sequence. Gorilla also has a five bp insertion in the putative promoter region ('GGGCC') (Fig. 3.14b).

Human (Fig. 3.14b) has a two bp deletion in the putative promoter region. This deletion disrupts the consensus sequence of CACCC-binding protein, which aids in transcriptional repression of several genes (Vliet et al. 2000, Funnell et al. 2012). The entire alignment is shown in Appendix 2.2.1.

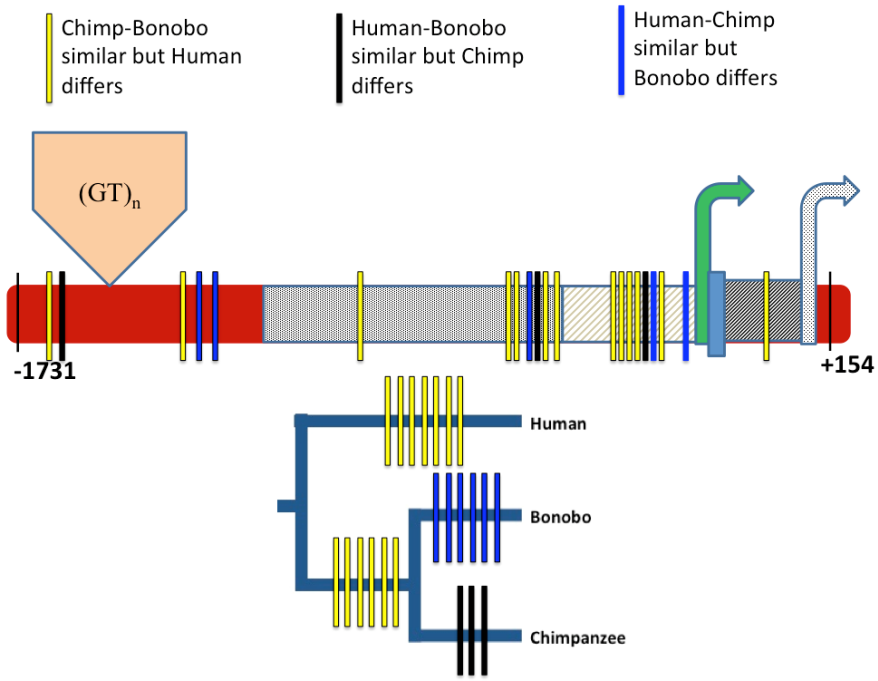


Figure 3.14a: Location of various nucleotide differences among human, chimpanzee and bonobo in the putative promoter region of *CRTAC1*, including the two 1bp deletions shared uniquely by chimpanzee and bonobo

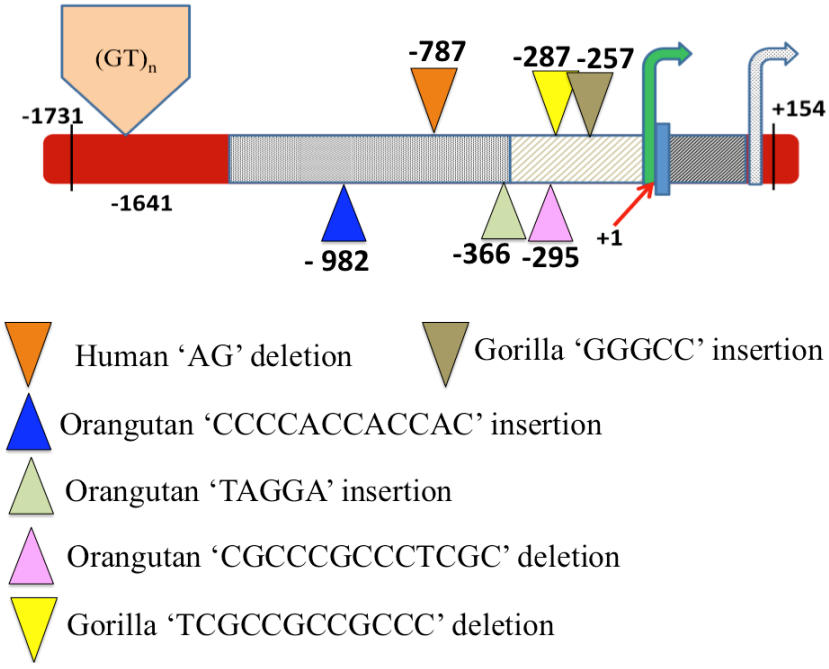


Figure 3.14b: Location of various insertion-deletions in the putative promoter region of *CRTAC1*

3.3.3 Genotyping 'GT' microsatellite repeat in *CRTAC1* putative promoter region

The Peak Scanner output files received after preliminary analysis contain allele sizes in decimals. The allele sizes were then manually 'binned' to nearest whole number (Table 3.6).

Table 3.6: The name and IDs of the hominoids used for 'GT' genotyping

Species	ID/Names	Alleles (Bin Version)	GT repeat no.
Human	NA15283	211 and 215	$(GT)_{23}$ and $(GT)_{25}$
	NA15047	218 and 228	$(GT)_{26}$ and $(GT)_{31}$
	NA15504	220 and 220	$(GT)_{27}$ and $(GT)_{27}$
	NA15230	206 and 222	$(GT)_{20}$ and $(GT)_{28}$
	NA15242	220 and 228	$(GT)_{27}$ and $(GT)_{31}$
	NA15216	220 and 220	$(GT)_{27}$ and $(GT)_{27}$
	NA15221	206 and 220	$(GT)_{20}$ and $(GT)_{27}$
	NA15245	220 and 220	$(GT)_{27}$ and $(GT)_{27}$
	NA15215	206 and 222	$(GT)_{20}$ and $(GT)_{28}$
	NA15341	206 and 222	$(GT)_{20}$ and $(GT)_{28}$
Chimpanzee	PR496	206 and 213	$(GT)_{20}$ and $(GT)_{24}$
	Pts Kobi	206 and 213	$(GT)_{20}$ and $(GT)_{24}$
	Pts Harriet	213 and 213	$(GT)_{24}$ and $(GT)_{24}$
	Ptv Lottie	206 and 220	$(GT)_{20}$ and $(GT)_{27}$
	Ptv Lowie	196 and 220	$(GT)_{15}$ and $(GT)_{27}$
	Ptv Colin	206 and 220	$(GT)_{20}$ and $(GT)_{27}$
	Ptt Dodo	206 and 206	$(GT)_{20}$ and $(GT)_{20}$
	Ptt Cheetah	206 and 213	$(GT)_{20}$ and $(GT)_{24}$
	Ptt Julie	206 and 222	$(GT)_{20}$ and $(GT)_{28}$
	Ptt Noemie	206 and 220	$(GT)_{20}$ and $(GT)_{27}$
Bonobo	Lomoko	193 and 211	$(GT)_{13}$ and $(GT)_{23}$
	Lenore	196 and 211	$(GT)_{15}$ and $(GT)_{23}$
	Matata	189 and 206	$(GT)_{11}$ and $(GT)_{20}$
	Kevin	211 and 211	$(GT)_{23}$ and $(GT)_{23}$
	Lody	206 and 215	$(GT)_{20}$ and $(GT)_{25}$
	Maringa	196 and 211	$(GT)_{15}$ and $(GT)_{23}$
	Bosonjo	196 and 211	$(GT)_{15}$ and $(GT)_{23}$
	PR261	196 and 211	$(GT)_{15}$ and $(GT)_{23}$
Gorilla	Frika	206 and 222	$(GT)_{20}$ and $(GT)_{28}$
	H	206 and 222	$(GT)_{20}$ and $(GT)_{28}$
	G	206 and 222	$(GT)_{20}$ and $(GT)_{28}$
	F	222 and 222	$(GT)_{28}$ and $(GT)_{28}$
	E	206 and 222	$(GT)_{20}$ and $(GT)_{28}$
	D	206 and 222	$(GT)_{20}$ and $(GT)_{28}$

	C B A Jphine	222 and 222 206 and 222 206 and 222 206 and 222	$(GT)_{28}$ and $(GT)_{28}$ $(GT)_{20}$ and $(GT)_{28}$ $(GT)_{20}$ and $(GT)_{28}$ $(GT)_{20}$ and $(GT)_{28}$

Gorilla and human possess longer size alleles compared to chimpanzee and bonobo (Fig. 3.15). Bonobos possess shorter size alleles with GT repeats ranging between 11 and 23. Chimpanzee has the most wide spread distribution of allele sizes. Interestingly all *Pan troglodytes troglodytes* individuals possess at least one allele of size 206 (GT_{20}), all *Pan troglodytes schweinfurthii* individuals possess at least one 213 (GT_{24}) allele and all *Pan troglodytes verus* individuals possess at least one allele of size 220 (GT_{27}). All gorillas possess at least one allele of size 222 (GT_{28}). Two humans possess the longest size allele of size 228 (GT_{31}) and one bonobo possess shortest size allele of size 189 (GT_{11}).

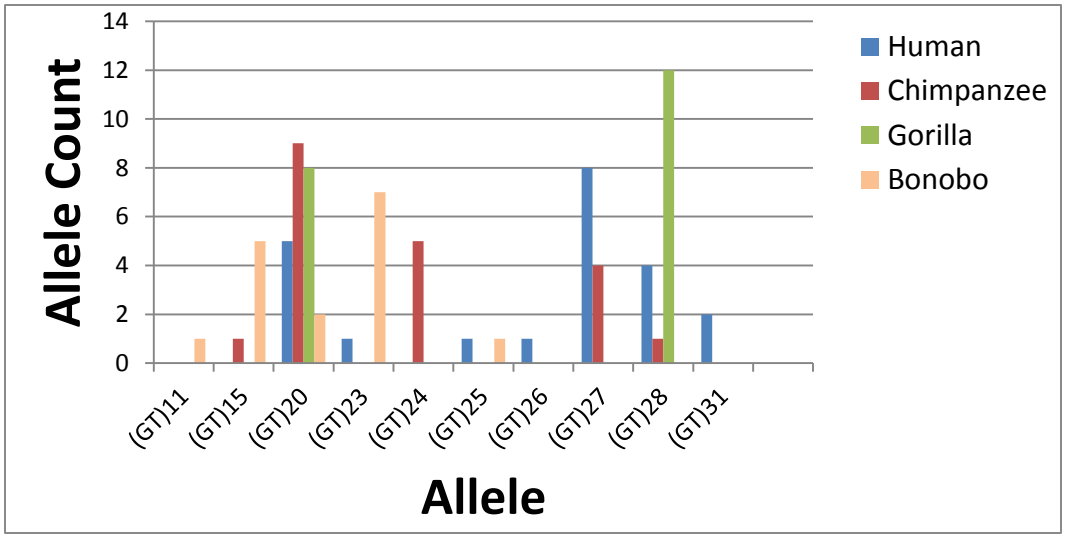


Figure 3.15: Allele Count vs. Allele Range graph showing various number of GT repeats in four hominoid species

The population genetics results are summarized in Table 3.6. All hominoids are in HWE for this locus except humans (P <0.01). Except bonobo, all hominoids have lower heterozygosity for this locus than expected. Humans have maximum variation in the alleles and gorilla has only two alleles for this locus.

Table 3.7 Population genetic analyses of the genotype data

Species[1]	**No. of Alleles**	**Observed Heterozygosity**	**Expected Heterozygosity**	**HWE P value**
Chimpanzee (2N = 20)	5	0.8	0.84	0.11
Human (2N = 22)	7	0.73	0.82	< 0.01*
Gorilla (2N = 20)	2	0.8	0.62	0.1
Bonobo (2N = 16)	5	0.87	0.78	0.1

[1]The number of chromosomes sampled is shown in parenthesis

The number of GT repeats present in the samples of different hominoids, used to make the expression constructs is shown in Fig. 3.16.

```
Human_TOPO      GAAAGTATACTCAAGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTAAATGTG
Chimp_TOPO      GAAAGTATACCCAAGGTGTGTGTGTGTGTGTGT--------GTGTGTGTAAGTAAATGTG
Bonobo_TOPO     GAAAGTATACTCAAGGTGTG------------------------TGTGTGTGTAAATGTG
Gibbon_TOPO     GAAAGTATACTCAAGGTGTGTGTGTGTGTGTG----------TGTGTGTGTGTAAaTGTG
Orang_TOPO      AAAGTATACTCAAGGT-----------------GGTGTGTGTGTGTGTGTGTAAATGTG
Gorilla_PCR     GAAAGTATATGCAGG-----GGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTAAATGTG
```

- Human GT_{28}: 3rd highest frequency
- Chimp GT_{24}: 2nd highest frequency
- Bonobo GT_{14}: 2nd highest frequency
- Gibbon GT_{23}
- Orangutan GT_{16}
- Gorilla GT_{26}

Figure 3.16: Partial alignment of the GT repeats present in *CRTAC1* putative promoter region from different hominoid clones. The allele size difference among human, chimpanzee and bonobo is even visible in 1% agarose gel

3.3.4 Transfection optimization

3.3.4.1 Time duration between transfection and cell lysis (24 hrs vs. 48 hrs) and optimum ratio of Fugene to DNA (3: 2 vs. 6: 2)

Before conducting actual transfection experiments we first optimized the various transfection parameters. First, we standardized the optimum amount of Fugene reagent that shows minimum well-to-well variation and the time duration between transfection and cell lysis. I transfected human *CRTAC1* pGL4.10 construct into LNCaP cells with two different amounts of Fugene transfection reagent: 1.5µl of Fugene for 1µg of DNA (3: 2) and 3µl of Fugene for 1µg of DNA (6: 2). Two identical plates were made. I lysed one plate of cells after 24 hours, and the other after 48 hours. A one-way ANOVA was

performed in GraphPad Prism v6 statistical software, followed by Tukey's multiple comparison as the *post hoc* analysis (Fig. 3.17). 3µl of Fugene for 1µg of DNA (6: 2) significantly increased the Relative Light Unit (RLU) of firefly luciferase compared to 1.5µl of Fugene for 1µg of DNA (3: 2) ($P < 0.05$).

There was no significant difference in RLU between 24 hours and 48 hours time duration. So, to determine which time duration (between transfection and cell lysis) is optimum for further experiments, I calculated the Coefficient of Variation (CV, Standard Deviation (σ)/Mean (µ)) for all four triplicates. I found human pGL4.10 constructs showed the least well-to-well variation for 3µl Fugene at 48 hours with CV of 0.049. 1.5µl Fugene at 24 hours showed the highest well-to-well variation (CV = 0.715), followed by 1.5µl Fugene at 48 hours (CV = 0.385) and 3µl Fugene at 24 hours (CV = 0.239).

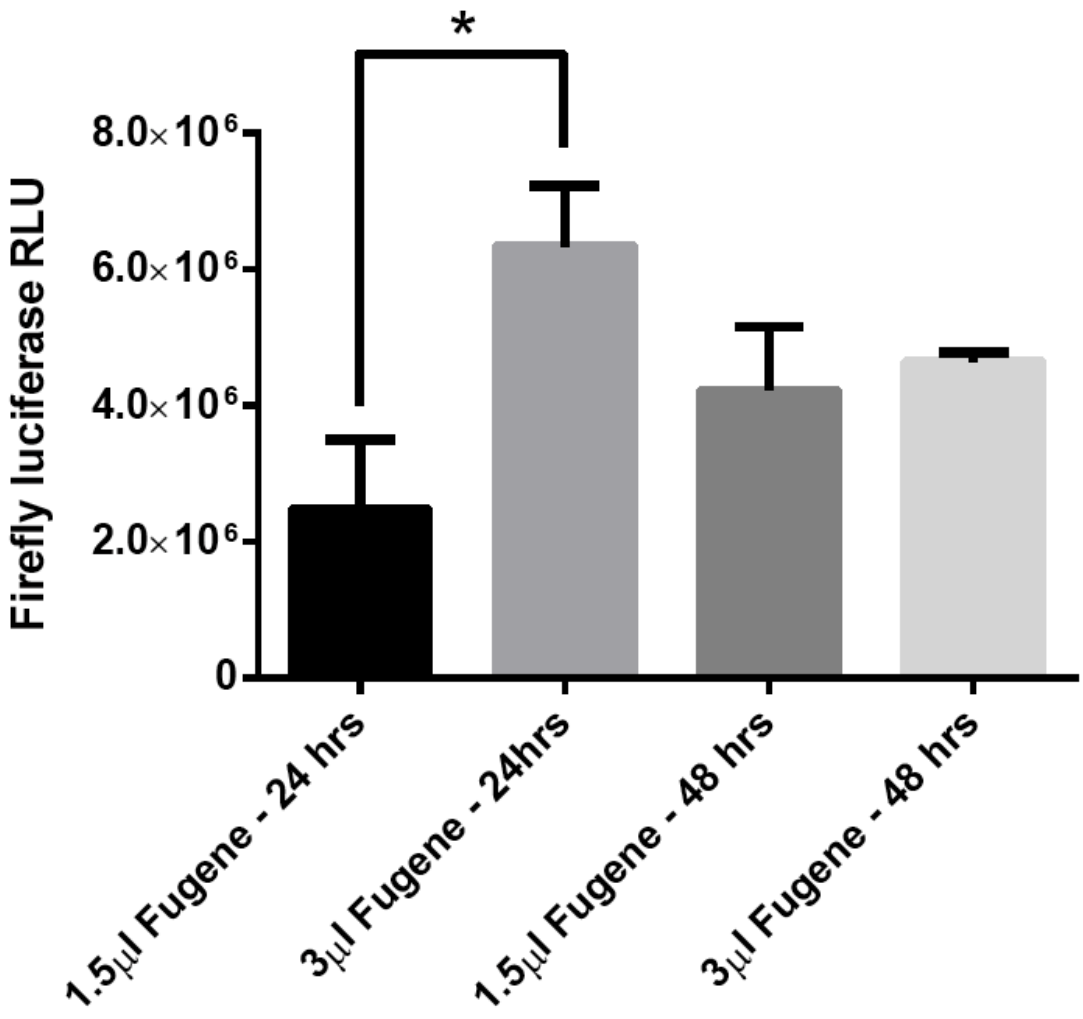


Figure 3.17: Transfection optimization experiment for time duration between transfection and cell lysis, and optimum Fugene to DNA ratio. One way ANOVA was non significant (P = 0.06) but Tukey's multiple comparison as *post hoc* test found significant difference between 3: 2 and 6: 2 Fugene to DNA ratio

3.3.4.2 Omission of Renilla luciferase vector

Our lab found that the use of *Renilla* luciferase as the control during transfection was not useful (Carnahan-Craig 2013). Two *Renilla* luciferase vectors were tried: the promoter-less pGL4.70, and one with the constitutive TK promoter (pGL4.74). Both vectors produced similar results. We found that the *Renilla* luciferase values increase or decrease with the increasing or decreasing firefly luciferase values. We found almost a

linear relationship between the increase of *Renilla* luciferase with the firefly luciferase. In other words, there appears to be cross talk between the two co-transfected vectors (Carnahan-Craig 2013). So, as a lab we decided to stop using *Renilla* as the control vector during transfection.

3.3.4.3 Confirming whether equal numbers of cells are plated in each well before transfection

Since we stopped using *Renilla* as the internal control during transfection, one of the major questions that came up was how to determine whether we are plating equal number of cells in each well on a particular day. I assessed the whole protein count of cells in each well of a plate using Bradford Assay (See Methods). I repeated this experiment for two consecutive weeks using a spectrophotometer.

There is no significant well-to-well variation for the number of cell lysate protein concentrations plated on a given day (Kruskal-Wallis test P =0.0756, Table 3.8, Fig. 3.18). If we consider the whole protein lysate concentration to be proportional to the total number of cells in each well, there is no significant well-to-well variation for the number of cells plated on a given day.

Table 3.8: Spectrophotometer readings of unknown protein lysates and the average

Species	**First reading at 595 nm**	**Second reading at 595nm**	**Third reading at 595nm**	**Average readings and lysate concentration from the standard curve**
Human_CRTAC1	0.333	0.308	0.331	0.324, 0.25mg/ml
Chimpanzee_CRTAC1	0.305	0.242	0.287	0.278, 0.22mg/ml
Bonobo_CRTAC1	0.272	0.296	0.282	0.283, 0.22mg/ml

No transfection control	0.316	0.338		0.327, 0.26mg/ml
Promoter-less pGL4.10	0.242			NA, 0.19mg/ml

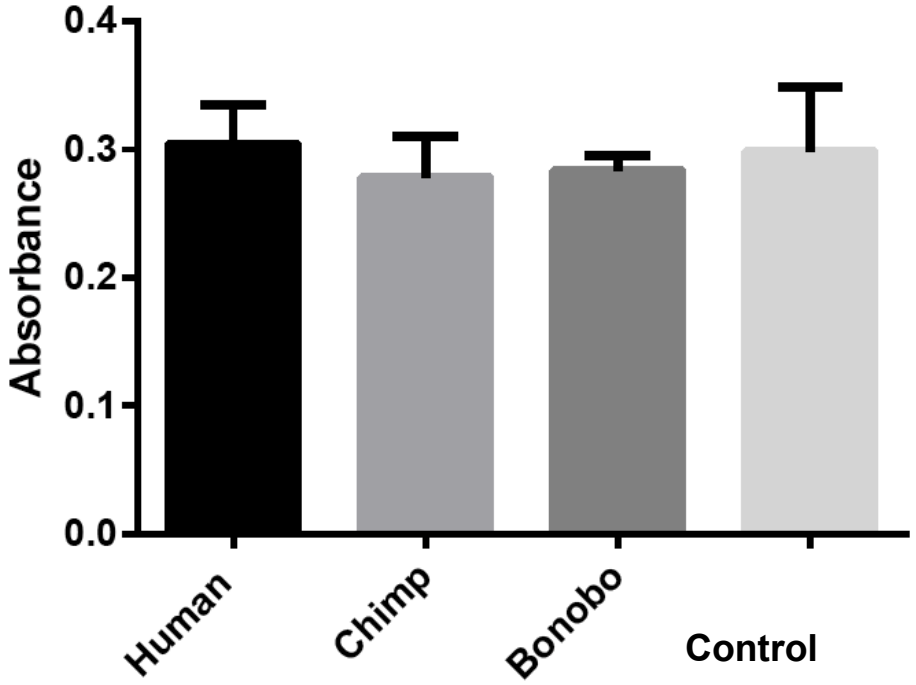


Figure 3.18: Protein concentrations of cell lysates. Kruskal-Wallis P = 0.0756. Dunn's multiple comparisons as the post hoc analysis did not find any significant difference between any groups

3.3.5 Transfection of pGL4.10 promoter-only constructs into LNCaP cells

After optimizing the transfection conditions, I transfected the pGL4.10 constructs from human, chimp, bonobo, and orangutan into the human prostate cell line (LNCaP cells), and quantified the luciferase activity. I found overall highly significant differential promoter activity among human, chimpanzee, bonobo, and orangutan (one-way ANOVA $P<0.0001$) (Fig. 3.19). Tukey's multiple comparison was performed as the *post hoc*

analysis. Human showed the highest promoter activity and orangutan showed the lowest. Bonobo, and chimpanzee showed intermediate activities. Human showed significantly higher promoter activity than chimpanzee ($P < 0.01$). This experiment was repeated thrice and we obtained similar results. The results combining all three experiments are shown in Fig. 3.20.

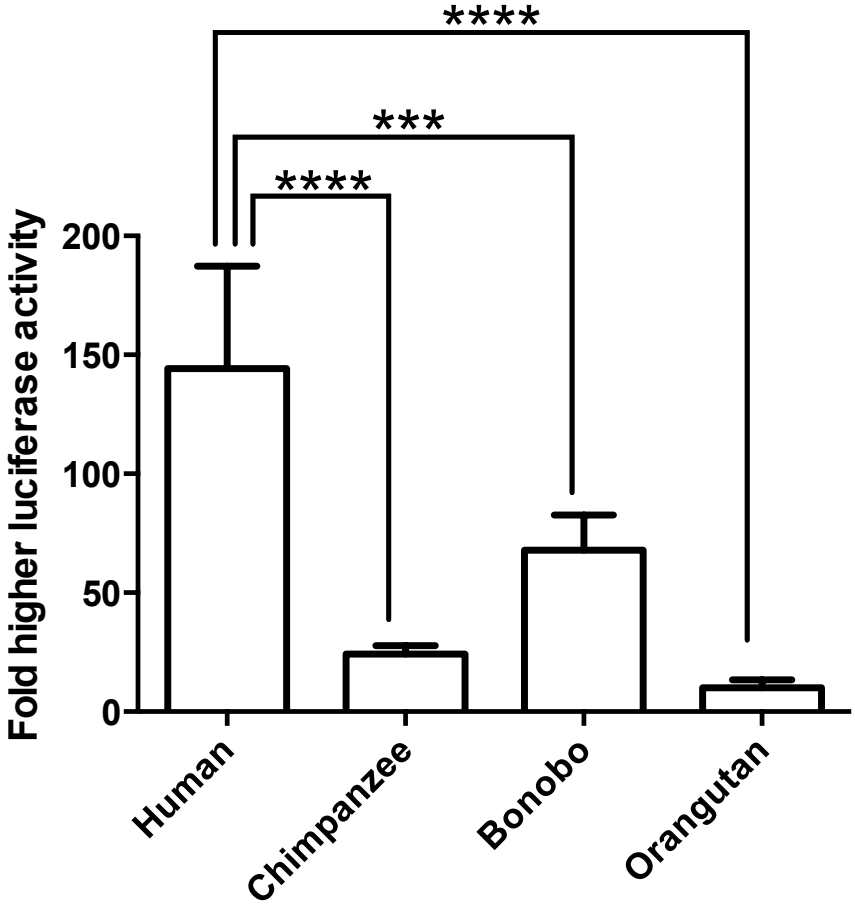


Figure 3.19: Single day transfection data of *CRTAC1* promoter-only constructs into LNCaP cells. Overall one-way ANOVA was highly significant ($P<0.001$). Human shows highly significantly higher promoter activity than both chimpanzee and bonobo ($P < 0.01$)

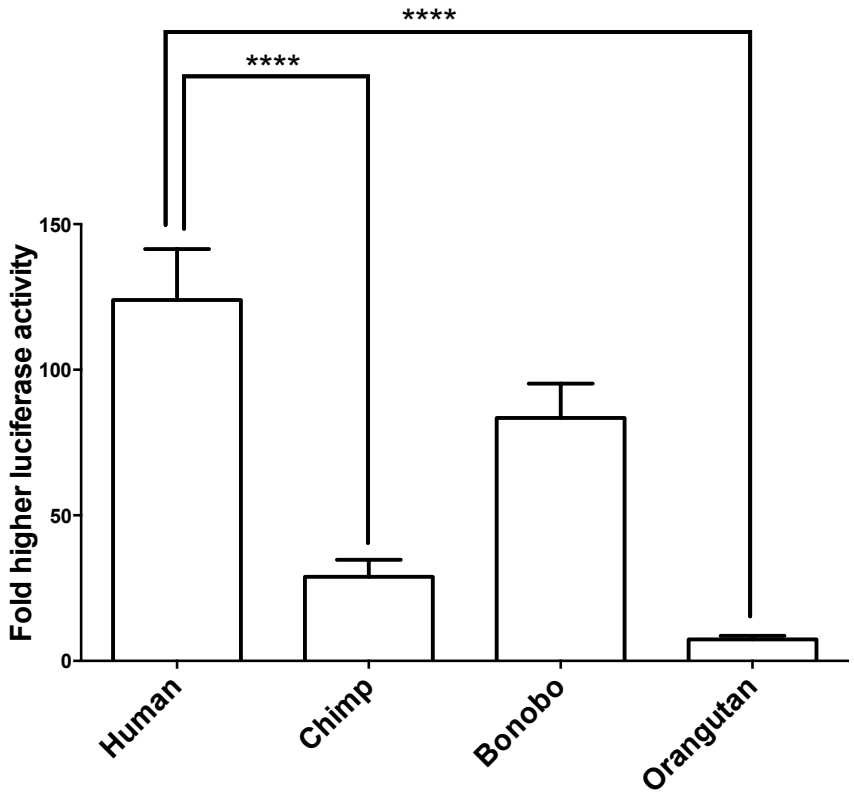


Figure 3.20: Combined transfection data of *CRTAC1* promoter-only constructs into LNCaP cells. Overall one-way ANOVA was highly significant (P<0.001). Human shows highly significantly higher promoter activity than both chimpanzee and bonobo (P <0.01)

3.3.6 Repeating transfection of human, chimpanzee and bonobo pGL4.10 constructs with new DNA midi-preps

To confirm that the above-mentioned transfection results are repeatable, I went back to my freezer stock of pGL4.10 clones, and made new preparations for human, chimp and bonobo pGL4.10 constructs (Batch 2 constructs). Like before, I then transfected the constructs into LNCaP cells and repeated the experiment thrice. The combined results are summarized in Fig. 3.21.

A two-way ANOVA was performed considering Day effect, Species effect, Batch effect and three interaction terms (Species-Batch, Species-Day and Batch-Day) using the following code:

```
ANOVA <- aov(Reading~(Species*Batch)+(Species*Day)+(Batch*Day),a)
```

where 'a' is the dataset. The ANOVA design is shown in Appendix 2.2.10.The results are summarized in Table 3.9.

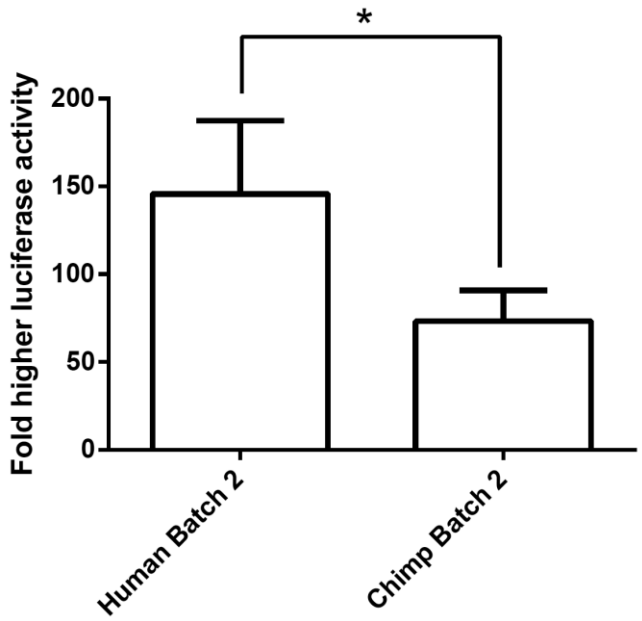


Figure 3.21: Combined transfection data of Batch 2 human and chimpanzee *CRTAC1* promoter-only constructs into LNCaP cells. Human shows significantly higher promoter activity than both chimpanzee and bonobo ($P < 0.05$)

Table 3.9: Summary of two-way ANOVA results

	Degrees of freedom	**Sum of Squares**	**Mean Squares**	**F value**	**P**
Species	2	57674	28837	63.537	1.99×10^{-13}
Batch	1	883	883	1.946	0.17031
Day	5	143987	28797	63.449	2×10^{-16}
Species:	2	7255	3628	7.993	0.00114

Batch					
Species: Day	10	32983	3298	7.267	1.57×10^{-6}

The batch effect was found to be non-significant (P = 0.17). Both older (Batch 1 constructs), and newer (Batch 2 constructs) pGL4.10 constructs showed similar trend in terms of differential promoter activity. However, the day-to-day variation for all three hominoids was highly significant (P <0.001). The interaction of Species with both Batch and Day was also highly significant (P <0.001). Species and Batch interaction plot is shown in Fig. 3.22.

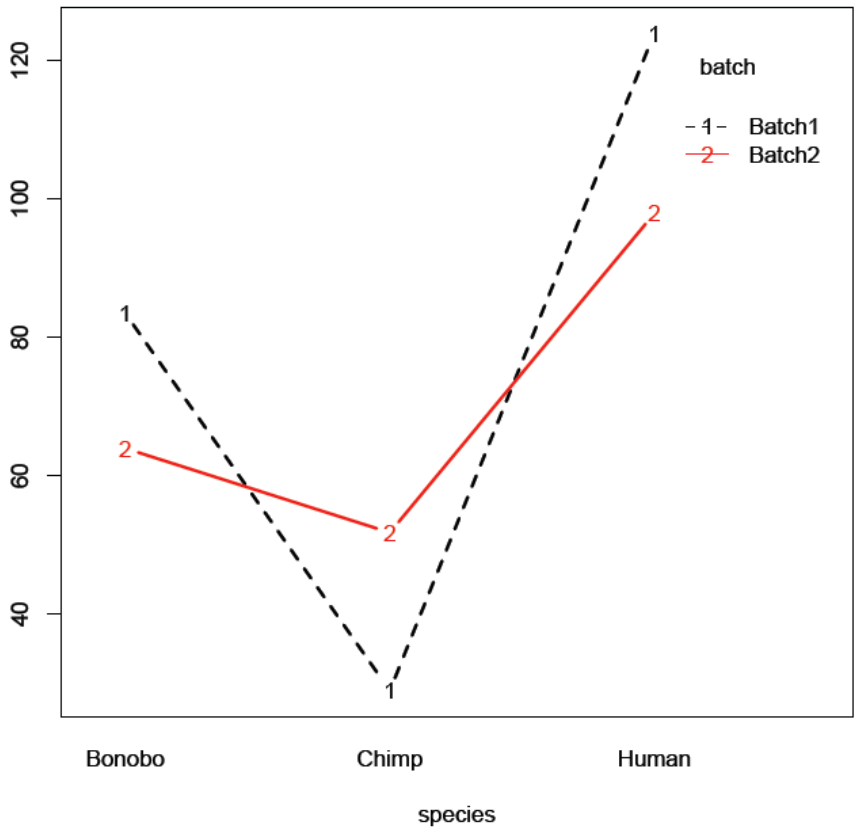


Figure 3.22: Two-way ANOVA interaction plot drawn in R v3.0.2. The Y-axis shows fold higher luciferase activity compared to the Empty vector. Both Batches show similar trend with human showing highest and chimpanzee showing least luciferase activity

3.3.7 Stimulating pGL4.10 promoter-only constructs with synthetic androgen (R1881)

Contrary to the protein data (section 3.1.7), the transfection results (section 3.3.5) showed that human promoter shows higher activity than the chimp promoter. Can androgen stimulation reverse the trend? To assess this, I stimulated human, chimp, and bonobo pGL4.10 constructs with 10 nM synthetic androgen R1881. There was no significant difference in promoter activity between the androgen-stimulated cells and non-stimulated cells in all three hominoids. Only human and chimpanzee are shown (Fig. 3.23). Human still showed significantly ($P < 0.05$) higher promoter activity than chimpanzee.

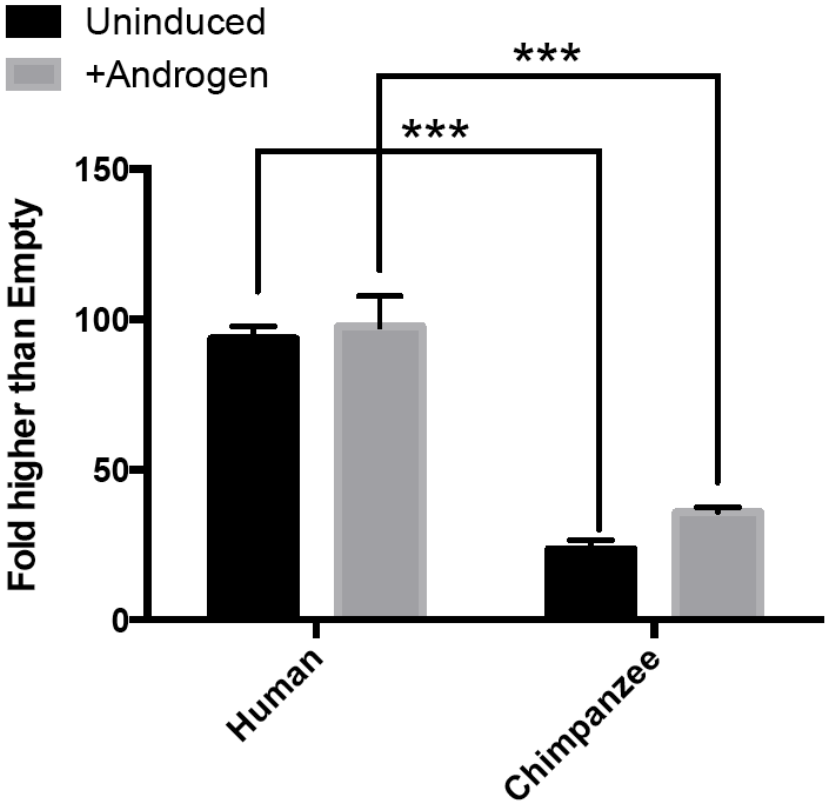


Figure 3.23: Transfection of *CRTAC1* promoter-only constructs from human and chimpanzee into LNCaP cells. One half of the constructs were stimulated with synthetic androgen (R1881), dissolved in 100% ethanol and the other half were supplemented with 100% ethanol as the vehicle control. No significant difference was observed between the stimulated and non-stimulated cells between human and chimpanzee

3.3.8 Transfection of human and chimpanzee pGL4.10 'promoter + additional *cis*-regulatory element' constructs into LNCaP cells

The possible additional *cis*-regulatory region, although highly conserved, has 40 nucleotide differences between human and chimpanzee along with one 1bp deletion, one 2bp insertion, and one 3bp insertion in human. Out of the 40 nucleotide differences, 17 are human specific and 21 are chimpanzee specific. Also, chimpanzee has a shorter [$(AC)_9$] microsatellite repeats compared to human [$(AC)_{18}$]. All excel spreadsheets containing raw values and additional graphs are added in the Appendix 2.2.7.

3.3.8.1 The additional cis-regulatory region helps in transcriptional repression

The addition of the additional *cis*-regulatory region to the pGL4.10 promoter constructs showed highly significant transcriptional repression in both human ($P<0.001$) and chimpanzee ($P<0.01$) compared to the promoter only constructs (Fig. 3.24). We repeated this experiment four times. Three times out of four we obtained similar transcriptional repression. One experiment was considered outlier and was discarded (see Appendix 2.2.7).

3.3.8.2 Human shows greater transcriptional repression compared to chimpanzee

After transcriptional repression (due to the addition of the additional *cis*-regulatory region) the difference in the transcriptional activity between human and chimp, although still significant, decreases drastically. We repeated this experiment four times. Three times out of four we obtained similar results. One experiment was considered outlier and was discarded (see Appendix 2.2.7). Human overall showed higher transcriptional repression (~11 fold) compared to chimpanzee (~2 fold). After transcriptional repression the average human luciferase activity was 13.20 fold higher

than the empty vector, very similar to the promoter-only average chimpanzee luciferase activity of 10.63 fold higher than the empty vector.

3.3.8.3 Androgen does not change the direction of the result

Like in case of promoter-only constructs, androgen stimulation does not aid in any further transcriptional repression in either human or chimpanzee (Fig. 3.24). We repeated this experiment four times. Three times out of four we obtained similar results. One experiment was considered outlier and was discarded (see Appendix 2.2.7).

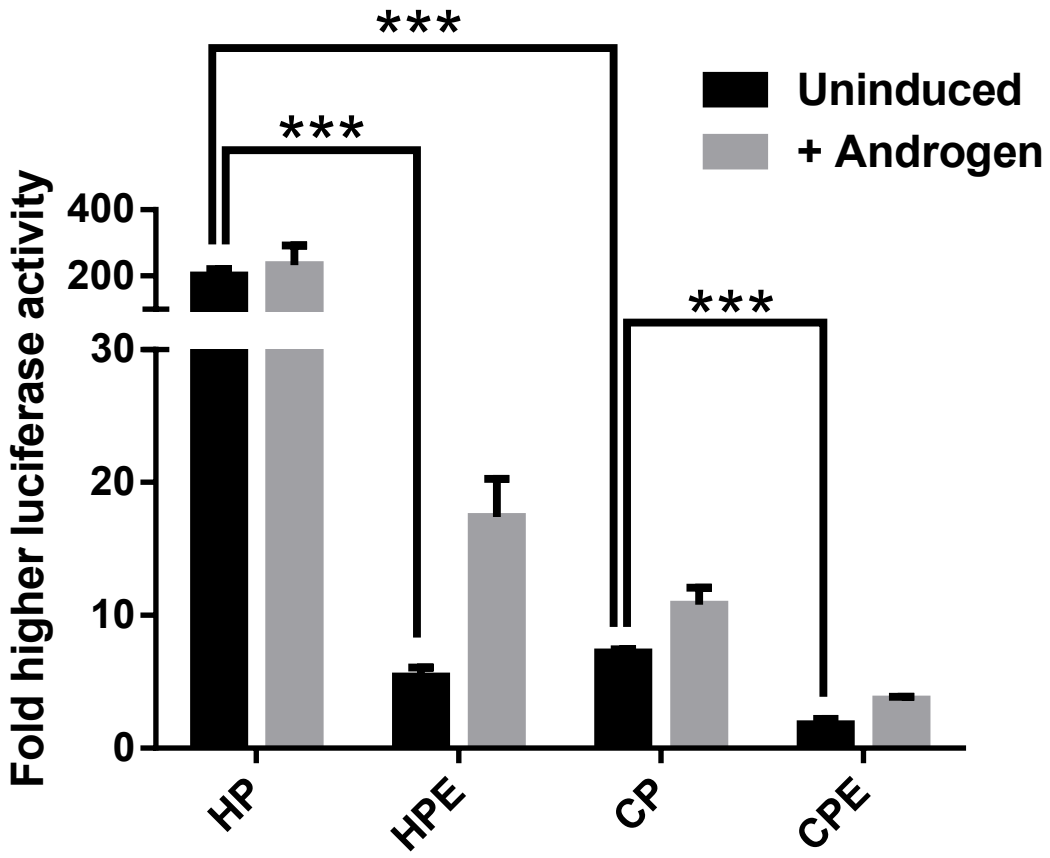


Figure 3.24: Transfection of human and chimpanzee *CRTAC1* 'promoter + *cis*-regulatory element' constructs into LNCaP cells. HP = Human Promoter-only, HPE = Human Promoter + *cis*-regulatory element, CP = Chimp Promoter-only and CPE = Chimp Promoter + *cis*-regulatory element. The additional cis-regulatory element aids in

3.3.9 Transfection of human and chimp pGL4.10 constructs into osteoblast cell line

By far we have seen that the human promoter shows higher activity than chimp promoter and the additional *cis*-regulatory element aids in transcriptional repression. But both observations were seen in human prostate cell line (LNCaP). However, as mentioned before, CRTAC1 is expressed in many human tissues. To find out whether the human promoter shows higher activity than chimp in all tissues or is it a prostate specific effect and to assess whether the *cis*-regulatory region is a universal silencer the following set of experiments were performed.

3.3.9.1 Human putative promoter potentially drives transcription significantly higher than chimp universally

To assess whether human promoter shows higher activity than chimp in all tissues or is it a prostate specific incidence, we transfected the human and chimp pGL4.10 promoter-only constructs into a human osteoblast cell line (MG63). Osteoblast was chosen because it is a non-reproductive tissue, where CRTAC1 is highly expressed. Human promoter still showed highly significantly higher promoter activity than chimp ($P < 0.01$) (Fig. 3.25). We repeated this experiment twice and obtained similar trends. All excel spreadsheets containing raw values and additional graphs are added in the Appendix 2.2.8.

3.3.9.2 The additional cis-regulatory region potentially drives repression universally

To assess whether the additional cis-regulatory region shows transcriptional repression in all tissues or is it a prostate specific incidence, we transfected the human

and chimp pGL4.10 'promoter + cis-regulatory element' and promoter only constructs into human osteoblast cell line (MG63). We found significant transcriptional repression in both human and chimpanzee (Fig. 3.25) in the osteoblast cell line like LNCaP cell line. We repeated this experiment twice and found similar results. All excel spreadsheets containing raw values and additional graphs are added in the Appendix 2.2.8.

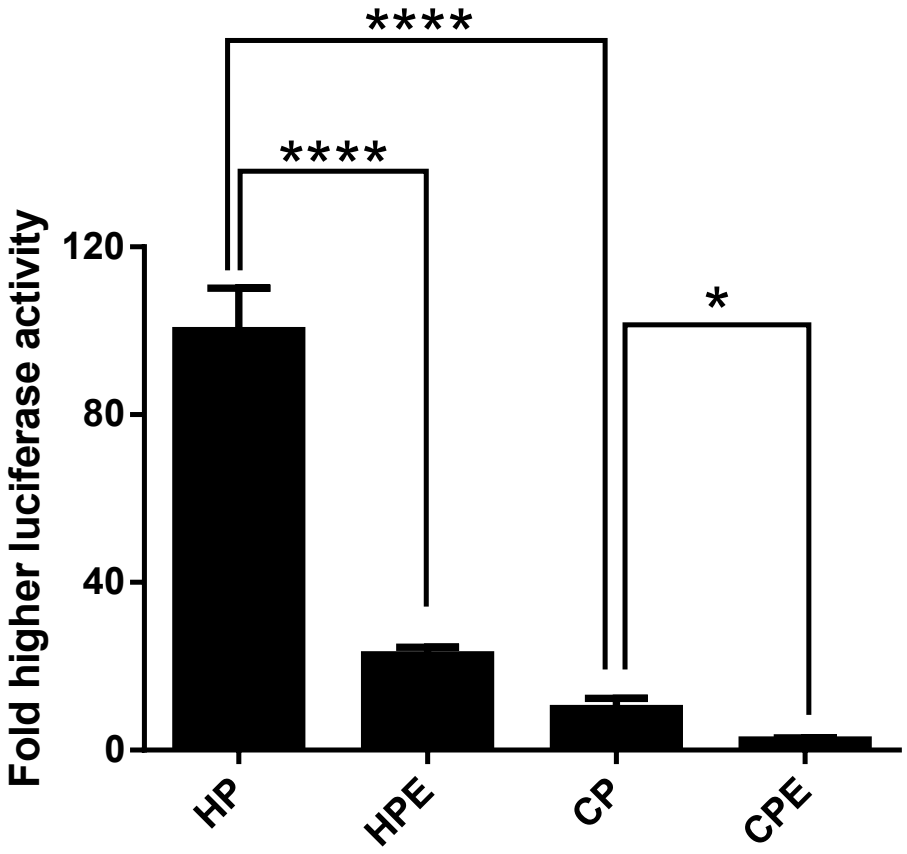


Figure 3.25: Transfection of human and chimpanzee *CRTAC1* 'promoter + *cis*-regulatory element' constructs into human osteoblast (MG63) cells. HP = Human Promoter-only, HPE = Human Promoter + additional *cis*-regulatory region, CP = Chimp Promoter-only, and CPE = Chimp Promoter + additional *cis*-regulatory region. Human promoter still shows higher activity than chimpanzee promoter. The additional cis-regulatory element still aids in transcriptional repression in both species ($P < 0.01$) and seems to be a universal repressor

3.3.10 Analysis of the protein-coding region of *CRTAC1* from four hominids

The consensus sequences from each exon were subsequently joined together to generate the complete virtual cDNA for each species (see Methods). The ClustalW alignment of the cDNA sequences, and inferred amino acid sequences from the four species is shown in Appendix 2.1.1 and 2.1.2. The alignment shows 98% sequence identity among the four hominid species.

Out of the 38 nucleotide substitutions in the cDNAs, only five are nonsynonymous. The amino acid substitutions observed include the presence of phenylalanine in the chimpanzee protein instead of valine at position 328, presence of glycine in human the place of valine at position 606 and presence of threonine in place of proline in gorilla at position 607. Although both are non-polar, phenylalanine (uniquely present in chimpanzee) contains a benzyl ring in its side chain while valine does not have a benzyl ring. Glycine (uniquely present in human) is distinctly smaller in size than valine. Glycene is unique because it not chiral and can fit to both hydrophilic and hydrophobic environments, due to its hydrogen atom side chain. Finally, both threonine and proline are α-amino acids but proline (uniquely absent in gorilla), is the only essential amino acid where the α-amino group is secondary (imino acid).

Both the rates of pair-wise nonsynonymous substitution (d_N) (Table 3.10) and the rates of pair-wise synonymous substitution (d_S) (Table 3.11) were very low for all hominid pairs.

Table 3.10: The rate of pair-wise nonsynonymous substitution (d_N)

	Chimpanzee	Gorilla	Human
Chimpanzee			
Gorilla	0.00208		
Human	0.00278	0.00208	
Orang	0.00208	0.00278	0.00347

Table 3.11: The rate of pair-wise synonymous substitution (d_S)

	Chimpanzee	Gorilla	Human
Chimpanzee			
Gorilla	0.01695		
Human	0.02970	0.02967	
Orang	0.05096	0.05090	0.05946

All d_S were higher than d_N values by one order of magnitude (Table 3.12). As a result all ω values were less than 1 indicating that purifying selection is operating at the coding region of *CRTAC1*.

Table 3.12: The pair-wise ω (d_N/d_S) values

	Chimpanzee	Gorilla	Human
Chimpanzee			
Gorilla	0.1227		
Human	0.0936	0.0701	
Orangutan	0.0408	0.0546	0.058

The likelihood ratio tests employed for all models (Branch, Site and Branch-site models) were non-significant suggesting the probable absence of positive selection in any branch of hominid phylogeny for *CRTAC1* protein coding region (Table 3.13). Although statistically non significant, the ω estimated under a free-ratio model, was highest on the chimpanzee branch after its isolation from humans with a lower d_N (0.0012) and d_S (0.0092) compared to humans (0.0016, and 0.0231 respectively), which may be biologically significant (Fig. 3.26). $\omega << 1$ suggests the operation of strong purifying selection in the coding region.

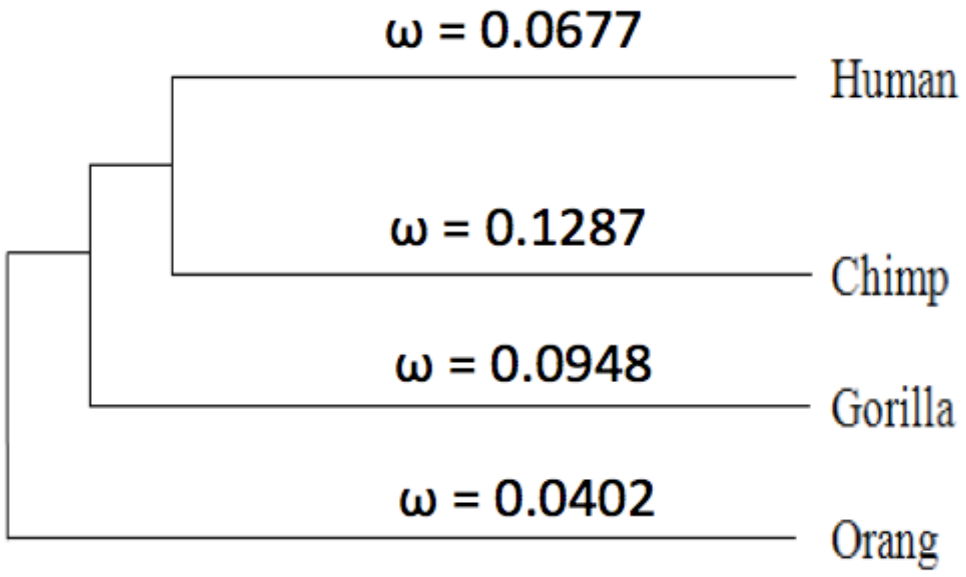


Figure 3.26: PAML4 output showing ω (d_N/d_S) values on the branches. ω was estimated using a free ratio model considering the true phylogenetic relationship among the four species

Table 3.13: Likelihood Ratio Tests (LRT) with different models for ω

H_0 vs. H_1	**Likelihood Values**	**$2\Delta l$ (χ^2) and df (df = the difference in no. of parameters)**	**LRT P value**
M0 (Uniform ω) vs. M1 (Free-ratio ω)	M0 = -2753.002 M1 = -2752.384	$2\Delta l$ (χ^2) = 1.236 df = 4	0.87
M1a (Neutral evolution) vs. M2a (Positive Selection)	M1a = -2750.452 M2a = -2750.068	$2\Delta l$ (χ^2) = 0.768 df = 2	0.68
M7 (Beta distributed Neutral model) vs. M8 (Beta distributed Positive Selection)	M7 = -2751.001 M8 = -2750.068	$2\Delta l$ (χ^2) = 1.866 df = 2	0.39
M8a (Beta distributed fixed ω=1) M8 (Beta distributed Positive Selection, ω>1)	M8a = -2750.452 M8 = -2750.068	$2\Delta l$ (χ^2) = 0.768 df = 1	0.38
M0 (Uniform selective pressure among sites) vs. M3 (Discrete selective pressure among sites)	M0 = -2753.002 M3 = -2750.068	$2\Delta l$ (χ^2) = 5.868 df = 14	0.96
M1a (Neutral evolution) vs. MA (Branch-Site model with fixed ω>1)	M1a = -2750.452 MA = -2750.452	$2\Delta l$ (χ^2) = 0 df = 1	NA

3.3.11 Population Genetic analysis of the protein-coding region of CRTAC1 from five hominids

Chimpanzee has four SNPs out of Hardy-Weinberg Equilibrium with high inbreeding coefficients. In all cases *P. t. verus* is fixed with the alternate allele. All gorilla, orangutan and bonobo SNPs are in Hardy-Weinberg Equilibrium. Chimpanzee

shows higher nucleotide diversity (0.0025) compared to gorilla (0.0016). Chimpanzee has 11 non-synonymous SNPs (out of 19, 58%) compared to 6 (out of 11, 55%) in gorilla, and 1 (out of 7, 14%) in human. Chimpanzee SNPs are summarized in Table 3.14 and 3.15. Human SNPs are summarized in Table 3.16. Gorilla SNPs are summarized in Table 3.17 and 3.18. The SNPs from other hominoids are summarized in Table 3.19.

Table 3.14: Population genetics of SNPs found in the protein-coding region of chimpanzee CRTAC1 (2N = 50)

Exon	SNP Position	Synonymous/Non-synonymous	MAF	Nucleotide substitution	Amino acid substitution	Observed Heterozygosity	HWE P	F
1	97517766	Synonymous	0.02	A>G	NA	1	0.92	-0.02
2	97498542	Non-synonymous	0.11	T>A	V>D	2	0.03	0.51
5	97404841	Non-synonymous	0.20	G>T	A>S	0	0.001	1
	97404811	Synonymous	0.06	T>C	NA	3	0.75	-0.64
	97404768	Non-synonymous	0.35	T>C	Y>C	11	0.98	-0.01
	97404725	Synonymous	0.10	G>A	NA	5	0.57	-0.16
7	97391145	Synonymous	0.24	C>T	NA	8	0.54	0.12
	97391126	Non-synonymous	0.06	G>C	S>T	3	0.75	-0.64
10	97382302	Non-synonymous	0.04	G>T	E>D	2	0.84	-0.42
	97382256	Synonymous	0.24	G>A	NA	2	0.001	0.78
11	97381702	Non-synonymous	0.25	G>A	D>N	9	0.32	0.19
12	97370543	Synonymous	0.04	G>A	NA	2	0.84	-0.42
	97370526	Non-synonymous	0.02	G>A	V>T	1	0.92	-0.02
	97370525	Non-synonymous	0.02	T>C	V>T	1	0.92	-0.02
14	97366523	Non-synonymous	0.04	G>A	S>N	2	0.84	-0.42
	97366519	Non-synonymous	0.25	G>C	S>T	9	0.38	0.18
	97366498	Non-synonymous	0.17	C>T	R>W	8	0.001	1
15	97351189	Synonymous	0.02	A>G	NA	1	0.92	-0.02
	97351164	Synonymous	0.02	T>C	NA	1	0.92	-0.02

Table 3.15: Pair-wise F_{st} for chimpanzee subspecies and Tajima's D test for Neutrality

Overall F_{st}	Pair-wise F_{st}	Overall Nucleotide Diversity (θ)	Overall Nucleotide Diversity (π)	Tajima's D
0.47	1. *P. t. ellioti – P. t. schweinfurthii*: 0.24	0.0025	0.0027	0.305 (P > 0.05)

	2. *P. t. ellioti – P. t. troglodytes*: 0.25 3. *P. t. ellioti – P. t. ver* 0.64 4. *P. t. schweinfurthii – P. t. troglodytes*: 0.17 5. *P. t. schweinfurthii - P. t. verus*: 0.69 6. *P. t. troglodytes - P. t. verus*: 0.75			

Table 3.16: Population genetics of SNPs found in the protein-coding region of human CRTAC1

Position	**Individual Genotype**	**Amino acid Substitution**	**Synonymous/Non-synonymous**	**Chimp and Neanderthal/ Denisovan**
10: 9962531 (Exon 15) (rs56007204)	C/T (84% C and 16% T)	E/K	Non-synonymous	Chimp has C Denisovan had Neanderthal ha
10: 9966445 (Exon 8) (rs7068503)	C/T (93% T and 7% C)	Q/Q	Synonymous	Chimp has C Denisovan had Neanderthal ha
10: 9969600 (Exon 3) (rs35027739)	G/A (91% G and 9% A)	I/I	Synonymous	Chimp has G Denisovan had Neanderthal ha
10: 9964012 (Exon 14) (rs2297935)	C/T (98% C and 2% T)	V/V	Synonymous	Chimp has C Denisovan had Neanderthal ha
10: 9965564 (Exon 10) (rs577537)	G/T (96% G and 4% T)	G/G	Synonymous	Chimp has G Denisovan had Neanderthal ha
10: 9966786 (Exon 6) (rs35853031)	C/T (97% C and 3% T)	A/A	Synonymous	Chimp has C Denisovan had Neanderthal ha

10: 9969596[illegible] (Exon 3) (rs3750608)	C/T (99.4% C and 0.6% T)	R/R	Synonymous	Chimp has C Denisovan had Neanderthal ha[illegible]

Table 3.17: Population genetics of SNPs found in the protein-coding region of gorilla CRTAC1 (2N = 56)

Exon	SNP Position	Synonymous/Non-synonymous	MAF	Nucleotid substitutio	Amino acid substitution	Observe Heterozy ity	HWE P value	F
3	111315839	Synonymous	0.07	C>T	NA	4	0.68	-0.09
5	111298570	Synonymous	0.39	T>G	NA	10	0.34	0.19
8	111282448	Non-synonymous	0.02	G>A	D>N	1	0.92	-0.02
9	111278450	Non-synonymous	0.28	A>G	NA	9	0.23	0.22
	111278437	Synonymous	0.13	C>G	NA	7	0.45	-0.18
10	111277462	Synonymous	0.11	A>G	D>G	7	0.45	-0.18
11	111276859	Non-synonymous	0.15	G>T	K>N	7	0.73	0.08
	111276763	Non-synonymous	0.02	T>G	L>R	1	0.92	-0.02
12	111259199	Non-synonymous	0.08	G>A	E>K	5	0.61	-0.09
	111259157	Non-synonymous	0.03	G>A	A>T	2	0.84	-0.07
15	111240408	Non-synonymous	0.11	C>G	NA	5	0.33	0.24

* For this SNP position the three Eastern gorilla sequenced in GAGP database also shows SNP

Table 3.18 Tajima's D test for Neutrality for gorilla

Overall Nucleotide Diversity (θ)	Overall Nucleotide Diversity (π)	Tajima's D
0.0016	0.0019	0.677 (P > 0.05)

Table 3.19: Population genetics of SNPs found in the protein-coding region of CRTAC1

Species	SNP position[1]	Exon	Observed heterozygosity	Expected hetrozygosity	HWE P	Inbreeding-coefficient (F)

Bonobo (*Pan paniscus*) (2N = 26)	97422588 97394484 97382305	3 6 10	1 2 2	0.96 1.85 1.85	0.88 0.76 0.76	0.00 -0.04 -0.04
Orangutan (*Pongo abelii*) (2N = 10)	97012262 96992755 96878086 96877124 96876430	1 2 9 10 11	1 1 4 2 2	0.88 0.90 2.40 1.60 1.50	0.77 0.81 0.13 0.58 0.51	-0.19 -0.11 -0.66 -0.25 -0.33

3.4 Discussion

3.4.1 Amplification, sequencing and transfection of the cis-regulatory elements of CRTAC1

3.4.1.1 Amplification of the putative promoter region of CRTAC1 using Polymerase Chain Reaction (PCR)

The putative promoter region of *CRTAC1* is highly GC rich (>70%). The highly GC rich DNA fragments tends to form secondary structures and are very difficult to denature. As a result, it is very difficult to PCR amplify GC rich DNA fragments and during this kind of PCR many short, non-specific DNA fragments are generated instead of the desired products (McDowell et al. 1998, Mamedov et al. 2008, Tindall et al. 2009). In the past several methods have been employed to amplify the GC rich DNA fragments. Some methods include addition of different chemical reagents such as dimethylsulfoxide (DMSO) (Sun et al. 1993), betaine (Henke et al. 1997), glycerol (Choi et al. 1999) and 7-deaza-dGTP (Frey et al. 2008). Many authors tried to combine two or more of the above-mentioned reagents for example DMSO and betaine (Kang et al. 2005, Ralser et al. 2006, Sahdev et al. 2007) or DMSO, betaine and 7-deaza-dGTP (Musso et al. 2006).

Techniques such as 'Slowdown PCR' (Frey et al. 2008) and 'Two-Step PCR' (Schuchard et al. 1993) have also shown to improve the amplification of GC rich DNA fragments.

I tried all of the above-mentioned techniques and different combinations of them for example combining 'Touchdown PCR' with DMSO and betaine, but none of them could successfully amplify the putative promoter region of *CRTAC1*. I finally adapted and modified the 'SAFE PCR' protocol (Wei et al. 2010). This technique included a combination of modification of PCR cycles, change in thermal cycler ramp speed, and simultaneous use of 5% DMSO, 5% betaine and 5% glycerol (See Results 3.3.1). A comparative analysis of the 'SAFE PCR' technique and my adaptation of the same are shown in Fig. 3.11. My adaptation was faster than the original technique since it did not include 'Hot start' at the beginning of each step. Also my adaptation reduced the use of expensive *Taq* polymerase since it was use only once compare to thrice in the original technique. My adaptation can be used for any commonly used thermal cycler machine and does not need an external computer support. On the other hand, the original 'SAFE PCR' technique requires external computer support to regulate the ramp speed. In my adaptation the slow ramp speed is maintained by programming the thermal cycler itself.

The most interesting aspect of 'SAFE PCR' is the successful adaptation of 'Slowdown PCR' (Frey et al. 2008), which runs at a reduced ramp rate of 2.5°C/sec. The cooling rate to reach the annealing temperature is especially slow at 1.5°C/sec. It is thought that 'Slowdown PCR' optimizes primer annealing by decreasing the annealing temperature every third cycle and using slow cooling rate (Frey et al. 2008). This technique is thought to increase the likelihood of primer binding specificity and enable primer binding at the optimal annealing temperature (T_a), at which primers are supposed

to bind only to their specific templates (Frey et al. 2008). The slow ramp speed also prevents the formation of interfering secondary structures in GC rich DNA fragments that also aid in successful binding of the primers to their specific templates (Frey et al. 2008). However, to change the ramp speed one needs a special thermal cycler that allows the alteration heating (2.5°C/sec) and cooling ramp rates (1.5°C/sec) for example TGradient Cycler (Biometra) used by Frey et al. (2008) or Bio-Rad Mycycler Thermal Cycler used by Wei et al. (2010). On the contrary, my adaptation of 'SAFE PCR' was performed in GeneAmp ® 9700 PCR System (Applied Biosystem), which does not allow the exact change in ramp rate during heating and cooling. It, however, allows changing the ramp speed in percentage (the default ramp speed being 100%). The 100% ramp speed of the GeneAmp ® 9700 PCR System is ~ 3-3.5°C/sec (GeneAmp ® 9700 PCR System manual). I maintained 90% overall ramp speed during my adaptation of 'SAFE PCR' (~ 2.7°C/sec) and 60% ramp speed during cooling (~1.8°C/sec), which are approximately equal to the heating and cooling ramp rate used in 'Slowdown PCR' and 'SAFE PCR'. The successful adaptation of 'Slowdown PCR' and 'SAFE PCR' that requires complex and expensive thermal cyclers to a comparatively cheaper, and commonly available thermal cycler makes my adaptation much more superior to the above-mentioned techniques.

3.4.1.2 Sequencing the putative promoter region of CRTAC1

As mentioned in Section 3.3.2 *CRTAC1* entire putative promoter region was sequenced twice: once after PCR amplification and another after TOPO cloning using BigDye cycle sequencing chemistry on capillary ABI-3100 auto sequencer. Sequencing the same regions twice and comparing with the UCSC Genome Browser confirmed the

integrity of the sequences. UCSC Genome browser had long gaps in *CRTAC1* putative promoter regions of chimpanzee and gorilla. I could PCR amplify and sequence through the gaps. The gaps turned out to be not real and artifacts of missing data. Thus my chimpanzee and gorilla putative promoter sequences are of superior quality compared to the public database. These sequences will be submitted to the Gen Bank and can be used to fill the gaps. The nucleotide differences in the putative promoter region observed between my sequences and UCSC Browser sequences, all correspond to real SNPs for that particular species (See Results section 3.3.2). This reconfirms the integrity of my sequences, showing that the observed nucleotide differences are not just artifacts of PCR or sequencing error.

CRTAC1 putative promoter region has only eight nucleotide differences among human, chimpanzee and bonobo. As mentioned before (See Results section 3.3.2) Orangutan putative promoter looked most different among all hominoids. Orangutan has two unique sequence insertions and one unique deletion in this region. The insertions are the binding sites of two transcriptional repressors Msx-1 and EBF1. Msx-1 directly interacts with TATA-binding protein (TBP) and aids in transcriptional repression in mouse (Zhang et al. 2002). EBF1 has also been shown to help in transcriptional repression (Timblin and Schlissel 2013; Banerjee et al. 2013). The unique deletion in orangutan putative promoter region corresponds to the binding site of Activating enhancer Protein-2 alpha (AP-2 alpha) binding sequence. AP-2 alpha overexpression has shown to inhibit chondrocyte differentiation (Huang et al. 2004). On the other hand humans have a 2bp deletion in the putative promoter region. This deletion disrupts the consensus sequence of CACCC-binding protein, which aids in transcriptional repression

of several genes (Vliet et al. 2000, Funnell et al. 2012). The *in vitro* luciferase assay showed that the human putative promoter is stronger than other hominids in driving the transcription and the difference in transcription activity. This may be due the unique nucleotide differences or the unique deletion found in the human putative promoter region. The orangutan promoter drove transcription weakest among all hominoids may be because of the distinct nature of its promoter with unique binding sites for MSX1 and EBF1 both of which can cause transcriptional repression.

3.4.1.3 In vitro luciferase assay with hominoid pGL4.10 constructs

For many years authors thought that most morphological adaptations take place through changes in non-coding DNA sequences (Haygood et al. 2010) and a great deal of gene regulation takes place at the transcriptional level (Wray et al. 2003) due to the variation in the *cis*-regulatory elements. One of the best approaches to assess *cis* regulatory variation is to investigate differential transcriptional activity by developing reporter gene assays (Wray 2007). *In vitro* cell culture and luciferase assay, which has successfully been used by many authors for the reporter gene assay, (Huby et al. 2001, Rockman et al. 2005, Inoue-Murayama et al. 2006, Loisel et al. 2006, Chabot et al. 2007), can provide convincing results and may be the most appropriate technique for studying transcriptional regulation.

One of the major issues in *in vitro* luciferase assays is to use the suitable internal control. The most commonly used internal control is co-transfecting the pGL4.10 constructs with a *Renilla* control vector. It has been used previously for transfection assays of hominoids (Babbit et al. 2009). It should be expressed at a lower level regardless the concentration of the other vector and control for the transcription

efficiency (Schagat et al. 2007). However, during transfection optimization we found that *Renilla* expression level changed with the strength of the pGL4.10 vector (Carnahan-Craig 2013). We found almost a linear relationship between the increase of *Renilla* luciferase with the firefly luciferase. In other words, there appears to be cross talk going on between the two co-transfected vectors (Carnahan-Craig 2013). It has been shown recently (Shifera and Hardin 2010) that in experimental conditions *Renilla* luciferase vector can be suppressed or induced by one or more experimental factors including the co-transfected vector. So, additional normalization may be required for these luciferase assays (Shifera and Hardin 2010). We as a lab decided to stop using *Renilla* as the internal control for the luciferase assays and in stead increase the number of replications of the experiments. Since we stopped using *Renilla* as the internal control during transfection, one of the major questions that came up was how to determine whether we are plating equal number of cells in each well on a particular day. So, I used the whole protein concentration as the internal control, considering the whole protein concentration will be proportional to the number of cells plated in a given plate. This approach has been used previously (Smale 2008) for *in vitro* luciferase assays. I found no significant well-to-well variation for the number of cell lysate protein concentrations plated on a given day (Kruskal-Wallis test P =0.0756; See Table 3.8; Fig. 3.18), suggesting that there is no significant well-to-well variation for the number of cells plated on a given day

Once the transfection conditions were optimized, I cloned ~1.9kb upstream region of the transcriptional start site (TSS) of *CRTAC1* as the putative promoter region into luciferase containing pGL4.10 vector for *in vitro* luciferase assays. I used four hominoid species- human, chimpanzee, bonobo, and orangutan for *in vitro* luciferase assay. I chose

~2kb region upstream of TSS because several genome wide studies have shown that most of the *cis* regulatory elements, functional regions and putative promoter regions concentrate within ~2kb upstream of TSS (Farré et al 2007, Trinklein et al. 2003). The transfection experiments paradoxically showed the exact opposite outcome compared to the proteomic data. Human putative promoter drove transcription significantly ($P < 0.01$) higher than chimpanzee. However, I found highly significant species specific variation in transcription (ANOVA $P < 0.01$) with orangutan putative promoter driving transcription significantly less than all other hominoids. The results of the *in vitro* luciferase assays suggested that the putative promoter is not the only *cis* regulatory factor driving the transcription of *CRTAC1* and potentially there are additional transcriptional enhancers/repressors involved in *CRTAC1* transcription.

My *in vitro* luciferase assay results are much more reliable than previous such assays for hominoids. Although multiple replicates have been used previously (Chabot 2007), the luciferase activity of the empty vector was calculated just once and the human and chimpanzee promoters were compared to the same empty vector every time. As a result their replicates cannot be considered as true replicates but are just pseudo-replicates of the same experiment. I calculated luciferase activity of the empty vector for all experiments separately not just once as done previously (Chabot 2007). This made all my transfection assays true experimental replicates of each other since the fold higher luciferase activity was calculated by comparing the luciferase activity of a particular species with the luciferase activity of the empty vector of that experiment. Also, no previous *in vitro* luciferase assays for hominids included the 'batch' factor in their experiment. The use of the 'day' and 'batch' factors further strengthens the integrity and

reproducibility of my luciferase assays, showing that the raw number may vary on a day-to-day basis but the trend will remain the same. Beside the species-to-species variation, I found significant ($P < 0.01$) day-to-day variation in the luciferase assays. This may be seen due to the day-to-day variation in the experimental conditions (physical conditions of the cells, fluctuations in the humidity and CO_2 in the incubators, and other manual errors). If the 'day' factor is not included in the *in vitro* luciferase assays, one can never encounter these issues, which can lead to misinterpretations of the results. I strongly recommend using multiple true replicates and including the 'day' and 'batch' factors in *in vitro* luciferase assays to avoid over/under estimations of the assay results.

3.4.1.4 Discovery and transfection of potential additional cis-regulatory region of CRTAC1

I identified a ~2.2kb putative *cis* regulatory region in intron 11 of *CRTAC1*. This region had strong H3K27Ac signal, which is associated with active regulatory elements outside the promoter region (ENCODE Project Consortium, Dunham et al. 2012). This approach of identifying a regulatory region based on histone marks has been recently employed by authors to identify (Ong and Corces 2011, Fernandez and Saavedra 2012) and *in vitro* characterize (Blum et al. 2012) enhancers. This region is also a strong DNase hypersensitive site (ENCODE Project Consortium, Dunham et al. 2012). DNase hypersensitive sites are one that have lost their condensed nature with exposed DNA and are highly accessible to transcription factors. ~600bp of this regulatory region is highly conserved among mammals. I cloned this region from human and chimpanzee into pGL4.10 constructs already containing human and chimpanzee putative promoters respectively. Similar *in vitro* luciferase assays with 'promoter + enhancer' constructs

have been previously performed and shown to cause transcriptional activation (Latham et al. 2000, Iwagawa et al. 2013). This *cis* regulatory region turned out to be an active silencer. It significantly repressed transcription in both human and chimpanzee. However, the repression was stronger in human (~11 fold repression) compared to chimpanzee (~2 fold). The addition of the repressor element to human *CRTAC1* putative promoter construct repressed transcription to the level of chimpanzee promoter-only construct. Both human and chimpanzee constructs showed significant transcriptional repression after the addition of the element. So, potentially there are additional enhancers/repressor elements and/or tissue specific alternative splicing involved in *CRTAC1* transcriptional regulation.

Eukaryotic gene regulation is complex and cannot be not explained by the simple model of one promoter and one enhancer. The regulation takes place at the transcriptional level, post-transcriptional level, translational level as well as post-translational level. So, the differential CRTAC1 expression in the seminal plasma of chimpanzees and humans may be due to a differential regulation activity in any of the above-mentioned stages. Recent studies have shown the combinatorial effect of multiple enhancers during regulation of gene expressions (Perry et al. 2011, Corradin et al. 2013) and many enhancers are very tissue specific (Ong and Corces 2011). The problem with identification of the right enhancer is the fact that eukaryotic transcription can be regulated by long-range (>10kb) cis regulatory elements (Ong and Corces 2011, Spivakov 2012, Vadnais 2013) and often involve long range interaction with the promoter **(**Ong and Corces 2011). So, the cis regulatory element in the intron 11 of *CRTAC1* may not be the only regulatory element for *CRTAC1* transcriptional regulation

or may not even participate in the transcriptional regulation al all *in vivo*. Even if it participates in *CRTAC1* transcriptional regulation, the final outcome is potentially aided by additional long-range regulatory element s and/or tissue specific regulatory element for final transcriptional outcome. Also in recent years various different non-coding RNAs have been identified that can actively participate in gene regulation. Some of these regulatory RNAs such as enhancer RNAs (eRNA) (Kim et al. 2010), and linkRNAs (Orom et al. 2010) can actually aid in the long-range interaction between the promoter and enhancer. Techniques like chromatin conformation capture (3C) or the various variations of the technique and/or fluorescent in situ hybridization (FISH) could be employed to identify the right gene specific or tissue specific enhancers by showing physical association between genomic elements within the nucleus (Ong and Corces 2011).

3.4.2 Molecular Evolution of CRTAC1

I sequenced all 15 exons of *CRTAC1* from four hominids - human, chimpanzee, gorilla and orangutan and virtually joined the exons together to create the entire protein coding DNA sequence (CDS) for each species. I found 99% sequence identity among the four hominids. The CDS is under strong purifying selection with ω (d_N/d_S) for all branches << 1. There was no insertion or deletion in the CDS of any of the four hominids. Also, there were only five non-synonymous substitutions in the entire 1911bp long CDS.

Likelihood Ratio Tests (LRT) with various models for ω was performed to find out whether any of the branches in the hominid phylogeny is evolving under positive selection for *CRTAC1* CDS. LRTs have been used by several authors to investigate the

action of positive selection in a given branch of a phylogenetic tree (Swanson et al. 2003, Clark and Swanson 2005). Previous pairwise ω estimation between human and chimpanzee has showed seven (*TGM4, KLK2, ACPP, DBI, PIP, TMPRSS2, MSMB*) seminal plasma genes to be under positive selection (Clark and Swanson 2005). Clark and Swanson (2005) employed three different comparisons of neutral vs. positive selection models (M1 vs. M2, M7 vs. M8 and M8a vs. M8) using LRT in PAML package. All comparisons rejected the model of neutral evolution over that of positive selection. Several mammalian reproductive genes (PH20, Fertilin, CD9, Zonadhesin, and SP17) are under strong positive selection (Swanson et al. 2003) and LRTs performed with these genes (M7 vs. M8, M8a vs. M8) rejected neutral evolution over positive selection. I employed similar comparisons of neutral vs. positive selection models for *CRTAC1* CDS. All LRTs were non-significant suggesting probable absence of positive selection in any branch of hominid phylogeny for *CRTAC1* CDS (See Results Section 3.3.10). My results indicate that *CRTAC1* CDS is very different than other reproductive genes or seminal plasma genes and evolves much slowly than the above-mentioned tissue-specific genes.

A $\omega << 1$ is commonly seen for most genes in human genome (Zhang and Li 2004, Zhu et al. 2008, Wang et al. 2011). For *CRTAC1* CDS human and mouse shows ~ 90% sequence similarity, higher than that of the genomic mean (84%), reconfirming the slower evolution of the gene (Wang et al. 2011). A genome-wide study with 1581 genes using PAML package showed that the genes expressed in multiple tissues (≥16) ("housekeeping genes") are under strong evolutionary constraint (evolves slowly) and are under strong purifying selection (human-mouse mean $\omega = 0.093$, median = 0.114) than

tissue-specific genes (Zhang and Li 2004). Although some prostate-specific genes are under strong positive selection in hominids (Clark and Swanson 2005), the overall mean (0.108) and median (0.125) ω for prostate specific genes between human and mouse are lower than all other tissues (lung being the highest with mean and median of 0.259 and 0.172 respectively) (Zhang and Li 2004). The non-synonymous rates of substitution are highly significantly lower for housekeeping genes than tissue-specific genes (human-mouse mean d_N = 0.046, P <0.001) (Zhang and Li 2004). For *CRTAC1* the human-mouse pairwise ω (0.129) is comparable to the highly expressed genes (mean = 0.097) and prostate-specific genes (mean = 0.108) but much lower than lung (mean = 0.259), liver (mean = 0.233) and kidney (mean = 0.166) specific genes (Zhang and Li 2004). Interestingly, the non-synonymous rate of substitution (d_N) for *CRTAC1* between human and mouse (0.068; 0.069 between mouse and chimp) is very much like the mean non-synonymous rate of substitution observed in case of prostate specific genes (0.066) and higher than that of housekeeping genes (0.046) (Zhang and Li 2004). The synonymous rate of substitution (d_S) for *CRTAC1* between human and mouse (0.477; 0.482 between mouse and chimp), however, is lower than that of prostate specific genes (mean = 0.540) and lies in between the mean of housekeeping genes (0.447) and overall tissue specific genes (0.492) (Zhang and Li 2004). A more recent study (Zhu et al. 2008) with more genes (3140 housekeeping genes and 885 tissue specific genes) has suggested similar human-mouse pairwise ω for housekeeping genes (mean = 0.09) and tissue-specific genes (mean = 0.21). *CRTAC1* human-mouse pairwise ω (0.13) falls between that of the housekeeping genes and tissue specific genes, suggesting it is evolving slowly than most tissue specific genes but evolving faster than most housekeeping genes. However, the

evolution of a certain gene can be highly clade, lineage and species specific. A gene that is evolving faster in one lineage can evolve slowly in another (Wang et al. 2011). So, I recommend calling CRTAC1 a slowly evolving gene among mammals and not generalizing the evolution rate of the gene.

3.4.3 Population Genetics of CRTAC1

Great Ape Genome Project (GAGP) database (Prado-Martinez et al. 2013) is an excellent addition to the available public databases. The authors have sequenced the entire genome of several individuals from the genera *Pan, Gorilla* and *Pongo*. The data is available as raw sequence files and variant calling files (VCFs). The VCF files contain the SNP and indel data of all great apes. I downloaded the VCF files for *Pan troglodytes, Pan paniscus, Gorilla gorilla, Gorilla beringei,* and *Pongo abelii*. I converted those file to usable formats using UNIX codes and Shell scripts. I found 19 SNPs in *Pan troglodytes* (11 non-synonymous), three SNPs in *Pan paniscus* (all synonymous), 12 SNPs in *Gorilla gorilla* (7 non-synonymous), and one SNP in *Gorilla beringei* (non-synonymous).

Pan troglodytes have four SNPs out of Hardy-Weinberg Equilibrium. Three of these are non-synonymous (See Results Section 3.3.11). For all four SNPs *Pan troglodytes verus* are fixed for the alternate allele. For the SNP position 97498542 (Exon 2) in chimpanzee all ten *P. t. ellioti* have the minor allele, two out of six *P. t. schweinfurthii* are heterozygous, one has the major allele, and one has the minor allele; three out of four *P. t. troglodytes* has the minor allele and one is heterozygous and all five *P. t. verus* have the minor allele. For SNP position 97404841 (exon 5), all *P. t. ellioti, P. t. schweinfurthii* and *P. t. troglodytes* have the minor allele, while all *P. t. verus* have the

major allele. For SNP position 97382256 (exon 10), nine *P. t. ellioti* have the minor allele and one is heterozygous, five *P. t. schweinfurthii* have the minor allele and one is heterozygous, all *P. t. troglodytes* have the minor allele, and all *P. t. verus* have the major allele. Finally, for SNP position 97366498 (exon 14), all *P. t. ellioti, P. t. schweinfurthii* and *P. t. troglodytes* have the minor allele, while all *P. t. verus* have the major allele. *P. t. verus CRTAC1* CDS is highly differentiated from the other three subspecies (F_{st} 0.64 – 0.75) compared to the differentiation of the other three subspecies among each other (F_{st} 0.17 – 0.25). *P. t. verus* got separated from *P. t. troglodytes* and *P. t. schweinfurthii* ~1 MYA (Bjork et al. 2010). So, *CRTAC1* CDS diversity among them is not unexpected. But the separation between *P. t. ellioti* and *P. t. verus* is comparatively newer ~500,000 years ago (Bjork et al. 2010). So, *CRTAC1* CDS diversity between them is rather surprising. This result reconfirms the usefulness of using multiple genes while building phylogenetic relationship among taxa since a single gene often does not reflect the true relationships among organisms.

3.4.4 Protein domains and potential function of CRTAC1

CRTAC1 has three functional domains. EGF-like calcium-binding domain in the C- terminal end (Redruello et al. 2010), FG-GAP domain that folds into a β-propeller structure, and UnbV_ASPIC conserved protein domain. The last two domains are found in integrin-like proteins. Vertebrate CRTAC1 was generated by combining UnbV_ASPIC and calcium binding EGF domains (EGF-CA) (Kawashima et al. 2009). Vertebrate CRTAC1 uniquely possess the calcium binding EGF domain. Echinoderms and Cephalochordates possess FG-GAP and UnbV_ASPIC domains but lack the EGF-CA domain (Keeley and Mecham 2013). The combination of FG-GAP and UnbV_ASPIC

probably took place in the common ancestors of all bilaterians. Later during vertebrate evolution the EGF-CA domain combined to the above-mentioned domains to generate the functional cartilage acidic protein (Keely and Mecham 2013).

As mentioned before, the function of CRTAC1 is unknown. But one can speculate its function based on its functional domain – the EGF-CA domain. EGF domain is one of the most abundant extracellular protein modules (Boswell et al. 2006). There are three known possible functions of EGF domains. EGF domain can be used as spacer units that provide right distance among other domains. This type of function is especially observed in blood coagulation factors, where two EGF domains are required to position the active site of the protease domain at the right distance from the cell membrane to aid in cofactor interaction and substrate activation (Husten et al. 1987, Brandstetter et al. 1995, Banner et al. 1996). Secondly, calcium binding EGF domains (EGF-CA) can help to increase the rigidity of the protein. As in case of fibrillin-1, EGF-CA plays a stabilizing role and helps to increase the rigidity of the protein (Cardy and Handford 1998). Third, EGF-CA can participate directly into protein-protein interactions in a Ca^{2+} dependent manner. EGF-like domains have been shown to interact with coagulation factors such as protein S, factor IX, factor X, and the low-density lipoprotein receptor (LDL receptor) (Stenflo et al. 2000). The EGF domain of Notch regulates the interaction of Notch with its ligands (Haltiwanger and Stanley 2002). Due to the presence of EGF-CA in CRTAC1, it can provide rigidity to the protein. Providing rigidity can be an important function of CRTAC1, especially in case of tissues like cartilage and lung alveoli. Being an extracellular matrix protein with EGF-CA, CRTAC1 can interact with other proteins including the coagulation factors.

In Integrin like proteins, FG-GAP domain has also shown to help in protein-protein interactions and aid in ligand binding in presence of Ca^{2+} (Springer 1997). Recently it has been shown that Integrin-α FG-GAP Repeat-Containing Protein 2 (Itfg2) plays important role in B-cell differentiation and helps in the development of autoimmunity (Al-Shami et al. 2013). Like many Integrin-like proteins, CRTAC1 also has FG-GAP repeat. It can be speculated that this domain serves as an interaction site with other proteins. Additionally, like Itfg2, this domain may aid in immune function. FG-GAP domain is also found in many leukocyte integrin family proteins such as Lymphocyte function-associated antigen 1 (LFA-1), which is found on all leukocytes. This protein is important in adhesive interactions in immune and inflammatory responses (Huang and Springer 1997). Like in LFA-1, FG-GAP domain in CRTAC1 can help in ligand binding and aid in immune responses. The immune function of CRTAC1 can be used to protect the sperm against parasites in female reproductive tract. Many human seminal plasma proteins such as interleukins (e.g. IL12) and tumor necrosis factors (e.g. TNFα) show this type of immune function (Hussenet et al. 1993, Huleihel et al. 1999). They provide protection against parasites, and also damage sperms from other males (Poiani 2006).

When the virtual cDNAs (See Methods) of human and chimpanzee are translated into proteins, only four amino acid differences were observed between human and chimpanzee CRTAC1. None of these amino acid differences fall within any of the above-mentioned domains. However, the *in vivo* scenario can be completely different. As mentioned before *CRTAC1* has seven splice variants. Two of these variants (ENST00000298819 and ENST00000413387) has spliced out exon 13 and thus have lost

the EGF like Ca^{+2} binding domain. The FG-GAP domain, however, is present in all splice variants. It is possible that CRTAC1 is expressed in tissue specific and/or species-specific manner, which may result in *in vivo* differences in protein domains. It can be speculated that the ligand binding ability of FG-GAP domain is utilized in all tissues but the protein-protein interaction function of EGF-CA domain is only utilized in very few tissues.

3.4.5 CRTAC1 is potentially a housekeeping gene

The genes that are highly expressed in many tissues and are found in the gene-dense region of the human genome are known as 'housekeeping genes' (Caron et al. 2001, Lercher et al. 2002). The regulation of these highly expressed genes often include long-range (up to 150 kb) looping interactions between promoters and enhancers, and require proper repositioning of the target loci (Noordermeer et al. 2008). Housekeeping genes are generally expressed in many human tissues (>18) (Vinogradov 2004) and their CDS tend to evolve much slower than the tissue specific genes (Zhang and Li 2004). Housekeeping genes have shown to be much older than the tissue specific genes with a highly GC-rich core promoter and have multiple GC boxes bound by the transcription factor SP1 (Zhu et al. 2008). Housekeeping genes have CpG dependent core promoters, which can initiate transcription in both TATA Box dependent and independent manner (Zhu et al. 2008). Another sticking feature of housekeeping genes is their size. The genomic length (mean = 28,792bp), transcript length (mean = 2801bp), CDS length (mean = 1380bp), and number of exons (mean = 11) of the housekeeping genes all tend to be larger than the tissue specific genes (genomic length mean = 7191bp, transcript

length mean = 1601bp, CDS length mean = 963, and mean number of exons = 4) (Zhu et al. 2008).

Considering the above-mentioned parameters, *CRTAC1* looks more like a housekeeping gene than a tissue specific gene. *CRTAC1* is found in a gene-dense region of the genome and actually shares its last exon with *GOLGA7B*. It is expressed in at least 20 tissues (EST Database) and is highly conserved among the metazoans with very little sequence divergence (Redruello et al. 2010). *CRTAC1* has 15 exons with a CDS length of 1911bp, both correspond to the mean values for housekeeping genes. The genomic length (165,829bp) and transcript length of *CRTAC1* are higher than the mean values for the housekeeping genes. The putative promoter region of *CRTAC1* is highly GC rich with several GC different boxes (CGG, CCG, GC repeats) present in this area. There are several SP1 target sites in the putative promoter region of *CRTAC1*. The putative TATA Box is surrounded by GC repeats but the region just upstream of the TATA box (~100bp) is free from them. All of these characters correspond to the CpG mediated TATA dependent promoters, which is a character of a housekeeping gene. Also the human promoter of *CRTAC1* shows only 12% sequence divergence from the mouse promoter, which is much lower than the mean for the housekeeping genes between human and mouse (35%) (Zhu et al. 2008). Finally, as mentioned before the CDS of *CRTAC1* is highly conserved among hominids and shows the operation of strong purifying selection with an overall human-mouse ω (D_N/D_S) of 0.13 much lower than that of the mean for the tissue-specific genes between human and mouse (0.21). Also, *CRTAC1* promoter shows similar expression pattern in both prostate cell line (LNCaP cell line) and osteoblast cell line (MG63) (See Results), which supports its 'housekeeping' nature. A recent study

(Altintas et al. 2013) has shown that seminal-plasma specific genes such as prostate specific antigen (*PSA*), and Kallikrein related peptide 2 (*KLK2*) respond to synthetic androgen (R1881) stimulation in the prostate tissue, but the 'housekeeping genes' such as Hydroxymethylbilane synthase (*HMBS*), Ribosomal protein 13a (*RPL13A*), and Ubiquitin B (*UBB*) do not show any over/under expression in the presence of the same. When transfected into the prostate cell line (LNCaP cells), like other housekeeping genes, *CRTAC1* is not affected with androgen stimulation, which further suggests that it is not a seminal plasma specific gene but more like a housekeeping gene.

All of the above-mentioned characters suggest that CRTAC1 is more like a housekeeping gene than a tissue specific gene and this makes its regulation much more complicated (Noordermeer et al. 2008) and *CRTAC1* promoter may be involved in long-range interaction with distant activators or repressors which may be >150kb away from the gene.

3.4.6 Housekeeping genes can get up/down regulated in certain tissues, under certain conditions

It has been shown that housekeeping genes can show differential expression among tissues, cell lines, and disease state (Huggett et al. 2005). This finding form a RT-PCR study, however, is not new. It has been shown previously that hypoxanthine-guanine phosphoribosyl transferase (*HPRT*), which is generally considered as a housekeeping gene, is constitutively expressed in most human tissues but is over-expressed in central nervous system (Stout et al. 1985). Another common 'housekeeping gene' β-actin was found differentially expressed among different leukaemia patients (Blomberg et al. 1987). Glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*), another commonly referred

housekeeping gene, was found to generate different amounts of mRNAs during transcription in different rat tissues (Piechaczyk et al. 1984).

A more recent paper has shown another interesting character of housekeeping genes. The housekeeping genes sometimes get down regulated in certain tissues due to functional reasons (Thorrez et al. 2011). This so-called "tissue-specific disallowance" of housekeeping genes has been shown for 13 genes in mouse with a fold change of >30 in all cases. Among these 13 identified genes five were only down regulated in testes (*Malat1, Lpl, Atrx, Stag2, Birc4*), five were only down regulated in liver (*Oxct1, Scd2, Enah, Slc25a4, Tspan13*), and three were only down regulated in pancreatic islets (*Mct1, Maf, Ldha*). As mentioned before, functional reasons are involved in this tissue specific down regulation of housekeeping genes. For example, if not repressed in the beta cells of pancreatic islets, MCT1 expression may cause hypoglycemia after physical exercise due to improper release of insulin (Otonkoski et al. 2003). OXCT1 is an enzyme that helps in ketone body degradation (which provides an alternative energy source). Liver is the major storehouse of ketone that supplies ketone to other tissues and supports their metabolism (Cahill and Veech 2003). So, liver specific down regulation of OXCT1 protects the ketones, stored in liver, from being degraded. ATRX plays important role in the sex differentiation during early embryogenesis and helps in the early development of testes. But in adults ATRX is found in very small amount in Leydig cells and plays a housekeeping role of maintenance of spermatogenesis (Tang et al. 2011). So, in this case the tissue specific down regulation is more temporal where a gene plays essential role in early life, gets down regulated in later stages. Various epigenetic histone modifications

(H3K27Me3, H3K9Ac) and/or tissue specific microRNAs are involved in the tissue specific down regulations of the housekeeping genes (Thorrez et al. 2011).

An opposite incidence regarding the expression of housekeeping gene has also been reported. In this case a so-called housekeeping gene *GABPα*, is over expressed in the fibroblast cells of human Down syndrome patients (O'Leary et al. 2004). GABPα together with GABPβ forms the functional GABP transcription factor complex and thus plays an essential housekeeping role (O'Leary et al. 2004). However, due to an increased gene dosage GABPα is expressed ~2 fold higher in Down syndrome patients compared to healthy individuals (O'Leary et al. 2004).

As discussed above housekeeping genes can show over/under expression in certain tissue or in certain individuals under certain conditions. One can speculate that a similar incident is happening for *CRTAC1*. An otherwise housekeeping gene, *CRTAC1*, is getting overexpressed in chimpanzee seminal plasma for some functional reasons. As discussed in section 3.4.4, CRTAC1 can serve as a coagulation factor and/or help in immune function. Chimpanzee has higher sperm competition than human (See Introduction Section 3.1.2) and so it may require the autoimmune function of CRTAC1 to damage or kill the competitor's sperms present in female reproductive tract. Humans with potentially lower sperm competition may not require this function any more, and so CRTAC1 expression may have been down regulated in human seminal plasma. Alternatively chimpanzee can also utilize the coagulatory function of CRTAC1. Chimpanzee seminal plasma is more viscous and forms rigid copulatory plug soon after ejaculation. CRTAC1 can provide rigidity to the plug and can potentially help in the formation of the plug by interacting with other proteins. Humans, unlike chimpanzee, do

not form any copulatory plug and have less viscous sperm. So, CRTAC1 have been down regulated in human seminal plasma in a tissue-specific disallowance manner due to its lack of importance in this tissue.

3.4.7 Concluding remarks: a note on the apparent anomaly between proteomic data and luciferase assay

Mass Spectrometry study showed CRTAC1 is expressed ~142 fold higher in chimpanzee seminal plasma compared to human and this result was supported by Western Blot analysis, showing CRTAC1 expression ~4 fold higher in chimpanzee seminal plasma compared to human. Paradoxically the *in vitro* luciferase assay showed that human putative promoter drives transcription highly significantly ($P<0.01$) than chimpanzee.

Although seems contradictory, this kind of results have been encountered before while comparing human and chimpanzee genes. A previous *in vitro* luciferase assay showed that the putative chimpanzee promoter of Golgi SNAP Receptor Complex Member 1 (*GOSR1*) drives transcription significantly higher ($P<0.05$) compared to human promoter in the human liver line (HEP) (Chabot et al. 2007). Paradoxically, *GOSR1* mRNA is highly expressed in human liver compared to chimpanzee, detected by microarrays and confirmed by quantitative RT-PCR (Chabot et al. 2007). Out of the 10 genes they studied for both RT-PCR and *in vitro* luciferase assays, only three genes showed similar direction of expression in both. Chabot et al. (2007) concluded that this anomaly might be due to the compensatory changes in the *cis*-regulatory region, which was not part of the ~1kb DNA segment they used as the putative promoter in the luciferase assay. Compensatory changes in the transcription factor binding sites have

been observed previously among fruit flies (Ludwig et al. 2005) and between humans and mice (Dermitzakis and Clark 2002). Similar results were obtained by Heissig et al. (2005) while comparing the promoter activity and the gene expression of 12 genes, between human and chimpanzee. They found four (*CGI-51, SH3BGR, UNG, TERF*) out of seven genes that showed significant difference in promoter activity, drove transcription to the opposite direction from what was expected by gene-expression study. Like Chabot et al. (2007), they also concluded the possibility of the presence of additional regulatory elements that aid in the final outcome of the gene-expression.

CRTAC1 possibly behaves like the above-mentioned genes and may involve additional tissue specific and/or species-specific regulatory elements during its regulation. As mentioned before this regulatory elements can be far away (up to 150kb) from the gene itself. Additionally there are many non-coding RNAs such as miRNAs, linkRNAs, and eRNAs may be involved in gene regulation in tissue specific manner. So, it may not be possible to identify the actual regulatory regions that are responsible for CRTAC1 expression. Finally, although not completely conclusive in determining why chimpanzee seminal plasma expresses CRTAC1 in such high amount compared to humans, this study threw light on the complexity of eukaryotic gene regulation, which in most cases, can not be explained by the simple promoter and/or promoter + single enhancer mediated luciferase assays.

References

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ 2004. Bmp4 and morphological variation of beaks in Darwin's finches. Science 305:1462-1465.

Al-Shami A, Crisostomo J, Wilkins C, Xu N, et al. 2013. Integrin-α FG-GAP repeat-containing protein 2 is critical for normal B cell differentiation and controls disease development in a lupus model. J Immunol. 191:3789-3798.

Altintas DM, Allioli N, Decaussin M, et al. 2013. Differentially Expressed Androgen-Regulated Genes in Androgen-Sensitive Tissues Reveal Potential Biomarkers of Early Prostate Cancer. PLoS ONE 6:e66278.

Anderson MJ, Chapman SJ, Videan EN, Evans E, Fritz J, Stoinski TS, Dixson AF, Gagneux P 2007. Functional evidence for differences in sperm competition in humans and chimpanzees. Am J Phys Anthropol. 134:274-280.

Anderson MJ, Dixson AF 2002. Motility and the midpiece in primates. Nature 416:496.

Babbitt CC, Silverman JS, Haygood R, Reininga JM, Rockman MV, Wray GA 2009. Multiple functional variants in cis modulate PDYN expression. Mol Biol Evol. 27:465-479.

Banerjee A, Northrup D, Boukarabila H, Jacobsen SEW, Allman D 2013. Transcriptional Repression of Gata3 Is Essential for Early B Cell Commitment. Immunity 38:930-942.

Banner DW, D'Arcy A, Chene C, Winkler FK, Guha A, Konigsberg WH, Nemerson Y, Kirchhofer D 1996. The crystal structure of the complex of blood coagulation factor VIIa with soluble tissue factor. Nature 380:41-46.

Berger SL 2000. Gene regulation. Local or global? Nature 408:412-413.

Birtle Z, Goodstadt L, Ponting C 2005. Duplication and positive selection among hominin-specific PRAME genes. BMC Genom. 6:120.

Bjork A, Liu W, Wertheim JO, Hahn BH, Worobey M 2011. Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. Mol Biol Evol. 28:615-623.

Blomberg J, Andersson M, Faldt R 1987. Differential pattern of oncogene and beta-actin expression in leukaemic cells from AML patients. Br J Haematol. 65:83-86.

Blum R, Vethantham V, Bowman C, et al. 2012. Genome-wide identification of enhancers in skeletal muscle: the role of MyoD1. Genes Dev. 26:2763-2779.

Boswell E, Kurniawan ND, Downing AK. 2006. Calcium-binding EGF-like domains: Wiley Online Library.

Bradley BJ, Doran-Sheehy DM, Lukas D, Boesch C, Vigilant L 2004. Dispersed male networks in western gorillas. Curr Biol. 14:510-513.

Brandstetter H, Bauer M, Huber R, Lollar P, Bode W 1995. X-ray structure of clotting factor IXa: active site and module structure related to Xase activity and hemophilia B. Proc Natl Acad Sci. 92:9796-9800.

Brillard-Bourdet M, Réhault s JLFMMTGF 2002. Amidolytic activity of prostate acid phosphatase on human semenogelins and semenogelin-derived synthetic substrates. Eur J Biochem. 269:390-395.

Buckland PR, Coleman SL, Hoogendoorn B, Guy C, Smith SK, et al. 2004a. A high proportion of chromosome 21 promoter polymorphisms influence transcriptional activity. Gene Expr. 11:233-239.

Buckland PR, Hoogendoorn B, Guy CA, Coleman SL, Smith SK, et al. 2004. A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity. Biochim Biophys Acta. 1690:238-249.

Cahill GF Jr VR 2003. Ketoacids? Good medicine? Trans Am Clin Climatol Assoc. 114:149-161.

Cardy CM, Handford PA 1998. Metal ion dependency of microfibrils supports a rod-like conformation for fibrillin-1 calcium-binding epidermal growth factor-like domains. J Mol Biol. 276:855-860.

Carnahan SJ, Jensen-Seaman MI 2008. Hominoid seminal protein evolution and ancestral mating behavior. Am J Primatol. 70:939-948.

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, et al. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291:1289-1292.

Chabot A, Shrit RA, Blekhman R, Gilad Y 2007. Using reporter gene assays to identify cis regulatory differences between humans and chimpanzees. Genetics 176:2069-2076.

Choi JS, Kim JS, Joe CO, Kim S, Ha KS, Park YM 1999. Improved cycle sequencing of GC-rich DNA template. Exp Mol Med. 31:20-24.

Clark NL, Swanson WJ 2005. Pervasive adaptive evolution in primate seminal proteins. PLoS Genet. 1:e35.

Clark RM, Wagler TN, Quijada P, Doebley J 2006. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet. 38:594-597.

Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, et al. 2013. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 24:1-13.

Cresko WA, Amores A, Wilson C, Murphy J, Currey M, et al. 2004. Parallel genetic basis for repeated evolution of armor loss in Alaskan three spine stickleback populations. Proc Natl Acad Sci. 101:6050-6055.

Darwin C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, Albemarle Street: London.

Darwin C. 1871. The descent of man and selection in relation to sex. John Murray, Albemarle Street: London.

de Lamirande E 2007. Semenogelin, the main protein of the human semen coagulum, regulates sperm function. Semin Thromb Hemost. 33:60-68.

Dermitzakis ET, Clark AG 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol. 19:1114-1121.

Dixson AF 1993. Sexual selection, sperm competition and the evolution of sperm length. Folia Primatol. 61:221-227.

Dixson AF 1998. Sexual selection and evolution of the seminal vesicles in primates. Folia Primatol. 69:300-306.

Dixson AF. 2012. Primate sexuality: comparative studies of the prosimians, monkeys, apes, and humans. New York, NY: Oxford University Press.

Dixson AL, Anderson MJ 2002. Sexual selection, seminal coagulation and copulatory plug formation in primates. Folia Primatol. 2:63-69.

Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat Genet. 36:1326-1329.

Dubbink HJ, de Waal L, van Haperen R, Verkaik NS, Trapman J, Romijn JC 1998. The human prostate-specific transglutaminase gene TGM4: genomic organization, tissuespecific expression, and promoter characterization. . Genomics 51:434-444.

Farre D, Bellora N, Mularoni L, Messeguer X, MarAlba M 2007. Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol. 8:R140.

Fernández M M-SD 2012. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic Acids Res. 40:e77.

Fleagle JG. 1999. Primate adaptation and evolution. San Diego, CA: Academic Press.

Flicek P, et al. 2013. Ensembl 2013. Nucleic Acids Res. 41:D48-D55.

Ford MJ 2001. Molecular evolution of transferrin: evidence for positive selection in salmonids. Mol Biol Evol. 18:639-647.

Frey UH, Bachmann HS, Peters J, Siffert W 2008. PCR-amplification of GC-rich regions:'slowdown PCR'. Nat Protoc. 3:1312-1317.

Fung KYC, Glode LM, Green S, Duncan MW 2004. A comprehensive characterization of the peptide and protein constituents of human seminal fluid. Prostate 61:171-178.

Funnell AP, Norton LJ, Mak KS, Burdach J, et al. 2012. The CACCC-binding protein KLF3/BKLF represses a subset of KLF1/EKLF target genes and is required for proper erythroid maturation in vivo. Mol Cell Biol. 32:3281-3292.

Gaubin M, Autiero M, Basmaciogullari S, Metivier D, Mis hal Z, et al. 1999. Potent inhibition of CD4/TCR-mediated T cell apoptosis by a CD4-binding glycoprotein secreted from breast tumor and seminal vesicle cells. J Immunol. 162:2631-2638.

Gibbs RA, Rogers J, Katze MG, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222-234.

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature 433:481-487.

Hahn MW 2007. Detecting natural selection on cis regulatory DNA. Genetica 129:7-18.

Hahn MW 2008. Toward a selection theory of molecular evolution. Evolution 62:255-265.

Haltiwanger RS, Stanley P 2002. Modulation of receptor signaling by glycosylation: fringe is an O-fucose-beta1,3-N-acetylglucosaminyltransferase. Biochim Biophys Acta. 1573:328-335.

Hammock EA, Young LJ 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308:1630-1634.

Harcourt AH, Harvey PH, Larson SG, Short RV 1981. Testis weight, body weight and breeding system in primates. Nature 293:55-57.

Harvey PaH, AH 1984. Sperm competition, testes size, and breeding systems in primates. In: Smith RL, editor. Sperm competition and the evolution of animal mating systems. Orlando: Academic Press.

Hasegawa T, Hiraiwa-Hasegawa M. 1990. Sperm competition and mating behavior. In: Nishida T, editor. The chimpanzees of the Mahale mountains. Tokyo: University of Tokyo Press.

Haudek SB, Natmessnig BE, Redl H, Schlag G, Giroir BP 1998. Genetic sequences and transcriptional regulation of the TNFA promoter: comparison of human and baboon. Immunogenetics 48:202-207.

Haygood R, Babbitt CC, Fedrigo O, Wray GA 2010. Contrasts between adaptive coding and noncoding changes during human evolution. Proc Natl Acad Sci. 107:7853-7857.

Heissig F, Krause J, Bryk J, Khaitovich P, Enard W, Pääbo S 2005. Functional analysis of human and chimpanzee promoters. Genome Biol. 6:R57.

Henke W, Herdel K, Jung K, Schnorr D, Loening SA 1997. Betaine improves the PCR amplification of GC-rich DNA sequences. Nucleic Acids Res. 25:3957-3958.

Huang C, Springer T 1997. Folding of the β-propeller domain of the integrin αL subunit is independent of the I domain and dependent on the β2 subunit. Proc Natl Acad Sci. 94:3162-3167.

Huang Z, Xu H, Sandell L 2004. Negative Regulation of Chondrocyte Differentiation by Transcription Factor AP-2. Bone Mineral Res J. 19:245-255.

Huby T, Dachet C, Lawn RM, Wickings J, Chapman MJ 2001. Functional analysis of the chimpanzee and human apo(a) promoter sequences:identification of sequence variations responsible for elevated transcriptional activity in chimpanzee. J Biol Chem. 276:22209-22214.

Hudson RR, Kreitman M, Aguad 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.

Huggett J, Dheda K, Bustin S, Zumla A 2005. Real-time RT-PCR normalisation; strategies and considerations. Genes Immun. 6:279-284.

Huleihel M, Lunenfeld E, Horowitz S, Levy A, Potashnik G MMGM 1999. Expression of IL-12, IL-10, PGE2, sIL-2R and sIL-6R in seminal plasma of fertile and infertile men. Andrologia 31:283-288.

Hussenet F, Dousset B, Cordonnier JL, Jacob C, Foliguet B, Grignon G, Nabet P 1993. Tumour necrosis factor alpha and interleukin 2 in normal and infected human seminal fluid. Hum Reprod. 8:409-411.

Husten EJ, Esmon CT, Johnson AE 1987. Its distance from the phospholipid surface and its conformational sensitivity to components of the prothrombinase complex. J Biol Chem. 262:12953-12961.

Inoue-Murayama M, Mishima N, Hayasaka I, Ito S, Murayama Y 2006. Divergence of ape and human monoamine oxidase A gene promoters: Comparative analysis of polymorphisms, tandem repeat structures and transcriptional activities on reporter gene expression. Neurosci Lett 405:207-211.

Iwagawa T, Tanaka Y, Lida A, Itoh T, Watanabe S 2013. Enhancer/Promoter Activities of the Long/Middle Wavelength-Sensitive Opsins of Vertebrates Mediated by Thyroid Hormone Receptor β2 and COUP-TFII PLoS ONE 8:e72065.

Jensen-Seaman MI, Li WH 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. J Mol Evol. 57:261-270.

Kang J, Lee MS, Gorenstein DG 2005. The enhancement of PCR amplification of a random sequence DNA library by DMSO and betaine: application to in vitro combinatorial selection of aptamers. J Biochem Bioph Meth. 64:147-151.

Kano T. 1992. The last ape: pygmy chimpanzee behavior and ecology. Stanford, CA: Stanford University Press.

Kawashima T, Kawashima S, Tanaka C, et al. 2009. Domain shuffling and the evolution of vertebrates. Genome Res. 19:1393-1403.

Keeley F, Mecham R. 2013. Evolution of extracellular matrix: Springer Publications.

Kim TK, Hemberg M, Gray JM, Costa AM, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182-187.

Kimura M. 1983. The neutral theory of molecular evolution: Cambridge University Press.

Kingan SB, Tatar M, Dm 2003. Reduced polymorphism in the chimpanzee semen coagulating protein Semenogelin I. J Mol Evol. 57:159-169.

Kingan SB TM, Rand DM 2003. Reduced polymorphism in the chimpanzee semen coagulating protein Semenogelin I. J Mol Evol. 57:159-169.

Kouprina N, Mullokandov M, Rogozin IB, et al. 2004. The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids. Proc Natl Acad Sci. 101:3077-3082.

Kreitman M 2000. Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet. 1:539-559.

Latham JP, Searl PF, Mautner V, et al. 2000. Prostate-specific antigen promoter/enhancer driven gene therapy for prostate cancer: construction and testing of a tissue-specific adenovirus vector. Cancer Res. 60:334-341.

Lercher MJ, Urrutia AO, Hurst LD 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet. 31:180-183.

Li B, Carey M, Workman JL 2007. The role of chromatin during transcription. Cell 128:707-719.

Lilja H 1985. A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. J Clin Invest. 76:1899-1903.

Lilja H, Oldbring J, Rennevik G, Laurell CB 1987. Seminal vesicle-secreted proteins and their reactions during gelation and liquefaction of human semen. J Clin Invest. 80:281-285.

Lin HJ, Luo CW, Chen YH 2002. Localization of the transglutaminase cross-linking site in SVS III, a novel glycoprotein secreted from mouse seminal vesicle. J Biol Chem. 277:3632-3639.

Loisel DA, Rockman MV, Wray GA, Altmann J, Alberts SC 2006. Ancient polymorphism and functional variation in the primate MHC-DQA1 59 cis-regulatory region. Proc Natl Acad Sci. 103:16331-16336.

Lövgren J AKLH 1999. Enzymatic action of human glandular kallikrein 2 (hK2): substrate specificity and regulation by Zn2+ and extracellular protease inhibitors. Eur J Biochem. 262:781-789.

Low SB. 2007. Ecological and socio-cultural impacts on mating and marriage systems. In: Dunbar R, Barret L, editors. The Oxford handbook of evolutionary psychology.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, et al. 2005. Functional evolution of a cis-regulatory module. PLoS Biol. 3:e93.

Lukk M, Kapushesky M, Nikkil¨a JPHGAHWUEBA 2010. A global map of human gene expression. Nat Biotechnol. 28:322-324.

Lundwall Å PA, Lövgren J, Lilja H, Malm J 1997. Chemical characterization of the predominant proteins secreted by mouse seminal vesicles. Eur J Biochem. 249:39-44.

Maggioncalda AN, Czekala NM, Sapolsky RM 2002. Male orangutan subadulthood:a new twist on the relationship between chronic stress and developmental arrest. Am J Phys Anthropol. 118:25-32.

Malm J, Hellman J, Magnusson H, Laurell CB, Lilja H 1996. Isolation and characterization of the major gel proteins in human semen, semenogelin I and semenogelin II. Eur J Biochem. 238:48-53.

Mamedov TG, Pienaar E, Whitney SE, TerMaat JR, Carvill G, Goliath R, Subramanian A, Viljoen HJ 2008. A fundamental study of the PCR amplification of GC-rich DNA templates. Comput Biol Chem. 32:452-457.

Martin DE, Gould KG. 1977. The male ape genital tract and its secretions. In: Graham CE, editor. Reproductive biology of the great apes: comparative and biomedical. New York: Academic Press.

Maston GA, Evans SK, Green MR 2006. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet. 7:29-59.

McDonald JH, Kreitman M 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.

McDowell DG, Burns NA, Parkes HC 1998. Localised sequence regions possessing high melting temperatures prevent the amplification of a DNA mimic in competitive PCR. Nucleic Acids Res. 26:3340-3347.

Mitani JC 1985. Sexual selection and adult male orangutan long calls. Anim Behav. 33:272-283.

Møller AP 1988. Ejaculate quality, testes size and sperm competition in primates. J Hum Evol. 17:479-488.

Møller Ap BTR 1989. Copulation behavior in mammals: evidence that sperm competition is wide-spread. Biol J Linn Soc. 38:119-131.

Musso M, Bocciardi R, Parodi S, Ravazzolo R, Ceccherini I 2006. Betaine, Dimethyl Sulfoxide, and 7-Deaza-dGTP, a Powerful Mixture for Amplification of GC-Rich DNA Sequences. J Mol Diagn. 8:544-550.

Nascimento JM, Shi LZ, Meyers S, Gagneux P, Loskutoff NM, Botvinick EL, Berns MW 2008. The use of optical tweezers to study sperm competition and motility in primates. J R Soc Interface 5:297-302.

Nielsen R 2001. Statistical tests of selective neutrality in the age of genomics. Heredity 86:641-647.

Nielsen R, Yang Z 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Noordermeer D, Branco MR, Splinter E, Klous P, van Ijcken W, et al. 2008.

Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region. PLoS Genet. 4:e1000016.

O'Leary DA, Pritchard MA, Xu D, Kola I, Hertzog PJ, Ristevski S 2004. Tissue-specific overexpression of the HSA21 gene GABPalpha: implications for DS. Biochim Biophys Acta. 1739:81-87.

Oda A 1999. Female choice in the opportunistic mating of wild chimpanzees (Pan troglodytes schweinfurthii) at Mahale. Behav Ecol Sociobiol. 46:258-266.

Ong CT, Corces VG 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 12:283-293.

Orlov SV, Kuteĭkin-Tepliakov Kb, Grishin A. V. Dizhe E. B. Prokhorchuk E. B. Perevozchikov A. P. 2006. Transcription factor ZF5 regulates expression of mammalian gene containing GCC-triplet repeats in 5'-regulatory region in human hepatoma HepG2 cells. Tsitologiia 48:246-252.

Orom UA, Derrien T, Beringer M, Gumireddy K, et al. 2010. Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46-58.

Orphanides G, Reinberg D 2002. A unified theory of gene expression. Cell 108:439-451.

Otonkoski T, Kaminen N, Ustinov J, Lapatto R, et al. 2003. Physical exercise-induced hyperinsulinemic hypoglycemia is an autosomal-dominant trait characterized byabnormal pyruvate-induced insulin release. Diabetes 52:199-204.

Parker GA 1970. Sperm competition and its evolutionary consequences in the insects. Biol Rev. 45:525-567.

Perry MW, Boettiger AN, Levine M 2011. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. Proc Natl Acad Sci. 108:13570-13575.

Piechaczyk M, Blanchard JM, Marty L, et al. 1984. Post-transcriptional regulation of glyceraldehyde-3-phosphate-dehydrogenase gene expression in rat tissues. Nucleic Acids Res. 12:6951-6963.

Pilch B, Mann M 2006. Large-scale and high-confidence proteomic analysis of human seminal plasma. Genome Biol. 7:R40.

Podlaha O, Zhang J 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. Proc Natl Acad Sci. 100:12241-12246.

Poiani A 2006. Complexity of seminal fluid: a review. Behav Ecol Sociobiol. 60:289-310.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, et al. 2013. Great ape genetic diversity and population history. Nature 499:471-475.

Project Consortium E, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

Pugh BF 2000. Control of gene expression through regulation of the TATA-binding protein. Gene 255:1-14.

Raiber EA, Kranaster R, Lam E, Nikan M, Balasubramanian S 2012. A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. Nucleic Acids Res. 40:1499-1508.

Ralser M, Querfurth R, Warnatz HJ, Lehrach H, Yaspo ML, Krobitsch S 2006. An efficient and economic enhancer mix for PCR. Biochem Biophys Res Commun. 347:747-751.

Redruello B, Louro B, Anjos L, Silva N, Greenwell RS, Canario AV, Power DM 2010. CRTAC1 homolog proteins are conserved from cyanobacteria to man and secreted by the teleost fish pituitary gland. Gene 456:1-14.

Robbins MM 1999. Male mating patterns in wild multimale mountain gorilla groups. Anim Behav. 57:1013-1020.

Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB 2005. Ancient and recent positive selection transformed opioid cis-regulation in humans. PLoS Biol. 3:e387.

Ross MT, LaBrie S, McPherson J, Stanton Jr VP. 1999. Screening large-insert libraries by hybridization. In: A B, editor. Current protocols in human genetics: New York: Wiley. p. 5. 6.1-5.6.52.

Sahdev S, Saini S, Tiwari P, Saxena S, Singh Saini K 2007. Amplification of GC-rich genes by following a combination strategy of primer design, enhancers and modified PCR cycle conditions. Mol Cell Probes 21:303-307.

Schagat T, Paguio A, Kopish K, Promega C 2007. Normalizing genetic reporter assays: approaches and considerations for increasing consistency and statistical significance. Cell Notes 17:9-12.

Schuchard M, Sarkar G, Ruesink T, Spelsberg TC 1993. hot. PCR amplification of GC-rich avian c-myc sequences. Biotechniques 14:390-394.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, et al. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature

428:717-723.

Shifera AM, Hardin JA 2010. Factors modulating expression of Renilla luciferase from control plasmids used in the luciferase reporter gene assays. Anal Biochem. 396:167-172.

Smale T. 2010. Luciferase assay In.

Spivakov M, Akhtar J, Kheradpour P, et al. 2012. Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol. 13:R49.

Springer TA 1997. Folding of the N-terminal, ligand-binding region of integrin alpha-subunits into a beta-propeller domain. Proc Natl Acad Sci. 94:65-72.

Steck E, Braun J, Pelttari K, Kadel S, Kalbacher H, Richter W 2007. Chondrocyte secreted CRTAC1: a glycosylated extracellular matrix molecule of human articular cartilage. Matrix Biol. 26:30-41.

Stern DL 1998. A role of Ultrabithorax in morphological differences between Drosophila species. Nature 396:463-466.

Storgaard T, Christensen J, Aasted B, Alexandersen S 1993. cis-acting sequences in the Aleutian mink disease parvovirus late promoter important for transcription:comparison to the canine parvovirus and minute virus of mice. J Virol. 67:1887-1895.

Stout JT, Chen HY, Brennand J, Caskey CT, Brinster RL 1985. Expression of human HPRT in the central nervous system of transgenic mice. Nature 317:250-252.

Sun Y, Hegamyer G, Colburn NH 1993. PCR-direct sequencing of a GC-rich region by inclusion of 10% DMSO. Application to mouse c-jun. Biotechniques 15:372-374.

Swanson WJ, Nielsen R, Yang Q 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol. 20:18-20.

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci. 98:2509-2514.

Tajima F 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123:585-595.

Takahashi H, Mitani Y, Satoh G, Satoh N 1999. Evolutionary alterations of the minimal promoter for notochord-specific Brachyury expression in ascidian embryos. Development 126:3725-3734.

Tang P, Argentaro A, Pask AJ, O'Donnell L, Marshall-Graves J, Familari M, Harley VR 2011. Localization of the chromatin remodelling protein, ATRX in the adult testis. J

Reprod Dev. 57:317-321.

Thorrez L, Laudadio I, Van Deun K, Quintens R, et al. 2011. Tissue-specific disallowance of housekeeping genes: the other face of cell differentiation. Genome Res. 21:95-105.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. The accessible chromatin landscape of the human genome. Nature 489:75-82.

Tilson RL 1981. Family Formation Strategies of Kloss Gibbons Hylobates klossii. Folia Primatol. 25:259-287.

Timblin GA, Schlissel MS and c-Myb repress rag transcription downstream of Stat5 during early B cell development. J Immunol. 191:4676-4687.

Tindall EA, Petersen DC, Woodbridge P, Schipany K, Hayes VM 2009. Assessing high-resolution melt curve analysis for accurate detection of gene variants in complex DNA fragments. Hum Mutat. 30:876-883.

Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM 2003. Identification and functional analysis of human transcriptional promoters. Genome Res. 13:308-312.

Utami SS, Goossens B, Bruford MW, de Ruiter JR, van Hooff J 2002. Male bimaturism and reproductive success in Sumatran orangutans. Behav Ecol. 13:643-652.

Vadnais C, Awan AA, Harada R, Clermont PL, et al. 2013. Long-range transcriptional regulation by the p110 CUX1 homeodomain protein on the ENCODE array. BMC Genom. 14:258.

van Vliet J, Turner J, Crossley M 2000. Human Krüppel-like factor 8:a CACCC-box binding protein that associates with CtBP and represses transcription. Nucleic Acids Res. 28:1955-1962.

Venters BJ, Pugh FB 2009. How eukaryotic genes are transcribed. Crit Rev Biochem Mol Biol. 44:117-141.

Vinogradov AE 2004. Compactness of human housekeeping genes: selection for economy or genomic design. Trends Genet. 20:248-253.

Voight BF, Kudaravalli S, Wen XQ, Pritchard JK 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:446-458.

Wang D, Liu F, Wang L, Huang S, Yu J 2011. Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. Biol Direct. 6:13.

Wang Y, Wang Q 1999. Coexistence of monogamy and polygyny in Black-crested Gibbon (Hylobates concolor). Primates 40:607-611.

Watts DP 1990. Ecology of gorillas and its relationship to female transfer in mountain gorillas. Int J Primatol. 11:21-45.

Wei M, Deng J, Feng K, Yu B, Chen Y 2010. Universal method facilitating the amplification of extremely GC-rich DNA fragments from genomic DNA. Anal Chem. 82:6303-6307.

Williamson SH, Hubisz MJ, Clark AG, et al. 2007. Localizing recent adaptive evolution in the human genome. PLoS Genet. 3:e90.

Wolff C, Pepling M, Gergen P, Klingler M 1999. Structure and evolution of a pair-rule interaction element: runt regulatory sequences in D. melanogaster and D. virilis. Mech Dev. 80:87-99.

Wray GA, Hahn Mw, Abouheif E. Balhoff J. P. Pizer M. Rockman M. V. Romano L. A. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 20:1377-1419.

Yang Z 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol. 51:423-432.

Yang Z 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586-1591.

Yang Z, Bielawski JP 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496-503.

Yang Z, Nielsen R 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 46:409-418.

Yang Z, Nielsen R 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908-917.

Zelezetsky I, Pontillo A, Puzzi L, et al. 2006. Correlation between structural variations and antimicrobial activity. J Biol Chem. 281:19861-19871.

Zhang L, Li WH 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol. 21:236-239.

Zhang Z, Song Y, Zhao X, Zhang X, Fermin C, Chen Y 2002. Rescue of cleft palate in Msx1-deficient mice by transgenic Bmp4 reveals a network of BMP and Shh signaling in the regulation of mammalian palatogenesis. Development 129:4135-4146.

Zhu J, He F, Hu S, Yu J 2008. On the nature of human housekeeping genes. Trends Genet. 24:481-484.

Chapter 4: Evolution of miRNAs and their targets among hominoid primates

4.1 Introduction

4.1.1 miRNA biogenesis and their role in eukaryotic gene regulation

MicroRNAs (miRNAs) are one of the largest gene regulators found in the plant and animals branches of eukaryote phylogeny (Carthew and Sontheimer 2009). These are small RNAs (~21 nucleotides) that play a major role in many eukayotic genes (Carthew and Sontheimer 2009, Hausser et al. 2013). miRNAs are associated with the regulation of a broad range of biological functions including chromosome segregation, cell cycle regulation, RNA processing, growth and development, various diseases and aging (Carthew and Sontheimer 2009, Somel et al. 2010, 2011, Kloosterman and Plasterk 2006).

Like protein coding genes, miRNA genes can also originate evolutionaily in various ways (Fig. 4.1). miRNA genes can originate from intergenic or intronic sequences (Campo-Paysaa et al. 2011, Chen and Rajewsky 2007, Berezikov et al. 2011) or from the insertion of transposable elements (TE) and repeats (Shabalina and Koonin 2008). A classic example of the origin of miRNA gene from TE insertion is Mir-548 family. This family of miRNAs has originated due to the insertion of MADE1, a TcMar-Mariner family DNA transposon (Piriyapongsa et al. 2007). One of the major mechanisms of the origin of new miRNA genes is gene duplication, which increases the size of a particular miRNA family. The final means of the generation of miRNA genes is *de novo* generation. Interestingly de novo generation can be equally important

mechanism for the origin of miRNA genes like gene duplication (Campo-Paysaa et al. 2011, Chen and Rajewski 2007, Shabalina and Koonin 2008, Liu et al. 2008, Gu et al. 2009).

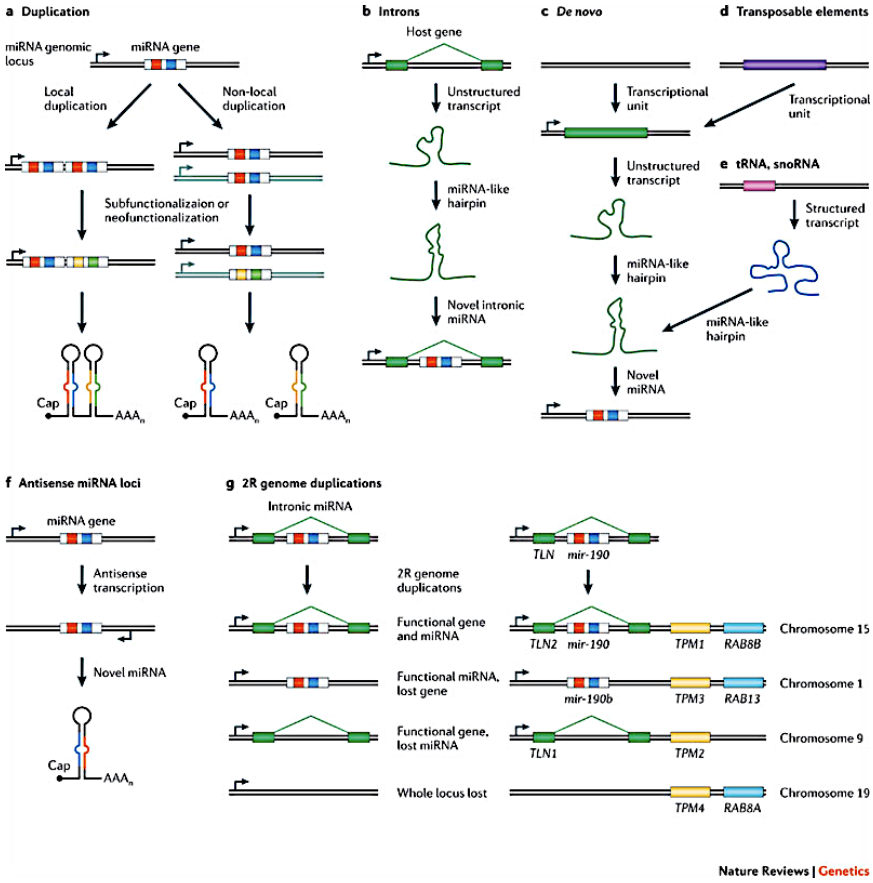


Figure 4.1: Various processes of the origin of miRNA genes. The figure is borrowed from Berezikov (2011) with permission from the journal

miRNA genes are transcribed by RNA polymerase II and like mRNAs, they are also capped and polyadenylated (Kim 2005). A transcript may code anywhere from a single miRNA to a cluster of many miRNAs (Carthew and Sontheimer 2009). This pri-miRNA transcript extends both 5' and 3' direction from the miRNA and two subsequent trimming of the transcript convert it into a mature miRNA (Carthew and Sontheimer 2009) (Fig. 4.2). A typical pri-miRNA in animals has a stem-loop structure with

imperfectly paired stem of ~33bp (Bartel 2004, Carthew and Sontheimer 2009). Pri-miRNA processing is catalyzed by Drosha, a nuclear member of the RNase III family (Lee et al. 2003, Kim 2005). This catalysis is aided by DGCR8, a cofactor containing two double-stranded RNA binding domains (dsRDB) (Denli et al. 2004). The resulting pre-miRNA is exported from the nucleus to cytosol and processed by Dicer to form the mature miRNA/miRNA* duplex of ~22 nucleotides (Bartel 2004) (Fig. 4.2). This second processing step excises the terminal loop of the pre-miRNA (Carthew and Sontheimer 2009). After the second processing step, the resulting miRNAs are assembled into miRISC (Fig. 4.2)

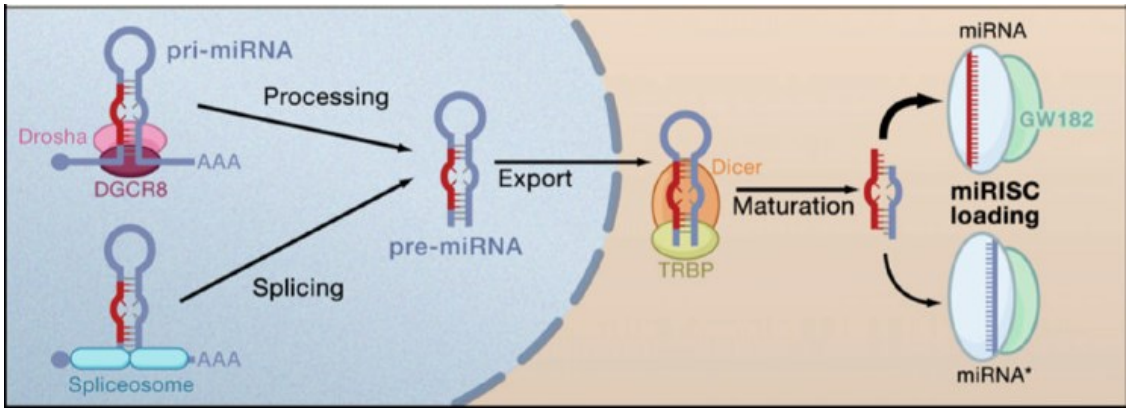


Figure 4.2: miRNA biogenesis. The figure is borrowed and modified from Carthew and Sontheimer (2009) with permission from the journal

The mature miRNA duplex (miRNA/miRNA*) has a very short life span and dissociates and unwinds as soon as they are associated with Argonaute proteins (Ago) (Carthew and Sontheimer 2009). Only one of the two strands is retained and gets associated with Ago (Okamura et al. 2008). The strand that is more commonly associated with Ago is called the miRNA strand and the other, less commonly associated strand is called the miRNA* strand (Carthew and Sontheimer 2009).

In animals miRNAs interact predominantly through their ‘seed region’ (nucleotides 2-8 from the 5' end of the miRNA) with 3’ untranslated regions (3’ UTRs) of mRNAs (Lewis et al. 2005). miRNAs in the majority of the cases either destabilizes the mRNAs or inhibit them from undergoing translation (Filipowicz et al., 2008). In the recent past many miRNA target sites have been also found within the coding region (CDS) of mRNA transcripts (Forman et al. 2008, Qin et al., 2010, Ott et al. 2011, Huang et al. 2010, Hausser et al. 2013).

The binding complementarity between the mRNA and the miRNA is essential for the mRNA regulation. If there is a perfect complementarity between the mRNA and the miRNA, Ago will catalyze a cleavage reaction that will destroy the mRNA strand. On the other hand, if there are mismatches between the mRNA and miRNA sequences, no cleavage will occur but the mRNA will be translationally silenced (Carthew and Sontheimer 2009).

How miRNA-Ago associated miRISC complex repress mRNA from translation is under debate. There are five potential models for mRNA silencing (Carthew and Sontheimer 2009). The binding of miRISC to the target mRNA can prevent the mRNA cap to bind with the initiation factor eIF4F and thus repress translation at the cap-recognition stage (Fig. 4.3). Alternatively, miRISC can interact with the eIF6-60S complex and prevent the 60S ribosomal subunit to assemble in the mRNA (Fig. 4.3). Another alternative is miRISC can induce deadenylation of the mRNA and thereby inhibit circularization of the mRNA (Fig. 4.3). According to the third model, miRISC can repress translation by inducing permanent removal of ribosomal subunits from the target

mRNA (Fig. 4.3). Finally, miRISC can cause complete degradation of the target mRNA by inducing deadenylation followed by decapping (Fig. 4.3).

mRNA degradation by the miRNAs can occur with or without translational repression. As mentioned before, miRNA-mediated mRNA degradation can occur by Ago-catalyzed mRNA cleavage. It has been shown *in vitro* that this process can take place without active translation (Wakiyama et al. 2007). On the other hand, the mRNA degradation during translation is an independent process that occurs because of deadenylation, decapping, and exonucleolytic digestion of the mRNA (Behm-Ansmant et al. 2006, Giraldez et al. 2006, Wu et al. 2006).

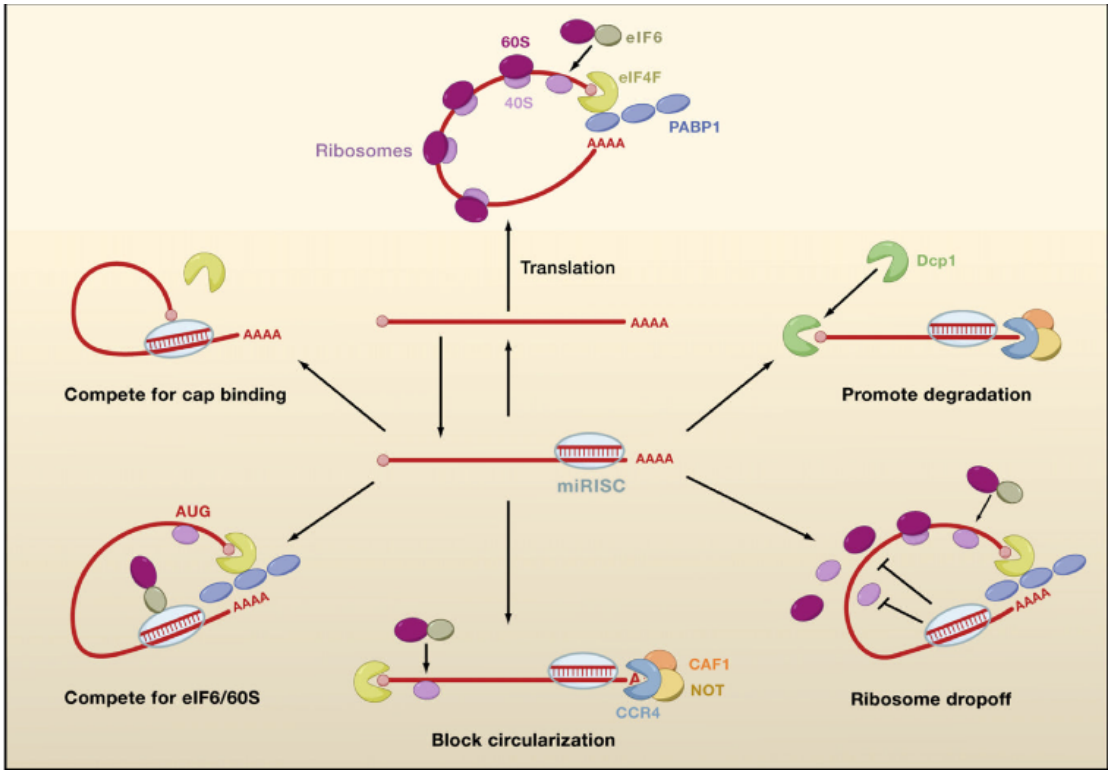


Figure 4.3: Potential models for translational repression by miRNAs. The figure is borrowed from Carthew and Sontheimer (2009) with permission from the journal

4.1.2 Evolution of miRNAs

Due to their essentiality in eukaryotic gene regulation, miRNA regulation may underlie many species specific or lineage specific adaptations (Lu et al. 2008, Zhang et al. 2007, Niwa and Slack 2007). miRNA provides combinatorial control power, flexibility, robustness, and buffering during eukaryotic gene regulation (Stark et al. 2005, 2007, Wu et al. 2009, Flynt and Lai 2008, Bartel and Chen 2004). This functional versatility makes miRNAs excellent target for adaptive evolution (Niwa and Slack 2007, Levine and Tjian, 2003, Hertel et al. 2006). It has been shown that the morphological complexity has a direct association with the number of miRNAs i.e. more complex animals have an expanded repertoire of miRNAs (Prochnik et al. 2007, Grimson et al. 2008, Heimberg et al. 2008, Peterson et al. 2009, Wheeler et al. 2009).

Strong selection pressure operates on miRNAs with essential functions and broader expression patterns. This model supports the fact that ~30% of human miRNAs with very little expression are under weak selection pressure and are less evolutionary constraint (Liang and Li 2009, Berezikov et al. 2006). The novel miRNAs are under weaker evolutionary constraint with very few conserved targets and low expression pattern, while the older miRNAs are under strong purifying selection and expressed broadly and robustly (Nozawa et al. 2010, Ruby et al. 2007, Lu et al. 2008, Stark et al. 2007).

Recent comparative genomic approaches have identified two primate specific miRNAs Zhang et al. (2007, 2008). More recently, a miRNA cluster was discovered on primate X-chromosome spanning ~33-kb region in human Xq27.3. The cluster consists of six distinct miRNAs and thought to have originated after the primate–rodent split but

before the divergence of New World and Old World monkeys (Li et al. 2010). Hu et al. (2011) with a combination of high-throughput sequencing, miRNA microarrays, and Q-PCR have shown that ~11% of miRNA expression differed significantly between human and chimpanzee brains and ~ 31% between human and macaque brains. They also identified a signature of positive selection in the upstream region of miR-34c-5p, a miRNA with human-specific expression.

Although a novel human specific miRNA (has-mir-941) has recently been identified (Hu et al. 2012), no one studied the uniquely gained and lost miRNAs within great apes. Since, these species share ~98% of their genome with each other, very few novel miRNA genes can be found in these species. The study of these novel miRNA genes within hominoids is very important since they can account for the unique species specific or lineage specific gene regulation, which aids in the evolution of the species. In this chapter I computationally predicted uniquely gained and lost miRNAs within hominoid primates and addressed the potential role of these miRNAs in unique species-specific gene regulation.

4.2 Methods

4.2.1 Investigation of the uniquely gained and lost miRNAs in hominoids

Currently identified miRNAs from human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*) and rhesus macaque (*Macaca mulatta*) were collected from miRBase (http: //www.mirbase.org/) along with their sequences. miRBase contains published miRNA sequences and their annotation and is updated on a regular basis whenever a novel miRNA is identified. Additional miRNA information was gathered from microRNAviewer (http:

//people.csail.mit.edu/akiezun/microRNAviewer/index.html). It contains a large set of miRNA genes from miRbase (Griffith-Jones et al. 2008) and their putative homologs using miRNAminer (Artzi 2008). So, this database possesses both experimentally identified and computationally predicted miRNAs. The sequences of the unique species-specific and lineage specific miRNAs were then mapped with hg19, panTro4, gorGor3, ponAbe2 and/or rheMac3 genomes using BLAT, implemented in UCSC Genome Browser (Kent 2002, Kent et al. 2002) to confirm the uniqueness of the sequence. Reciprocal best-hit BLAST technique was used in all cases. In Reciprocal BLAST the sequence from one species is BLASTed to another species. Then the highest-scoring sequence is taken and BLASTed back to the database of the first species. If the returned best hit is the same sequence originally used as the highest scorer, then the two genes are considered putative orthologs. The whole process of identification of novel miRNAs is summarized in Fig. 4.4.

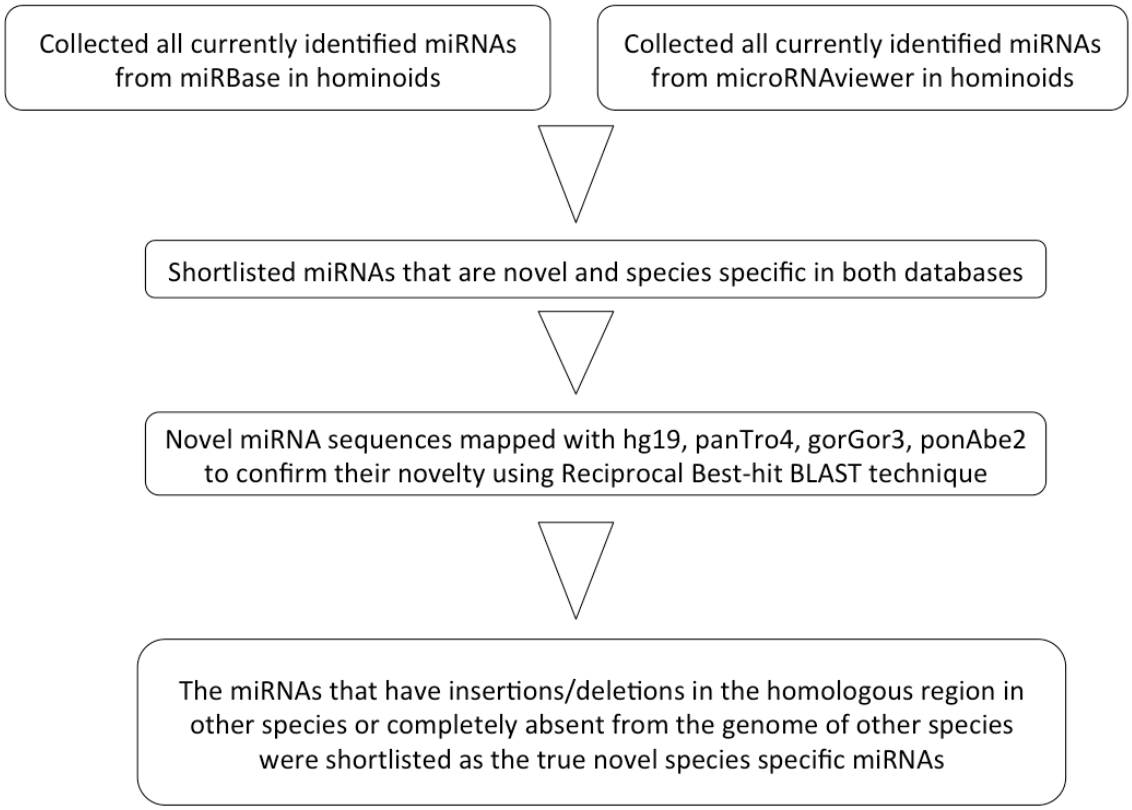


Figure 4.4: Flowchart outlining the process of identification of novel miRNAs

The alignment of the miRNA genes and their homologs in other species were generated using Clustal Omega web-based server (http: //www.ebi.ac.uk/Tools/msa/clustalo/). Phylogenetic analysis of orthologous and paralogous miRNA genes was carried out using MEGA 5.0 (Tamura et al. 2011). The structures of the uniquely gained human miRNAs were compared to the homologous region in chimpanzee using RNAstructure (http: //rna.urmc.rochester.edu/RNAstructureWeb/index.html) web-based server.

4.2.2 Investigation of tissue-specificity of the uniquely gained and lost miRNAs in hominoids

For annotation of the tissue expression of uniquely gained and lost miRNAs TAM (Tool for Annotations of miRNAs; Lu et al. 2010 http: //202.38.126.151/hmdd/tools/tam.html/) and mESAdb (miRNA sequence and expression database; Kaya et al. 2011 http: //konulab.fen.bilkent.edu.tr/mirna/mirna.php) web-based servers were used. Both servers are old and so do not have information of newly discovered miRNAs. mESAdb contains expression data from real experiments for some human miRNAs. TAM, on the other hand predicts the function of miRNAs based on available literature and its own algorithms. Also they only have human miRNA annotations. In mESAdb, I used Meiri et al. (2010) and Navon et al. (2009) miRNA expression datasets for annotating the tissue specificity of the uniquely gained and lost miRNAs. These are the only two most recent experimentally validated miRNA expression datasets for human available in the mESAdb server. Tissue specificity could not be determined for miRNAs absent in humans.

4.2.3 Investigation of potential disease association of the uniquely gained and lost miRNAs

The disease association of miRNAs was obtained from Human miRNA & Disease Database (HMDD) (Lu et al. 2008 http: //202.38.126.151/hmdd/mirna/md/). The creators of the HMDD database have manually retrieved the associations of miRNA and disease from literatures and built the database. Since the database is created mostly manually, I encountered some manual errors in the raw association file. I rectified those mistakes before further analysis. The whole miRNA-disease association file (`hmdd2012-09-`

09.txt) was downloaded from the server. The file was manually crosschecked before analysis. The file was renamed as humanDisease_miRNA.txt. The uniquely gained and lost miRNAs were compared to this file using 'grep -w' command in UNIX. The use of the whole word grep command confirmed that only the desired miRNAs would be extracted from the file. Multiple miRNAs were separated by '\|', while grepping, using the command line:

```
cat humanDisease_miRNA.txt|grep -w 'hsa-mir-*\|hsa-mir-
*\|....'
```

Like in the case of tissue-specificity, disease association could not be determined for miRNAs absent in humans.

4.2.4 Conservation of miRNA genes among hominoids

To determine conservation of miRNAs I used Genomic Evolutionary Rate Profiling (GERP) Scores (Sidow lab http://mendel.stanford.edu/SidowLab/downloads/gerp/). During calculation of GERP scores, each base position in the genome is scored independently in a maximum likelihood framework. Positive GERP score is proportional to evolutionary constraint [higher the score more the constraint] on that base position. GERP++ is implemented in the UCSC Genome Browser (Davydov et al. 2010). Maximum GERP score in UCSC browser is capped at 6.18 and minimum capped at -12.36. A 0 GERP score suggests a shallow sequence alignment at that base position and the algorithm could not estimate a constraint score. The GERP scores were determined from an alignment of 19-35 mammalian species. Scores were determined only for those positions that have at least three ungapped

species present. GERP++ uses the HKY85 model of evolution with the transition/transversion ratio set to 2.0.

In order to get the GERP scores, first, the miRNA gene coordinates were downloaded from UCSC browser for Chr 1, 2, 3, 4, 5, 10, X and Y. The downloaded files were then modified in UNIX using the command line:

```
cat chr*miRNA.txt| sed '/^#/d'|awk '{print
$2"\t"$3"\t"$4"\t"$5}'  > *miRNA.txt
```
[* = Chromosome number]

The GERP scores were obtained for all base positions within the miRNA gene coordinates through the UCSC browser. The obtained GERP score files were modified to delete the index lines using the following command line:

```
cat *miRNAgerp.txt|grep -v 'track'| grep -v '#'>
*miRNAgerpmod.txt
```

Average GERP scores for each miRNA gene coordinate was calculated using the following command line:

```
 awk '/^variableStep/ {if(NR!=1){print " " sum/cnt}; f=0;
sum=0; cnt=0; next} {sum+=$2; cnt++} END {print " "
sum/cnt}'  *miRNAgerpmod.txt > *miRNAavggerp.txt
```

Finally, the `*miRNA.txt` and `*miRNAavggerp.txt` files are combined using the following command line:

```
paste *miRNA.txt *miRNAavggerp.txt > *miRNA_gerp.txt
```
[* = Chromosome number]

To find out lineage specific uniquely gained miRNA gene GERP scores, the human (hg19) genome coordinates for the miRNA genes from Catarrhini, Great Apes,

African Apes, Homo-Pan clade and Uniquely Human were taken. Then a similar approach (as mentioned above) was taken to generate each lineage specific GERP score files. Graphs were generated in GraphPad Prism v6. Kruskal-Wallis Test, followed by Dunn's multiple comparison was carried out to investigate significant difference in conservation of miRNA genes among different Primate lineages.

4.2.5 Investigation of targets of uniquely gained miRNAs

The putative targets of the miRNAs was determined from DIANA microT-CDS v5.0 (Maragkakis et al. 2013 http: //diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index), TargetScanHuman v6.2 (Whitehead Institute for Biomedical Research, 2012 http: //www.targetscan.org/), and miRanda (Jhon et al. 2005 http: //www.microrna.org/microrna/home.do). All these webservers use a different statistic to predict miRNA targets. DIANA uses the 5th version of the microT algorithm created by Maragkakis et al (2012). The putative targets are ranked according to the miTG score generated by microT algorithm. The higher the miTG score the higher the probability of targeting. For the current study the cut off of miTG score was set at 0.7 (ranges 0-1). TargetScanHuman predicts miRNA targets by looking for the presence of conserved 8mer and 7mer sites in the targets that match the seed region of miRNAs. The putative miRNA targets are arranged by context-scores, that the sum of the contribution of these four feature: site-type contribution, 3' pairing contribution, local AU contribution, and position contribution. miRanda (microrna.org) uses mirSVR algorithm-generated scores (Betel et al. 2010). The putative targets are arranged on a neagive scale mirSVR scores (the lower the better). After preliminary analysis we decided to use the targets predicted by DIANA microT-CDS for further

analysis. DIANA microT-CDS has been recently used to predict human miRNA targets for comprehensive target network analysis (Satoh 2011). The raw file downloaded from the DIANA server was named as `mir-*_targets.txt`. The target gene lists were generated from this file using the following UNIX command line:

```
cat mir-*_targets.txt|tr ',' '\t'|tr ' ' '\t'| grep -v
'UTR3'| grep -v 'CDS'|awk '{print $3}'|tr '(' ' '|tr ')' '
' > mir-*_tragetGenes.txt
```

4.2.6 Investigation of potential biological function of the target genes of uniquely gained miRNAs

The target gene list was exported to the gene ontology PANTHER classification system server (http: //www.pantherdb.org/). A statistical overrepresentation test, implemented in the PANTHER server was performed with all gene lists to determine statistically over- and under-represented biological processes among the genes. The enrichment analysis file downloaded from the PANTHER server was divided into separate over and under represented biological function lists using the commands:

```
cat mir-*_analysis.txt| grep -v 'Biological Process'| grep
'+' > mir-*_positive-enrichment.txt

cat mir-*_analysis.txt| grep -v 'Biological Process'| grep
'-' > mir-*_negative-enrichment.txt
```

4.2.7 Investigation of the predominant site of the uniquely gained miRNA regulation – 3' UTR vs. CDS

As mentioned in 4.2.5, the raw file downloaded from the DIANA server was named as `mir-*_targets.txt`. The target genes were then grouped into two groups: genes targeted at the 3'UTR and the genes regulated at the coding region (CDS). No

target gene was regulated at the 5'UTR region. The files were generated from `mir-*_targets.txt` using the following UNIX commands:

```
 cat mir-*_targets.txt|tr ',' '\t'|tr ' ' '\t'| grep -v
'UTR3'|awk '{print $3}'|tr '(' ' '|tr ')' ' '|grep -B1
'0.0'|grep -v '0.0' > mir-*_CDStarget.txt

 cat mir-*_targets.txt|tr ',' '\t'|tr ' ' '\t'| grep -v
'CDS'|awk '{print $3}'|tr '(' ' '|tr ')' ' '|grep -B1
'0.0'|grep -v '0.0' > mir-*_UTR3target.txt
```

Venn diagrams were plotted using Draw Venn Diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/) web-based server to determine the number of genes that are targeted in either 3'UTR or CDS vs. the genes that are regulated at both 3'UTR and CDS.

4.2.8 Conservation of miRNA target sites

The coordinates of uniquely gained miRNA targets were obtained from the raw file (`mir-*_targets.txt`) downloaded from the DIANA server using the following UNIX command line:

```
cat mir-*_targets.txt|tr ',' '\t'|grep 'UTR3\|CDS'|awk
'{print "chr"$2}'| grep -v ';' > mir-
*_target_coordinates.txt
```

Once the coordinates of the target sites are obtained from the DIANA server, the GERP scores for each target site were generated from the UCSC genome browser. The target sites were grouped into four groups: *Homo* only target sites, *Homo-Pan* target sites, African Ape target sites, and Ape target sites. The conservation of these target sites were compared to the conservation of the targets of 10 randomly selected miRNAs, found among all Catarrhini Primates. Graphs were generated in GraphPad Prism v6. Kruskal-

Wallis Test, followed by Dunn's multiple comparison was carried out to investigate significant difference in conservation of miRNA targets among different Primate lineages.

A correlation test was performed between the GERP scores of the target sites of one *Homo-Pan* gained (has-mir-548) and one African Ape (has-mir-320A) gained miRNAs and their associated miRNA binding score (miTG score) to investigate whether the stronger miRNA target sites are more conserved. The correlation test was performed using 'Hmisc' package in R v3.0.2. Subsequently, a 'corrgram' was plotted using the 'corrgram' package in R v3.0.2.

4.3 Results

4.3.1 Uniquely gained and lost miRNAs in hominoids

The miRNA genes that have unique insertion or deletion in one species, but not found in the homologous region of other hominoids are considered species specific uniquely gained or lost miRNA. Similarly, the miRNA genes that have unique insertion or deletion in a specific hominoid lineage but not found in other hominoids are considered group-specific uniquely gained or lost miRNA. The uniquely gained and lost species specific as well as group specific miRNAs are summarized in Table 4.1. The alignments of uniquely gained and lost miRNA genes and their homologous regions in other hominoids are shown in Appendix 3.1. Two uniquely gained human miRNA and their chimpanzee counter part are shown in Figure 4.5. The rest are shown in Appendix 3.2.

Table 4.1: **Species specific and group specific uniquely gained and lost miRNAs**

	Human	Chimp	Gorilla	Orang	Human-Chimp-Gorilla	Great Apes	Human -Chimp	Human-Gorilla	Chimp - Gorilla
Gain	Mir-585 Mir-941 Mir-1289 Mir-1303 Mir-3118-5 Mir-3125 Mir-3679 Mir-3913 Mir-3916 Mir-3919 Mir-3938 Mir-3941 Mir-4327 Mir-4329	Mir-1935 Mir-3470			Mir-320A	Mir-588 Mir-617 Mir-634 Mir-645 Mir-1295 Mir-3156 Mir-3202	Mir-548-a,d,f,I,j,k,l,n,p Mir-635 Mir-935	Mir-1270 Mir-1470 Mir-3124 Mir-3130 Mir-3142	
Loss	Mir-1283	Mir-611 Mir-620 Mir-1287 Mir-2116 Mir-3153 Mir-3179 Mir-3925	Mir-1-1 Mir-184 Mir-320C1 Mir-515 family Mir-708 Mir-1237 Mir-1263 Mir-1299	Mir-718 Mir-1272 Mir-1278 Mir-1825 Mir-2117 Mir-3653			Mir-578		Mir-132 Mir-466 Mir-567

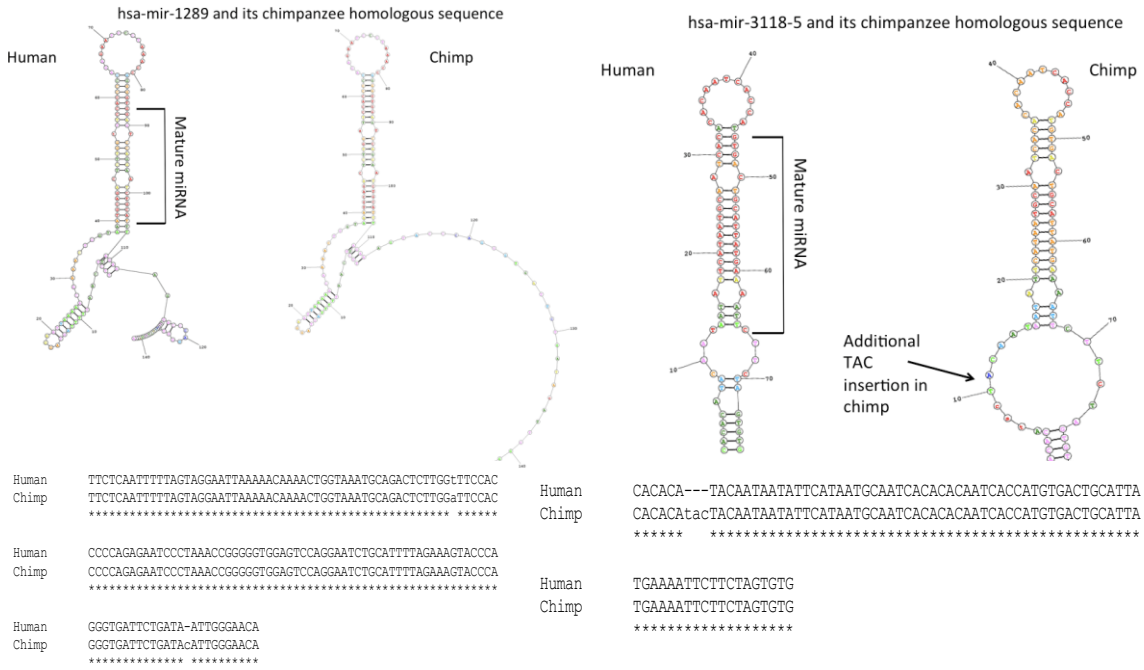


Figure 4.5: Comparison of the structure of human uniquely gained miRNA and its chimpanzee homolog

4.3.1.1 Insertion of MADE1 DNA transposon in mir-548 family

MADE1 is a TcMar-Mariner family DNA transposon, containing the consensus sequence GTTGGTGCAAAAGTAATTG (Fig. 4.5). In primates this transposon is found inserted in mir-548 family. In humans, paralogs of mir-548 are found in Chr 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, X. Humans have the highest number of paralogs of mir-548 family followed by chimpanzee. Gorilla has the lowest number of paralogs (Piriyapongsa and Jordan 2007).

```
hsa-mir-548a-1        --------------UGCAGGGAGGUAUU-AA---GUUGGUGCAAAAGUAA---UUGUG--
37

ptr-mir-548a-1        ---------------GCAGGGAGGUAUU-AA---GUUGGUGCAAAAGUAA---UUGUG--
36
ggo-mir-548a          ----AUUUAUGCACUGCAGGGAGGUAUU-AA---GUUGGUUCAAAAGUAA---UUGUG--
47
ppy-mir-548a          ----AUUUAUGCACUGCAGGGAGGUAUU-AA---GUUGGUGCAAAAGUAA---UUGUG--
47

mml-mir-548a          --------------UCCAGGGAGGUAUU-AA---GUUGGUGCAAAAGUAA---UUGUG--
37

ptr-mir-548a-2        ---------------GUGAUGUG-UAUU-AG---GUUUGUGCAAAAGUAA---CUGGG--
35
```

Figure 4.6: ClustalW alignment of mir-548a among Catarrhini Primates

4.3.1.2. Tandem duplication of mir-515 family in Chr19 of Catarrhini Primates

There are 44 tandem duplications in Chr19: 54,182,257-54,264,476 in humans (Fig. 4.6). Other hominoids have fewer duplications. All mir-515 family members uniquely share the consensus sequence: TGACTCTACAAAGG (Fig. 4.7). Gorilla has only two paralogs of mir-515 family: 516b and 517c. The phylogenetic relations of various human paralogs of mir-515 family are shown in Fig. 4.8.

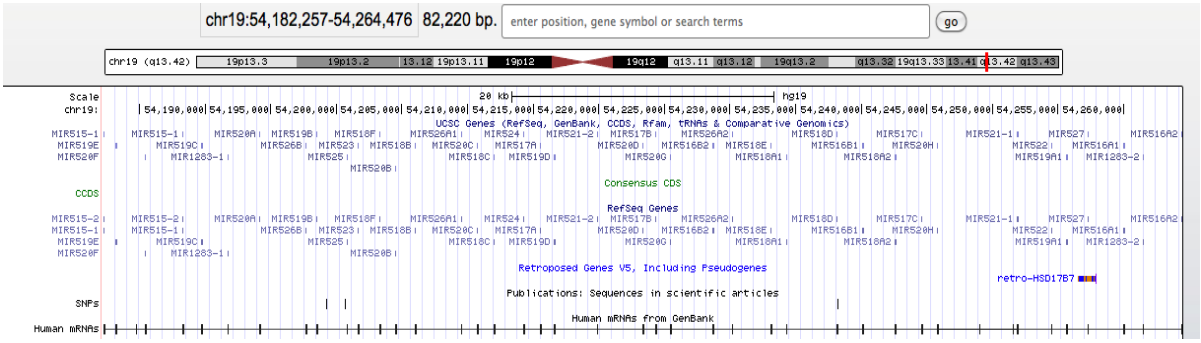


Figure 4.7: Screen-shot of UCSC Genome Browser, showing tandem duplication of mir-515 family in human

```
hsa-mir-519d        -----------------UCCCA---UGCUG------UGACCCUCCAAAGG---GA----- 26
hsa-mir-527         -----------------UCUCA---AGCUG------UGACU--GCAAAGG---GA----- 24
hsa-mir-520d        -----------------UCUCA---AGCUG------UGAGUCUACAAAGG---GA----- 26
hsa-mir-1283-1      -----------------CUCA---AGCUA------UGAGUCUACAAAGG---AA----- 25
hsa-mir-521-1       -----------------UCUCA---GGCUG------UGACCCUCCAAAGG---GA----- 26
hsa-mir-524         -----------------UCUCA---UGCUG------UGACCCUACAAAGG---GA----- 26
hsa-mir-525         -----------------CUCA---AGCUG------UGACUCUCCAGAGG---GA----- 25
hsa-mir-520a        -----------------CUCA---GGCUG------UGACCCUCCAGAGG---GA----- 25
hsa-mir-518b        -----------------UCA---UGCUG------UGGCCCUCCAGAGG---GA----- 24
```

Figure 4.8: ClustalW alignment of human 515-family containing ‘TGACTCTACAAAGG’ sequence

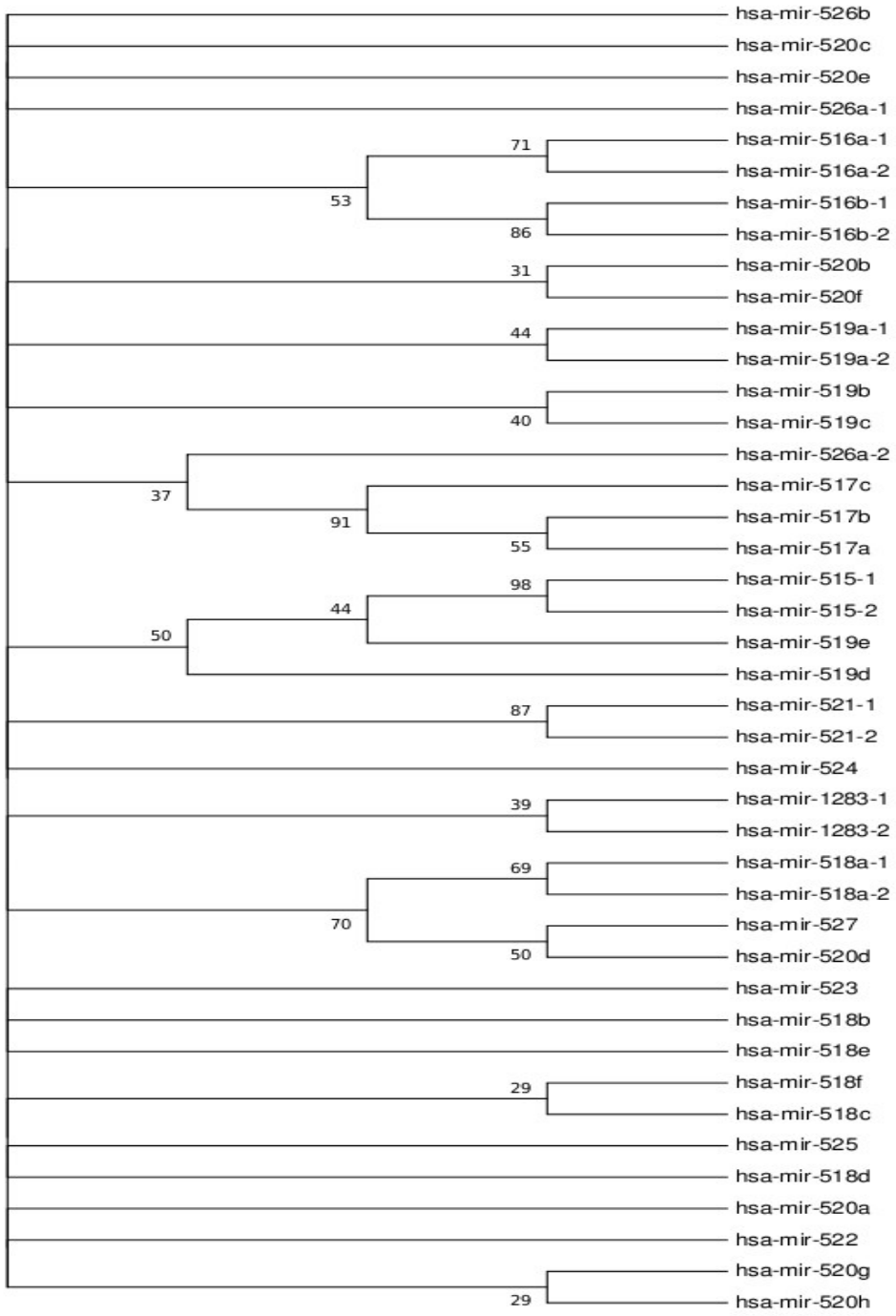


Figure 4.9: Maximum Parsimony Tree showing the interrelation between mir-515 family members

4.3.2 Most uniquely gained and lost miRNAs are brain specific

Three out of four miRNAs gained uniquely by human, two out of three miRNAs gained by human and chimpanzee, and both miRNAs gained by human and gorilla are brain specific (Table 4.2). Seven out of eight miRNAs lost in gorilla, two out of three miRNAs lost in chimpanzee, one of the two miRNAs lost in orangutan are brain specific (Table 4.2). Also, two out of three miRNAs lost uniquely in chimpanzee and gorilla are also brain specific. Gorilla also lost two prostate-specific miRNAs (Table 4.2).

Table 4.2: Tissue specificity of uniquely gained and lost miRNAs

	Brain	Heart	Liver	Kidney	Lung	Ovary	Testes	Endometrium	Other
Human	Mir-585 Mir-941 Mir-1289 Mir-1303	Mir-585 Mir-1289 Mir-1303	Mir-585 Mir-941 Mir-1289 Mir-1303	Mir-585 Mir-1289	Mir-941 Mir-1289 Mir-1303			Mir-585 Mir-1289 Mir-1303	Mir-1283 (Placenta)
Chimpanzee	Mir-611 Mir-620	Mir-611 Mir-620 Mir-1287	Mir-611 Mir-620	Mir-611 Mir-620	Mir-611 Mir-620	Mir-611 Mir-620		Mir-611 Mir-620	
Gorilla	Mir-184 Mir-320C Mir-519e Mir-708 Mir-1237 Mir-1263 Mir-1299	Mir-1-1 Mir-518e Mir-519e	Mir-184 Mir-518e Mir-519e Mir-1237	Mir-184 Mir-518e Mir-519e	Mir-320C Mir-519e Mir-1237		Mir-1-1 Mir-184 Mir-518e Mir-520h	Mir-708 Mir-1237 Mir-1299	Mir-1-1 (Prostate) Mir-515 (Placenta) Mir-1237 (Prostate)
Orangutan	Mir-1278								Mir-718 (Breast)
Human-Chimpanzee	Mir-578 Mir-635 Mir-935	Mir-578 Mir-635	Mir-578	Mir-578 Mir-935	Mir-578			Mir-578 Mir-635	Mir-578 (Prostate) Mir-548 (Breast)
Human-Gorilla	Mir-1270 Mir-1470	Mir-1270 Mir-1470	Mir-1270 Mir-1470	Mir-1270 Mir-1470	Mir-1270 Mir-1470			Mir-1270 Mir-1470	
Chimpanzee-Gorilla	Mir-132 Mir-567	Mir-567	Mir-466 Mir-567	Mir-567	Mir-567				
African Apes	Mir-320A	Mir-320A	Mir-320A				Mir-320A		

*Red miRNAs suggest uniquely lost miRNAs while black miRNAs suggest uniquely gained miRNAs

4.3.3 Disease association of uniquely gained and lost miRNAs

The miRNAs uniquely gained by humans either in a species-specific manner or in a group specific manner are associated with tumorogenesis and neoplasms (Table 4.3a). Mir-320A, uniquely gained in the African Ape lineage after its split from *Pongo*, is associated with diabetes (Table 4.3a). All miRNAs uniquely lost in gorilla are associated with cancer (Table 4.3b). The only uniquely lost miRNA in orangutan that shows disease association is mir-718. It too is associated with cancer (Table 4.3b).

Table 4.3a: **Disease association of uniquely gained miRNAs**

Uniquely gained miRNAs	**Disease Association**
hsa-mir-941 (Human Only)	Ulcerative Colitis
hsa-mir-1303 (Human Only)	Tumorogenesis
hsa-mir-548 (Human-Chimp)	Tumorogenesis, Tuberculosis
hsa-mir-635 (Human-Chimp)	Adenoviridae Infections
hsa-mir-935 (Human-Chimp)	Tumorogenesis
hsa-mir-3130 (Human-Gorilla)	Tumorogenesis
hsa-mir-320A (African Apes)	Tumorogenesis, Diabetes, Heart failure

Table 4.3b: Disease association of uniquely lost miRNAs

Species	**Uniquely lost miRNAs**	**Disease Association**
Chimpanzee and Gorilla	hsa-mir-132	Tumorogenesis, Heart failure
Chimpanzee	hsa-mir-611	Lupus Vulgaris
	hsa-mir-3179	Tuberculosis
Gorilla	hsa-mir-1-1	Tumorogenesis, Heart failure
	hsa-mir-184	Tumorogenesis
	hsa-mir-320c	Adenoviridae Infections, Aging, Tumorogenesis
	hsa-mir-515 family	Tumorogenesis, Gout
	hsa-mir-708	Tumorogenesis
	hsa-mir-1299	Tumorogenesis
Orangutan	hsa-mir-718	Tumorogenesis

4.3.4 Conservation of miRNAs

4.3.4.1 miRNAs are the most conserved non-coding region in the genome

The average GERP scores of all miRNA genes of all chromosomes show signature of strong evolutionary constraint on miRNA genes. The average GERP score was ~4 (UCSC bowser highest GERP score is capped at ~6) (Fig. 4.9). When compared to other non-coding regulatory regions like promoters and UTRs of the genome, miRNAs turned out to be the most conserved non-coding region of the genome. ($P < 0.05$).

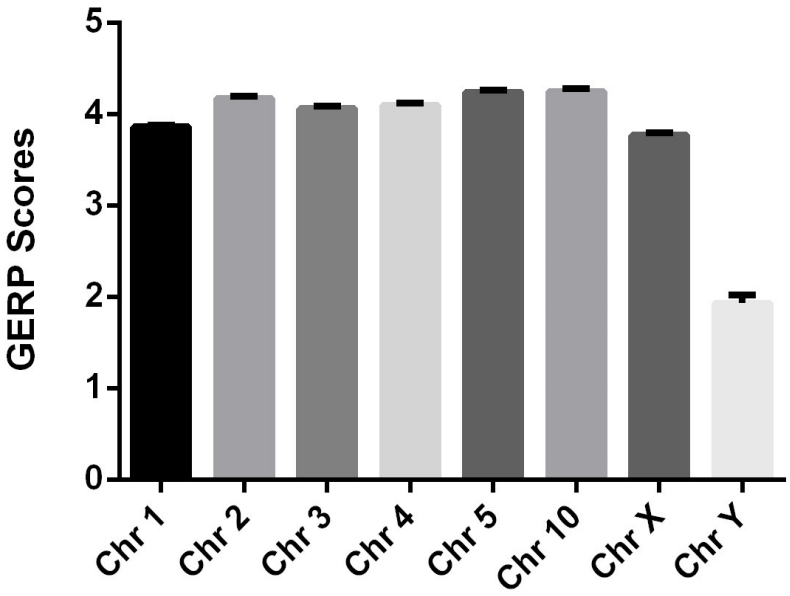


Figure 4.10: Evolutionary constraint on miRNA genes across the genome

4.3.4.2 Older miRNAs are more conserved than younger miRNAs

The miRNA genes that are found in all catarrhin primates (the control miRNA set) are under strongest evolutionary constraint ($P < 0.001$) (Fig. 4.10). The newer miRNA genes (<15 Ma) originated in the Great Ape lineage (*Homo-Pan-Gorilla*), *Homo-Pan* lineage, and uniquely in humans are evolving neutrally.

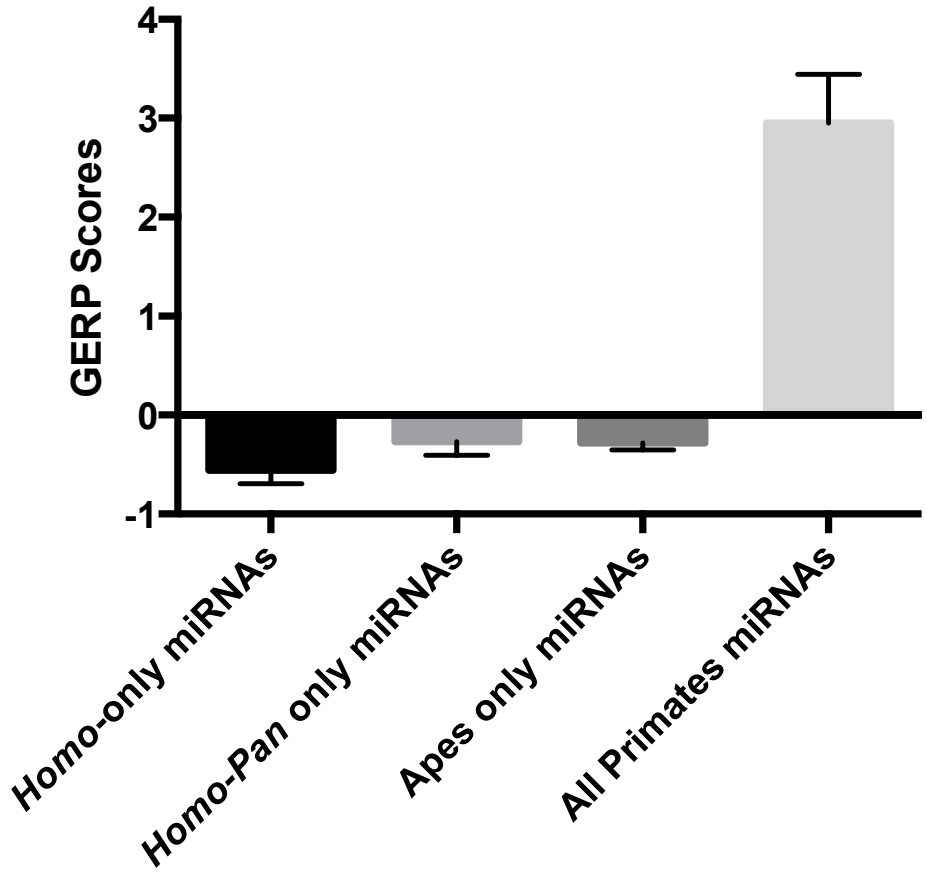


Figure 4.11: Evolutionary constraints on uniquely gained miRNAs

4.3.5 miRNA target prediction websites show disagreement over target site prediction

To compare the targets predicted by different servers, I used two human specific miRNAs (mir-466 and mir-625), one uniquely lost miRNA in gorilla (mir-1) and three universally expressed miRNAs (mir-25, mir-136 and mir-1179). Overall there were high discripencies among servers. DIANA TOOLS, which is by far the most cited target prediction server, worked the best with 63% overlap. microRNA (miRanda) predicted largest number targets, 90-95% of which were unique to that server. psRNATarget

performed the worst. Although it predicted least number of targets, >90% of those targets were unique to psRNATarget server (Fig 4.11).

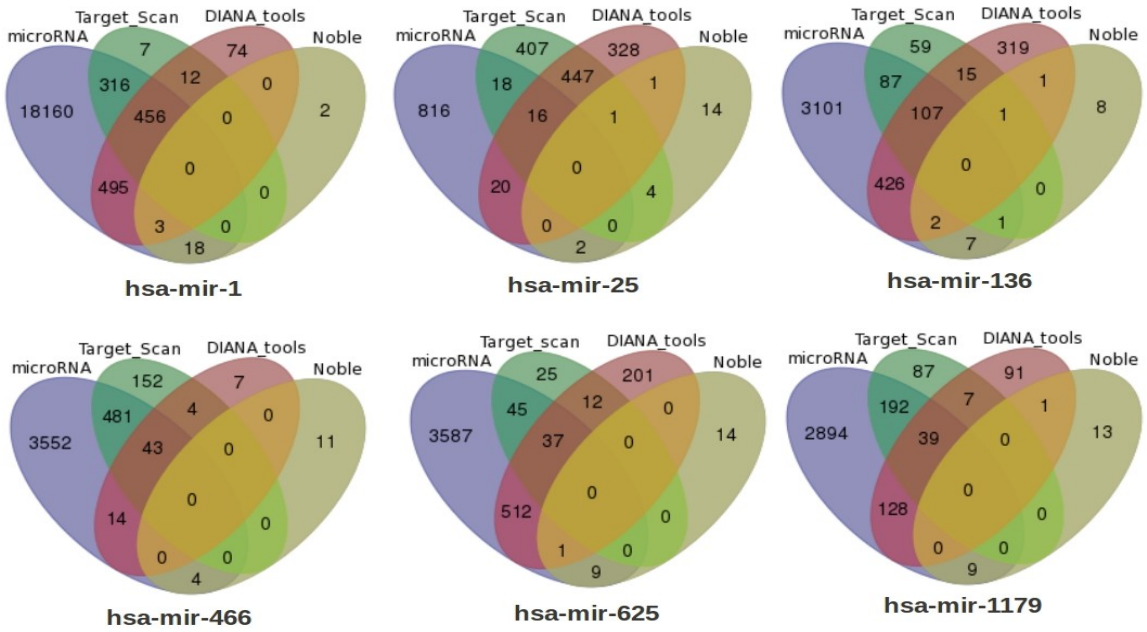


Figure 4.12: Venn diagram showing the overlap of target genes predicted by various target prediction websites

4.3.6 The majority of target genes of the uniquely gained miRNAs are regulated at the 3'UTR region

The target sites are found at both the 3'UTR region as well as the coding region (CDS). However, in all cases (targets of *Homo* only, *Homo-Pan* only, African ape, and Ape specific miRNAs) the majority of target genes are either regulated only at the 3'UTR region or both 3'UTR and CDS regions. Very few target genes are solely regulated at the CDS (Fig. 4.12). No target gene is targeted at the 5'UTR region.

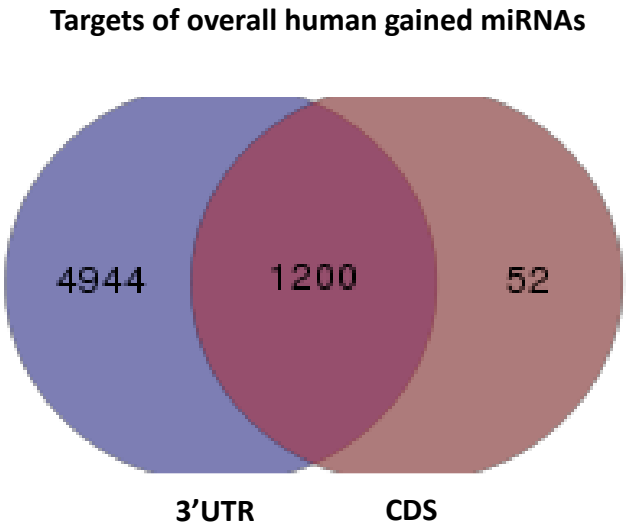


Figure 4.13: miRNA target sites 3'UTR vs. CDS. The orange circle shows the number of target genes regulated at CDS and the blue circle shows the number of target genes regulated at 3'UTR

4.3.7 Annotation of the targets of the uniquely gained miRNAs

The majority of the target genes of uniquely gained miRNAs are significantly (<0.05) enriched in housekeeping functions. The GO Biological functions of the target genes of uniquely gained miRNAs are summarized in Table 4.4.

Table 4.4: **GO Biological function of the target genes of uniquely gained miRNAs**

Target genes	**GO Biological function**
Homo-only	Cell signaling, transcription, synaptic transmission, nervous system development metabolic process, cell adhesion, cell communication, neurological system process, cell cycle, cellular process, ectoderm development, homeostatic process
Homo-Pan only	Metabolic process, biological regulation, transcription, developmental process, cell communication, mRNA processing, cell cycle, nervous system development, cellular component organization, mRNA splicing, cell signaling, neurological system process
African ape specific	Metabolic process, transcription, developmental process, cell communication, nervous system development, mRNA processing and splicing, embryo development
Ape specific	Metabolic process, cell death, biological regulation, development, cellular component organization, transcription, cell communication, neurological system process, neurotransmitter secretion, cellular component organization, cell cycle, synaptic vesicle, immune system process, chromatin organization, homeostasis gamete generation

4.3.8 Target sites of older miRNAs are significantly more conserved compared to Human only miRNA target sites

The miRNA target sites of older miRNAs are significantly more conserved than human-only gained miRNA target sites. Human only gained miRNA target sites are significantly less conserved than the other uniquely gained miRNA target sites (Homo-

Pan gained miRNA target sites, African ape specific miRNA target sites) and all primate specific miRNA target sites (control miRNAs) (Fig. 4.13).

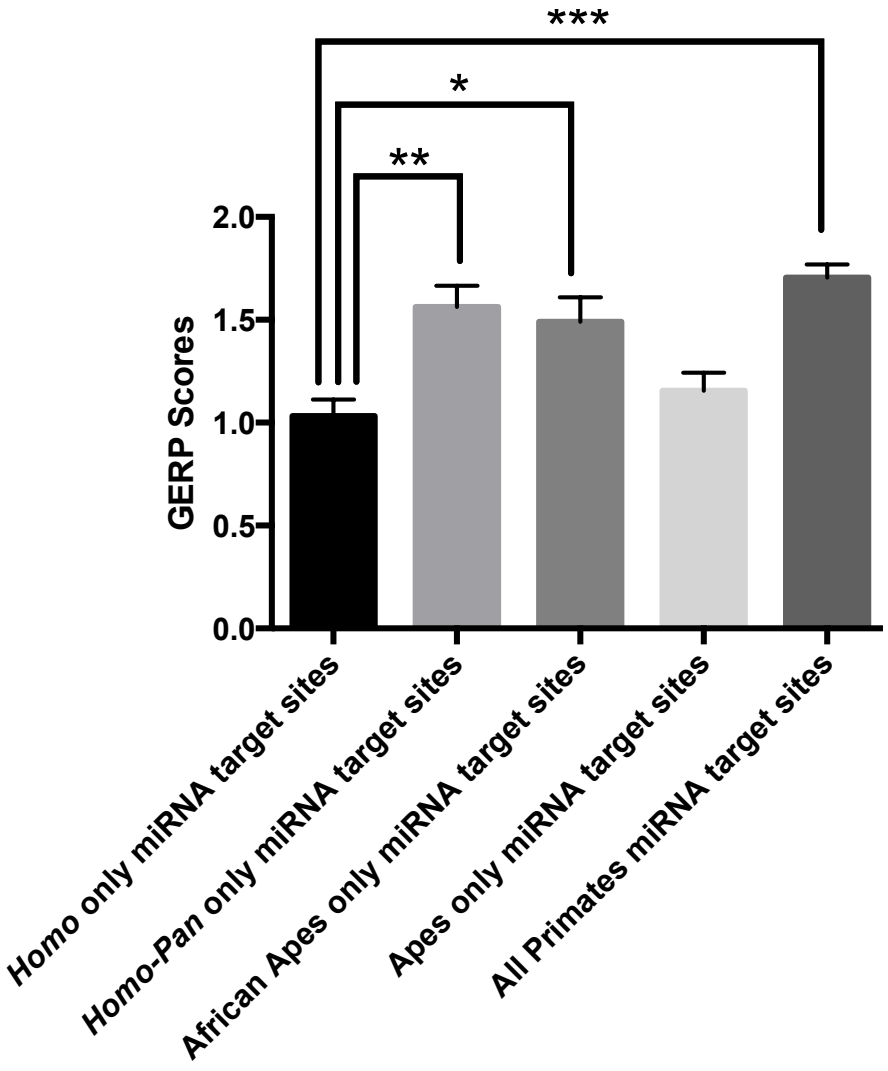


Figure 4.14: miRNA target site conservation among various primate lineages

4.3.9 The conservation of miRNA target sites is correlated with the binding score of the miRNAs

The stronger the binding of the miRNAs to their targets, the more conserved those target sites are ($P < 0.01$). The GERP scores are strongly correlated with the miTG scores in case of both mir-320A (Spearman's rho = 0.98) (Fig. 4.11) and mir-548 (Spearman's rho = 0.91) (not shown).

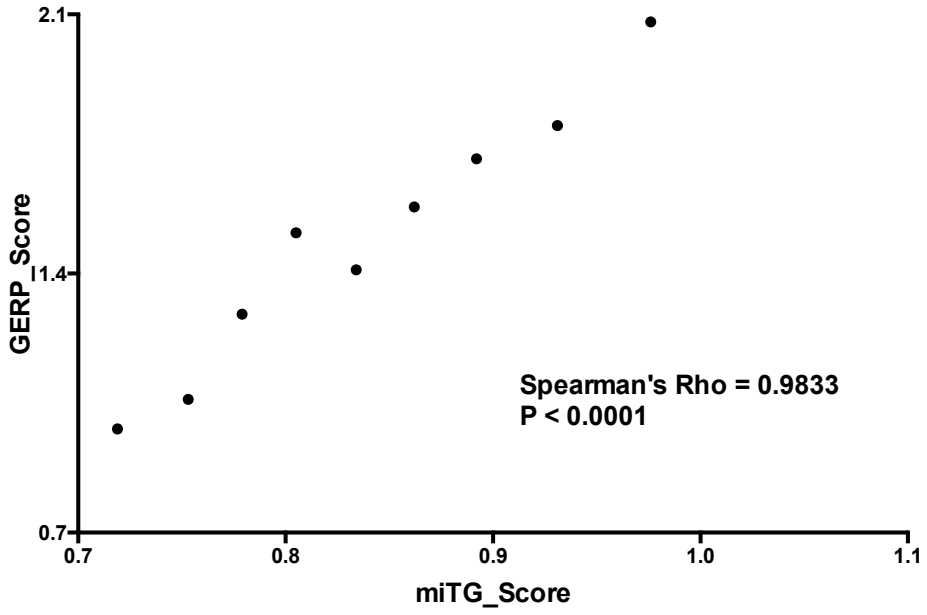


Figure 4.15: Corrgram between the GERP scores and miTG scores. The pie chart shows the degree of correlation between the two variables

4.3.10 The unique insertion/deletions responsible for generating human specific miRNAs are not fixed in the population

I found multiple SNPs at the miRNA genes, uniquely gained by humans. These 'MirSNPs' (Liu et al. 2012) were identified using 1000 Genomes data (McVean et al. 2012) implemented in UCSC genome browser. Only SNPs that are found ≥ 1% of the samples are listed (Table 4.5). In many cases, the unique insertions and deletions that are

responsible for generating human specific miRNAs are not fixed in the population (rs200868230 and rs5829384) (Table 4.5).

Table 4.5 SNPs in the uniquely gained miRNAs in human

miRNA Gene	SNP_ID and nucleotide changes (Reference allele first)	MAF	Comments
Mir-585	rs62376935 (C > T) rs62376934 (A > G)	0.099 0.197	
Mir-941	rs4809383 (C > T) rs2427556 (G > A)	0.140 0.388	
Mir-1289	rs199812733 (- > A)	0.015	
Mir-1303	rs77055126 (T > C) rs199839137 (- > T) rs200868230 (TA > -) rs33982250 (A > -)	0.032 0.097 0.479 0.261	Like 26.1% human population with SNP rs33982250, other non-human hominoids have a single nucleotide deletion instead of 'A', which does not change the stem-loop structure of the miRNA
Mir-3125	rs5829384 (- > A)	0.281	Like 28.1% humans, chimp has an 'A' at this position. Gorilla and orangutan have 'G' at this position. The deletion of 'A' marginally changes the stem-loop structure of the miRNA
Mir-3916	rs113974396 (GA > -)	0.053	Like 5% human population, chimp, gorilla, and orangutan have 2bp deletion instead of 'GA'. 95% human population has uniquely inserted 'GA' at this position. The 'GA' insertion does not change the stem-loop structure of the miRNA

Mir-3938	rs10575780 (AA > -)	0.179	82.1% humans have uniquely inserted 'AA' at this position. Chimp, gorilla, and orangutan have a 2bp deletion. The 2bp insertion significantly changes the miRNA stem-loop structure
Mir-4329	rs146184857 (C > T)	0.017	

* The SNPs that maintain the ancestral insertion/deletion are highlighted in red

4.4 Discussion

4.4.1 Prediction of uniquely gained and lost miRNAs among hominoids

Computational and bioinformatics approaches for the prediction of novel miRNAs in eukaryotes have been an invaluable tool for more than a decade and together with the experimental approaches, they becomes essential for studying miRNA biology (Yoon and Micheli 2006). Using comparative genomic approaches several lineage specific miRNAs has been discovered in *Drosophila* (Li et al. 2006, Lu et al. 2008), *C. elegans* (Lim et al. 2003a, b), rat, and mouse (Sewer et al. 2005, Xie et al. 2005). Moreover, several studies have used computational approaches for the identification of novel species specific and lineage specific miRNAs in primates (Berezikov et al. 2005, Zhang et al. 2007, 2008, Baev et al. 2008, Yuan et al. 2013). Human specific miRNA, hsa-mir-941, was also initially discovered using computational approaches and later validated experimentally (Hu et al. 2012).

Several different bioinformatics approaches are in practice for the identification of novel miRNAs. One of the oldest methods of discovering novel miRNAs is 'MiRscan' (Lim et al. 2003a, b). MiRscan and its later modified version (Ohler et al. 2004) are likelihood-based approaches that assign a log-likelihood score to each base position of the target sequence for its similarities to known miRNAs. 'Phylogenetic shadowing'

approach by Berezikov et al. (2005) is another powerful computational technique that can assess the degree of conservation of each nucleotide in the target sequence. They predicted ~300 novel miRNAs in human genome with this approach, almost 2 fold higher than those predicted by 'MiRscan'. 'miRseeker' approach by Lai et al. (2003) discovered ~150 novel miRNAs in Drosophila. 'ProMiR' algorithm is another strong bioinformatics approach (Nam et al. 2005). It is based on a probabilistic co-learning model that simultaneously compares the structure and sequence of miRNA precursors and can detect less abundantly expressed and less sequence conserved miRNAs, not possible to detect through previous approaches (Yoon et al. 2006).

For initial analysis, I used both MiRscan and miRseeker tools. Both approaches came up with dissatisfactory results. Since the 'ProMir' webpage and the algorithm have not been updated since 2005, I decided not to use this technique for the current study. After initial analysis, I decided to use the most commonly used computation method for detection of novel miRNAs - the comparative method, which provided the most satisfactory results. Comparative methods have been previously employed for the detection of novel miRNAs in human, mouse and other non-human primates (Weber 2005, Yuan et al. 2013, Hu et al. 2012). The reciprocal BLAST approach that I used in this study for the detection of novel miRNAs has been previously employed to detect human-specific and primate specific miRNAs (Yuan et al. 2013, Hu et al. 2012). Like previous studies, the approach was highly successful in detecting species specific and lineage specific miRNAs among hominoid primates.

The major limitation of this study is the lack of experimental evidence. In recent past, once the novel miRNAs were predicted in primates through computational

approaches, they were validated either by deep sequencing or by expression assays (Yuan et al. 2013, Hu et al. 2012). Although computational approaches are great for predicting novel miRNAs, they are limited by genome assembly problems. Since, the current chimpanzee (panTro4), gorilla (gorGor3), and orangutan (ponAbe2) genome assemblies are full of gaps and false deletions, the uniquely gained and lost miRNAs, although extensively filtered, may still not be truly unique. So, the uniquely gained and lost miRNAs, detected in this study, needs to be experimentally validated through deep sequencing to confirm whether they are really unique.

4.4.2 Differential tissue specific expression of uniquely gained and lost miRNAs

Although several previous studies have detected novel primate specific miRNAs (Li et al. 2009, Dannemann et al. 2012, Yuan et al. 2013), they never focused on the uniquely gained and lost miRNAs and their tissue specificity. A previous expression study has showed differential expression of miRNAs among the brains of human, chimpanzee and macaque (Hu et al. 2011) but did not detect the novel miRNAs that are differentially expressed in the primate brains. In this study I detected four uniquely gained human miRNAs expressed in the brain (hsa-mir-585, hsa-mir-941, hsa-mir-1289, and hsa-mir-1303). Recently the expression of hsa-mir-941 in the human brain has been experimentally validated (Hu et al. 2012). These uniquely gained miRNAs may be responsible for differential miRNA expression pattern of human and chimpanzee brains.

The differential miRNA expression between human and chimpanzee brains as mentioned by Hu et al. (2011) may also be because of the brain specific miRNAs that are lost in chimpanzee (hsa-mir-611, hsa-mir-620). Interestingly, most (21) miRNAs that are either uniquely gained or lost in a species specific or lineage specific manner among

hominoids are enriched in brain. Gorilla has lost the maximum number (7) of brain-specific miRNAs among all hominoids and orangutan has lost one brain specific miRNA. The unique gain and loss of multiple brain specific miRNAs is potentially associated with the rapid evolution of brain gene expression in the human lineage, not seen in other hominoids (Somel et al. 2011).

Gorilla has uniquely lost two prostate specific miRNAs, which may be associated with the differential regulation of seminal plasma genes in gorilla, aided by the low sperm competition in this species (See Chapter 3). Interestingly, mir-548d from the miRNA-548 family, uniquely gained in *Homo-Pan* lineage, is enriched in breast and associated with breast cancer (Buffa et al. 2011). Several mir-515 family members (hsa-miR-518b, hsa-miR-516a-5p, hsa-miR-525-5p, hsa-miR-515-5p, hsa-miR-520h, hsa-miR-520a-5p, hsa-miR-519d, and hsa-miR-526b), which are uniquely lost in gorilla, are associated with fetal growth restriction (FGR) placenta, a pregnancy complication commonly seen in humans (Higashijima et al. 2013).

There are two major limitations for detecting differential tissue specificity of miRNAs through computational approaches. Firstly, since the expression arrays used in this study (Meiri et al. 2010 and Navon et al. 2009) are old, they did not include recently detected human miRNAs (miRNA no. 2000 onwards). So, it is possible that the rest 10 uniquely gained human miRNAs, which are recently discovered (miRNA no. >3000), are also highly expressed in the brain and/or shows any other unique tissue specificity. The same thing is applicable to all uniquely lost miRNAs in non-human hominoids. The second major issue is the lack of availability of expression arrays for non-human hominoids. So, the tissue specificity determined for the uniquely gained and lost miRNAs

is solely based on human data. Although unlikely, but it is possible that the miRNAs are enriched in different tissues in different species. So, we should not completely rely on the predicted tissue specificity data unless they are experimentally verified.

4.4.3 Conservation of miRNAs and their targets: the use of GERP scores for detecting conservation of the non-coding region of the genome

This study has shown that the miRNA genes are highly conserved non-coding region of the genome. The globally expressed, primate miRNAs are under strong purifying selection and under significantly stronger evolutionary constraint ($P < 0.01$) (Fig. 4.10). On the contrary, the newer miRNAs, which are either species specific or lineage specific, are under virtually no constraint and evolving neutrally (Fig. 4.10). This result supports the previous studies showing the older, highly expressed miRNAs are under stronger purifying selection compared to the newer miRNAs (Liang and Li 2009, Berezikov et al. 2006, Nozawa et al. 2010, Ruby et al. 2007, Lu et al. 2008, Stark et al. 2007). However, I did not find any significant difference in evolutionary constraint among the *Homo*-only, *Homo-Pan* only, and Ape specific miRNAs. All of them are evolving neutrally. Since these miRNAs have evolved relatively recently and are expressed in limited number of species, they are probably not suitable target of purifying selection as previously shown in case of human specific miRNAs (Liang and Li 2009, Berezikov et al. 2006).

The miRNA target sites are under constraint in all cases (*Homo*-only, *Homo-Pan* only, African Ape specific, and Ape specific miRNAs). My result supports the previous study showing most mammalian mRNAs that are targets of miRNA regulation are highly conserved (Friedman et al. 2009). However, I found *Homo*-only miRNA target sites are

under significantly less ($P < 0.01$) evolutionary constraint compared to lineage specific miRNA targets and evolves almost neutrally. The potential absence of evolutionary constraint on *Homo*-only miRNA targets may account for the human specific gene regulation that evolved recently after its split from chimpanzee. For example, in case of the target sites of hsa-mir-941, the GERP score was only 0.49. hsa-mir-941 originated after human-chimpanzee split of ~6Mya and performs human brain specific gene regulation (Hu et al. 2012).

Another interesting finding of this study is that the conservation of miRNA target sites is highly significantly ($P < 0.01$) positively correlated to the binding of miRNAs to those target sites. The stronger the binding of miRNAs and their target sites, more constraint those target sites are. This observation supports the theory that miRNAs and their target genes have co-evolved (Barbash et al. 2014). The targets that have evolved with the miRNA, have stronger binding affinity for the miRNAs. These targets are highly conserved and are under strong evolutionary constraint. On the other hand, the targets where the miRNAs can bind theoretically, but may not bind *in vivo*, are under less constraint since these sites have not co-evolved with the miRNAs. As mentioned in section 4.2.5, all miRNA target prediction websites suffer from false positive problem and generate a huge list of miRNA targets, where miRNAs probably do not bind *in vivo*. The detection of the evolutionary constraints of those target sites can solve this false positive problem. The true miRNA target sites, since coevolved with the miRNA, will be under strong evolutionary constraint. Thus, we can consider the target sites with stronger evolutionary constraints as true miRNA targets and reject the rest.

One important aspect of this study is the use of Genomic Evolutionary Rate

profiling (GERP) scores for the detection of evolutionary constraint on miRNA genes and their targets. GERP score has been previously used for detecting the evolutionary constraint on CDS, UTRs, *cis*-regulatory elements, and intergenic regions across mammals (ENCODE project consortium 2012). Although used to detect the effect of purifying selection on non-coding regions of the genome (ENCODE project consortium 2012), the utility of GERP scores for determining the evolutionary constrained on miRNA genes and their targets, has never been explored. Instead 'branch-length score' (BLS), a multivalued statistic that accounts for phylogenetic relationships among the species studied (Kheradpour et al. 2007), has been used previously to detect the conservation of miRNA genes and their target sites (Friedman et al. 2009). I decided to use GERP scores for detecting evolutionary constraint because of two major reasons. Firstly, the GERP score is readily available through UCSC since GERP++ is implemented in UCSC Genome Browser (Davydov et al. 2010). The second reason is technical. GERP score is determined independently for each base position. Determining a constraint score for every base position is very important for short nucleotide sequences like miRNAs. If two sequences (miRNA and its homolog in another species) differ even in one base position the GERP score can change substantially. Also GERP score, unlike BLS score, is not susceptible to neighborhood nucleotide buffering (Spivakov et al. 2012). As a result, even if the surrounding area is highly conserved, a true constraint score can be obtained for a particular base position.

One major limitation of using GERP score for detecting the effect of selection on the non-coding region of the genome is its inability to detect positive selection. GERP scores were designed to detect evolutionary constraint on different coding and non-

coding elements of the genome (Cooper et al. 2005). So, GERP score can efficiently detect the effect of purifying selection on various elements of the genome. The higher the GERP score the more constraint the elements are and the stronger the purifying selection is. However, GERP considers any genomic element with a score ≤ 0 to be evolving neutrally. Even if the GERP score is $<< 0$, we still cannot designate that area to be under positive selection. So, GERP score is not a good statistic to detect the effect of selection for the fast evolving regions of the genome. It is a great statistic for highly conserved genome elements like miRNAs, where GERP can efficiently detect the degree of purifying selection operating in that region.

4.4.4 Prediction of miRNA targets and their potential biological function

As mentioned before, all miRNA target site prediction websites suffer from false positive problems. After preliminary analysis I chose DIANA server for predicting the miRNA targets. DIANA was chosen because it generates the highest number of correctly predicted targets than any other prediction tools (Maragkakis et al. 2009, Satoh and Tabunoki 2011). DIANA calculates the miRNA-targeted gene (miTG) score by measuring the weighted sum of the scores of all conserved and non-conserved miRNA targets on the 3'UTR and the CDS of the mRNA (Maragkakis et al. 2009) and this score correlates with the fold changes in suppression of protein expression (Satoh and Tabunoki 2011). This makes miTG score one of the strongest statistics for the determination of miRNA target sites. Also, unlike most other prediction servers, DIANA considers both 3'UTR and CDS as the miRNA targets (Satoh and Tabunoki 2011). This makes DIANA server more versatile and reliable in predicting novel miRNA targets. Although 0.2 miTG score has been shown to detect miRNA targets efficiently without

false positive problem (Satoh and Tabunoki 2011), I chose a more stringent cut off of 0.7 (default cut-off of the server) during the detection of miRNA targets of uniquely gained miRNAs.

Supporting the previous studies mentioned before, the target genes of uniquely human gained miRNAs are over-represented in various brain-related biological functions including Gene Ontology (GO) terms such as 'synaptic transmission', 'nervous system development', and 'neurological system process'. Beside brain-related functions, the target genes of human specific miRNAs are also over-represented in various housekeeping functions including cell cycle regulation, transcription, homeostasis, cellular and biological regulation, and developmental processes. Targets of the miRNAs, uniquely shared between human and chimpanzee are mostly over-represented in various housekeeping functions such as metabolic processes, cell-cell communication, cell proliferation, transcription, and splicing. Interestingly, the GO term 'nervous system development' was also found to be associated with the target genes of the miRNAs shared between human and chimpanzee, and human, chimpanzee and gorilla. The targets of the miRNAs shared uniquely among all apes were also significantly ($P < 0.01$) over-represented in various brain-related functions such as 'synaptic transmission', 'nervous system development', 'neurotransmitter secretion', and 'neurological system process' beside the essential housekeeping functions. These results clearly suggest the rapid evolution of brain specific gene regulation in the hominoid lineage after the apes split from the old world monkeys ~30Mya.

Since in recent years many miRNA target sites have been found within CDS of mRNA transcripts, more and more authors think the number of target sites within CDS is

thought to be as numerous as the target sites within 3'UTR (Forman et al. 2008, Qin et al., 2010, Ott et al. 2011, Huang et al. 2010, Hausser et al. 2013). However, I found that all uniquely gained miRNAs (*Homo* only, *Homo-Pan* only, African ape specific, and ape specific) preferentially target the 3'UTR region of the mRNAs. The target mRNAs are either regulated at both 3'UTR and CDS regions or solely at the 3'UTR region. Very few targets are uniquely regulated at the CDS. Also, I did not find any mRNA targeted at the 5'UTR region. Again, since the current study is completely computational, the target sites are not experimentally validated. This study is limited to the target coordinates provided by DIANA server. It is possible that more uniquely gained miRNAs solely target the CDS and even 5'UTR *in vivo*, which were not predicted *in silico*.

4.4.5 MirSNPs and the limitation of *in silico* prediction of novel miRNAs

As shown in Table 4.5, four SNPs: rs33982250, rs5829384, rs113974396, and rs10575780 found in the miRNA genes mir-1303, mir-3125, mir-3916, and mir-3938 respectively have retained their ancestral insertions or deletions in more than 5% human population. In this study, I have defined uniquely gained and lost miRNAs based on unique species specific or lineage specific insertion/deletions. According to this definition, if one species or lineage does not possess the same insertion and/or deletion like the miRNA gene in question, that miRNA is thought to be absent or non-functional in that species/lineage. But since more than 5% human population have retained the ancestral insertion/deletion state, it is very difficult to predict whether they possess a non-functional copy of the miRNA gene (by definition) or the one/two bp insertion/deletion does not affect the function of the miRNA *in vivo*. Interestingly, two out of four above-mentioned indels (rs33982250 and rs113974396) do not change the pre-miRNA stem-

loop structure and one (rs5829384) only marginally affects the stem-loop structure (See Appendix 3.2). Only rs10575780 significantly changes the miRNA stem-loop structure. In cases where the insertion-deletion does not change the stem-loop structure, it is very difficult to predict whether that indel has any impact on the miRNA function *in vivo*. I think this is the major limitation of *in silico* prediction of novel miRNAs and even more for the prediction of uniquely gained and lost miRNAs. The *in silico* predicted novel miRNAs should be experimentally validated by deep sequencing techniques to confirm their uniqueness in a particular species or lineage.

4.4.6 Future directions: application of selection based tests on miRSNPs

As mentioned in the previous paragraph, SNPs found in the uniquely gained miRNAs in humans may or may not change the structure of the miRNA and thus may or may not affect its function. It is interesting to investigate whether the SNPs that creates or disrupts miRNA genes are evolving neutrally or under selective pressure. Similar question has been addressed recently by accessing whether SNPs that disrupt or create miRNA recognition element (MRE) seed sites (MRESS) are under selection (Richardson et al. 2011). The authors identified more than 2700 SNPs that disrupt and more than 22000 SNPs that create MRESSs. To determine whether the SNPs that create or disrupt predicted MRESSs are under positive selection, they used genome wide F_{ST} calculations from HapMap Phase 3 data. They found that the MRESS SNPs and the SNPs that create novel MRESS (CNM) are under strong positive selection. Interestingly they identified a SNP that creates novel MRESS for has-mir-3916 (uniquely human gained miRNA) in the 3'UTR of *ENAM* under strong positive selection (F_{ST} = 0.8942). So, it can be speculated that not only novel miRNAs, but also their target sites show variation in the population.

Similar F_{ST} based approach can be employed to investigate whether the SNPs that create or disrupt novel miRNA genes are under positive selection. Additionally, extended haplotype homozogysity (EHH) approach can also be employed to detect whether these miRSNPs are under positive selection (Sabeti et al 2002, Sabeti et al. 2007). It would be very interesting to find out whether the above-mentioned miRSNPs are under any selective sweep and whether that sweep is associated with certain human population. Since the recent draft of 1000 genome project (McVean et al. 2012) contain huge amount of information about human population genetics, we are currently in a great position to employ the above-mentioned tests and find out the biological importance of the miRSNPs.

References

Artzi S, Kiezun A, Shomron N 2008. miRNAminer:a tool for homologous microRNA gene search. BMC Bioinformatics 9:39.

Baev V, Daskalova E, Minkov I 2008. Computational identification of novel microRNA homologs in the chimpanzee genome. Comput Biol Chem. 33:62-70.

Barbash S, Shifman S, Soreq H 2014. Global Coevolution of Human MicroRNAs and Their Target Genes. Mol Biol Evol. 31:1237-1247.

Bartel DP 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281-297.

Bartel DP, Chen CZ 2004. Micromanagers of gene expression:the potentially widespread influence of metazoan microRNAs. Nat Rev Genet. 5:396-400.

Behm-Ansmant I, Rehwinkel J, Doerks T, Stark A, Bork P, Izaurralde E 2006. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev. 20:1885-1898.

Berezikov E, Guryev V, van de Belt J, et al. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. Cell 120:21-24.

Berezikov E, Robine N, Samsonova A, et al. 2011. Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. Genome Res. 21:203-215.

Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RHA 2006. Diversity of microRNAs in human and chimpanzee brain. Nat Genet. 38:1375-1377.

Buffa FM, Camps C, Winchester L, Snell CE, Gee HE, Sheldon H, Taylor M, Harris AL, Ragoussis J 2011. microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. Cancer Res. 71:5635-5645.

Campo-Paysaa F, Semon M, Cameron RA, Peterson KJ, Schubert M 2011. microRNA complements in deuterostomes:origin and evolution of microRNAs. Evol Dev. 13:15-27.

Carthew RW, Sontheimer EJ 2009. Origins and Mechanisms of miRNAs and siRNAs. Cell 136:642-655.

Chen K, Rajewsky N 2007. The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet. 8:93-103.

Cooper GM, Stone EA, Asimenos G, Comparative Sequencing Program N 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15:901-913.

Dannemann M, Nickel B, Lizano E, Burbano HA, Kelso J 2012. Annotation of primate miRNAs by high throughput sequencing of small RNA libraries. BMC Genomics 13:116.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP. PLoS Comput Bio. 6:e1001025.

Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ 2004. Processing of primary microRNAs by the Microprocessor complex. Nature 432:231-235.

Filipowicz W, Bhattacharyya S, Sonenberg N 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight. Nat Rev Genet. 9:102-114.

Flynt AS, Lai EC 2008. Biological principles of microRNA-mediated regulation: shared themes amid diversity. Nat Rev Genet. 9:831-842.

Forman JJ, Legesse-Miller A, Coller HA 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc Natl Acad Sci. 105:14879-14884.

Friedman RC, Farh KK, Burge CB, Bartel DP 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 19:92-105.

Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF 2006. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science 312:75-79.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ 2008. miRBase:tools for microRNA genomics. Nucleic Acids Res. 36:D154-D158.

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature 455:1193-1197.

Gu X, Su Z, Huang Y 2009. Simultaneous expansions of microRNAs and protein-coding genes by gene/genome duplications in early vertebrates. J Exp Zool B Mol Dev Evol. 312B:164-170.

Hausser J, Syed AP, Bilen B, et al. 2013. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. Genome Res. 23:1615-1623.

Heimberg AM, Sempere LF, Moy VN, Donoghue PC, Peterson KJ 2008. MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci. 105:2946-2950.

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker I, Stadler P 2006. The expansion of the metazoan microRNA repertoire. BMC Genomics 7:25.

Higashijima A, Miura K, Mishima H, Kinoshita A, et al. 2013. Characterization of placenta-specific microRNAs in fetal growth restriction pregnancy. Prenat Diagn. 33:214-222.

Hu HY, Guo S, Xi J, Yan Z, et al. 2011. MicroRNA expression and regulation in human, chimpanzee, and macaque brains. PLoS Genet. 7:e1002327.

Hu HY, He L, Fominykh K, Yan Z, et al. 2012. Evolution of the human-specific microRNA miR-941. Nat commun. 3:1145.

Huaiyu Mi MATPD 2012. PANTHER in 2013:modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41:D377-386.

Huang FWD, Qin J, Reidys CM, Stadler PF 2010. Target prediction and a statistical sampling algorithm for RNA-RNA interaction. Bioinformatics 26:175-181.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS 2005. Human MicroRNA targets. PLoS Biol. 3:e264.

Kaya KD, Karakulah G, Yakicier CM, Acar AC, Konu O 2011. mESAdb:microRNA expression and sequence analysis database. Nucleic Acids Res. 39:D170-180.

Kent WJ 2002. BLAT-the BLAST-like alignment tool. Genome Res. 12:656-664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D 2002. The human genome browser at UCSC. Genome Res. 12:996-1006.

Kheradpour P, Stark A, Roy S, Kellis M 2007. Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res. 17:1919-1931.

Kim VN 2005. MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol. 6:376-385.

Kloosterman W, Plasterk RHA 2006. The diverse functions of microRNAs in animal development and disease. Dev Cell. 11:441-450.

Lai EC, Tomancak P, Williams RW, Rubin GM 2003. Computational identification of Drosophila microRNA genes. Genome Biol Evol. 4:R42.

Lee Y, Ahn C, Han J, Choi H, et al. 2003. The nuclear RNase III Drosha initiates microRNA processing. Nature 425:415-419.

Levine M, Tjian R 2003. Transcription regulation and animal diversity. Nature 424:147-151.

Lewis BP, Burge CB, Bartel DP 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120:15-20.

Li J, Liu Y, Dong D, Zhang Z 2009. Evolution of an X-Linked Primate-Specific Micro RNA Cluster. Mol Biol Evol. 27:671-683.

Li Y, Wang F, Lee JA, Gao FB 2006. MicroRNA-9a ensures the precise specification of sensory organ precursors in Drosophila. Genes Dev. 20:2793-2805.

Liang H, Li WH 2009. Lowly expressed human microRNA genes evolve rapidly. Mol Biol Evol. 26:1195-1198.

Lim LP, Glasner ME, Yekta S, et al. 2003. Vertebrate microRNA genes. Science 299:1540.

Lim LP, Lau NC, Weinstein EG, et al. 2003. The microRNAs of Caenorhabditis elegans. Genes Dev. 17:991-1008.

Liu C, Zhang F, Li T, Lu M, Wang L, Yue W, Zhang D 2012. MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. BMC Genomics 13:661.

Liu N, Okamura K, Tyler DM, Phillips MD, Chung WJ, Lai EC 2008. The evolution and functional diversification of animal microRNA genes. Cell Res. 18:985-996.

Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu CI 2008. The birth and death of microRNA genes in Drosophila. Nat Genet. 40:351-355.

Lu M, Shi B, Wang J, Cao Q, Cui Q 2010. TAM: A method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. BMC Bioinformatics 11:419.

Lu M, Zhang Q, Deng M, Miao J, Guo Y, et al. 2008. An analysis of human microRNA and disease associations. PLoS One 3:e3420.

McVean, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65.

Meiri E, Levy A, Benjamin H, Ben-David M, et al. 2010. Discovery of microRNAs and other small RNAs in solid tumors. Nucleic Acids Res. 38:6234-6246.

Nam JW, Shin KR, Han J, et al. 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res. 33:3570-3581.

Navon R, Wang H, Steinfeld I, Tsalenko A, Ben-Dor A, Yakhini Z 2009. Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types. PLoS One 4:e8003.

Niwa R, Slack FJ 2007. The evolution of animal microRNA function. Curr Opin Genet Dev. 17:145-150.

Nozawa M, Miura S, Nei M 2010. Origins and evolution of MicroRNA genes in Drosophila species. Genome Biol Evol. 2:180-189.

Ohler U, Yekta S, Lim LP, et al. 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA 10:1309-1322.

Okamura K, Phillips MD, Tyler DM, Duan H, Chou YT, Lai EC 2008. The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. Nat Struct Mol Biol. 15:354-363.

Ott CE, Grunhagen J, Jager M, Horbelt D, et al. 2011. MicroRNAs di_erentially expressed in postnatal aortic development downregulate elastin via 3' UTR and coding-sequence binding sites. PLoS One 6:e16250.

Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG 2013. DIANA-microT web server v5.0:service integration into miRNA functional analysis workflows. Nucleic Acids Res. 41:W169-173.

Peterson KJ, Dietrich MR, McPeek MA 2009. MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. BioEssays 31:736-747.

Piriyapongsa J, Mariño-Ramírez L JKI 2007. Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323-1337.

Prochnik S, Rokhsar D, Aboobaker A 2007. Evidence for a microRNA expansion in the bilaterian ancestor. Dev Genes Evol. 217:73-77.

Project Consortium E 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

Qin W, Shi Y, Zhao B, Yao C, Jin L, Ma J, Jin Y 2010. miR-24 regulates apoptosis by targeting the open reading frame (ORF) region of FAF1 in cancer cells. PLoS One 5:e9429.

Richardson K, Lai CQ, Parnell LD, Lee YC, Ordovas JM 2011. A genome-wide survey for SNPs altering microRNA seed sites identifies functional candidates in GWAS. BMC Genomics 12:504.

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome Res. 17:1850-1864.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Sabeti PC, Varilly P, Fry B, Lohmueller J 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-919.

Satoh J, Tabunoki H 2011. Comprehensive analysis of human microRNA target networks. BioData Min. 4:17.

Sewer A, Paul N, Landgraf P, et al. 2005. Identification of clustered microRNAs using an ab initio prediction method. BMC Bioinformatics 6:267.

Shabalina SA, Koonin EV 2008. Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol. 23:578-587.

Somel M, Guo S, Fu N, Yan Z, et al. 2010. MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. Genome Res. 20:1207-1218.

Somel M, Liu X, Tang L, Yan Z, et al. 2011. MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. PLoS Biol. 9:e1001214.

Spivakov M, Akhtar J, Kheradpour P, Beal K, Girardot C, Koscielny G, Herrero J, Kellis M 2012. Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol Evol. 13:R49.

Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM 2005. Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell 123:1133-1146.

Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, Kellis M 2007. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. Genome Res. 17:1865-1879.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol. 28:2731-2739.

Wakiyama M, Takimoto K, Ohara O, Yokoyama S 2007. Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. Genes Dev. 21:1857-1862.

Weber MJ 2005. New human and mouse microRNA genes found by homology search. Febs J. 272:59-73.

Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, Peterson KJ 2009. The deep evolution of metazoan microRNAs. Evol Dev. 11:50-68.

Wu CI, Shen Y, Tang T 2009. Evolution under canalization and the dual roles of microRNAs: a hypothesis. Genome Res. 19:734-743.

Wu L, Fan J, Belasco JG 2006. MicroRNAs direct rapid deadenylation of mRNA. Proc Natl Acad Sci. 103:4034-4039.

Xie X, Lu J, Kulbokas EJ, et al. 2005. Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals. Nature 434:338-345.

Yoon S, Micheli GD 2006. Computational Identification of MicroRNAs and Their Targets. Comput Biol Chem. 78:118-128.

Yuan Z, Liu H, Nie Y, Ding S, Yan M, Tan S, Jin Y, Sun X 2013. Identification of novel microRNAs in primates by using the synteny information and small RNA deep sequencing data. Int J Mol Sci. 14:20820-20832.

Zhang R, Peng Y, Wang W, Su B 2007. Rapid evolution of an X-linked microRNA cluster in primates. Genome Res. 17:612-617.

Zhang R, Wang YQ, Su B 2008. Molecular evolution of a primate specific microRNA family. Mol Biol Evol. 25:1493-1502.

Chapter 5: Final thoughts and future directions

All three chapters of my dissertation are independent studies yet they are linked to each other with the central concept of molecular evolution of hominoid primates. My dissertation focuses on the evolution of both coding and non-coding regulatory regions of the genome among hominoids. Also, the dissertation focuses on the evolution of hominoids at all levels: below species level (population genetics), species level (speciation and divergence), and above species level (molecular evolution).

The second chapter, although majorly focuses on the reconstruction of gorilla phylogeny, it compares and contrasts gorilla mitochondrial genome with other hominoid mitochondrial genomes, especially with that of chimpanzee and bonobo. It gives this chapter a broader perspective in respect to hominoid evolution. Although the overall degree of anatomical and molecular differentiation between eastern and western gorillas was clearly greater than between any chimpanzee subspecies, and equivalent to other sister species pairs in primates (Groves 2001), traditionally, the living gorilla populations were considered a single species (*Gorilla gorilla*) with three recognized subspecies (*G. g. gorilla, G. g. beringei, G. g. graueri*) (Groves 2003). The 1000 bootstrap replicate scatterplot of genetic distances between the chimpanzee subspecies and the same between *Gorilla* species (Fig. 2.11) clearly showed the difference in genetic divergence between the two with no overlap. However, the similar plot with *Pan* species and *Gorilla* species shows ~70% overlap in the genetic distances (Fig. 2.10). The two above-mentioned scatter plots (Fig. 2.10 and 2.11) strongly supported that Western-Eastern gorilla divergence is equivalent to sister species pairs in primates.

The third chapter, which comprises the core of my dissertation, highlights the evolution of *cis*-regulatory regions among hominoids. King and Wilson (1975) hypothesized that the differences among hominoid primates lie in the regulatory sequences. This hypothesis has been shown to be true in several previous studies (See Chapter 3). This study supported this hypothesis and showed highly significant (0.01) differential promoter activity among the hominoid primates for *CRTAC1*. *CRTAC1* is highly conserved among the hominoids and it shows several characters similar to a housekeeping gene (See Chapter 3 discussion). This chapter highlights the complexity of eukaryotic gene regulation, discussing how a potential housekeeping gene can get tissue specifically up/down-regulated in the presence of 'enhancer' and 'silencer' elements. Although this chapter showed that the 'silencer' brings down transcription of human *CRTAC1* construct to a level similar to chimpanzee promoter-only constructs, it could not conclusively explain the differential expression of CRTAC1 in human and chimpanzee seminal plasma (chimpanzee CRTAC1 expression more than 140 fold higher than human). There must be additional tissue specific enhancers and/or silencers involved in the regulation of *CRTAC1*. Since techniques like chromatin conformation capture (3C) was not employed in this study, it was not possible to pinpoint the right enhancer/silencer elements that are additionally involved in the regulation of *CRTAC1*. There are two major issues in this chapter, which should be experimentally addressed in the future. Firstly, from the current data, it seems like human 'silencer' is stronger than chimpanzee 'silencer' in driving transcriptional repression. To confirm this experimentally, the 'silencer' elements should be swapped between human and chimpanzee, i.e. human promoter with chimpanzee silencer and vice versa. Secondly, since gorilla putative promoter could not be cloned

into the pGL4.10 vector, we do not know *CRTAC1* expression pattern in gorilla. Using gorilla BAC library, I have successfully amplified gorilla *CRTAC1* putative promoter region. As a follow up, the amplified gorilla putative promoter should be cloned into pGL4.10 vector and transfected into LNCaP cells to investigate the expression of gorilla *CRTAC1* 'promoter'. Finally, the regulatory differences seen between human and chimpanzee came from *in vitro* luciferase assays. Considering the complexity of eukaryotic gene regulation *in vivo,* CRTAC1 regulation may be completely different than what showed in *in vitro* luciferase assays and may not be explained by the simple promoter and/or promoter + single enhancer constructs.

The fourth chapter of my dissertation is completely computational. The uniquely gained and lost miRNA that are listed in Table 4.1 are predicted computationally. Their uniqueness is not experimentally validated. As discussed in Section 4.4.5 (See Chapter 4 discussion), the computationally predicted novel miRNAs are susceptible to mirSNP issues and it becomes very difficult to predict whether one/two bp insertion/deletion affect the function of the miRNA *in vivo*. I strongly think that the *in silico* predicted novel miRNAs should be experimentally validated by deep sequencing techniques to confirm their uniqueness in a particular species or lineage as done for mir-941 (Hu et al. 2012). This study was also limited by the unavailability of chimpanzee, gorilla, and orangutan specific miRNA expression studies. It is possible that the miRNAs are enriched in different sets of tissues and/or regulate different set of genes in other non-human hoiminoids. So, we should not completely rely on the predicted tissue specificity and disease data for other hominoids unless they are experimentally verified.

References

Groves CP 2001. Primate Taxonomy. Washington and London: Smithsonian Institution Press.

Groves CP 2003. A history of gorilla taxonomy. In: Taylor AB GM, editor. Gorilla Biology: a multidisciplinary perspective: Cambridge University Press.

Hu HY, He L, Fominykh K, Yan Z, et al. 2012. Evolution of the human-specific microRNA miR-941. Nat Commun. 3:1145.

King MC, Wilson AC 1975. Evolution at two levels in humans and chimpanzees. Science 188:107-116.

Appendix 1: Data from Chapter 2

1.1 BEAST files (without sequences)

1.1.1 12 heavy strand genes and Whole mtDNA without D-Loop datasets (without GAGP Gorillas)

```
<?xml version="1.0" standalone="yes"?>
<!-- Generated by BEAUTi v1.7.5
-->
<!--       by Alexei J. Drummond and Andrew Rambaut
-->
<!--       Department of Computer Science, University of
Auckland and        -->
<!--       Institute of Evolutionary Biology, University of
Edinburgh        -->
<!--       http: //beast.bio.ed.ac.uk/
-->
<beast>


     <!-- The list of taxa analyse (can also include
dates/ages).                 -->
     <!-- ntax=15
-->
     <taxa id="taxa">
          <taxon id="Pongo_abelii"/>
          <taxon id="Pongo_pygmaeus"/>
          <taxon id="westerngorilla011120"/>
          <taxon id="westerngorillachipua"/>
          <taxon id="easterngorillamkubwa"/>
          <taxon id="P.t.verus"/>
          <taxon id="P.t.ellioti"/>
          <taxon id="P.t.troglodytes"/>
          <taxon id="P.t.schweinfurthii"/>
          <taxon id="bonoboisolatePP30"/>
          <taxon id="bonoboisolatePP75"/>
          <taxon id="human"/>
          <taxon id="neanderthal"/>
          <taxon id="gibbon"/>
          <taxon id="macaque"/>
```

```
</taxa>
<taxa id="African Apes">
     <taxon idref="P.t.ellioti"/>
     <taxon idref="P.t.schweinfurthii"/>
     <taxon idref="P.t.troglodytes"/>
     <taxon idref="P.t.verus"/>
     <taxon idref="bonoboisolatePP30"/>
     <taxon idref="bonoboisolatePP75"/>
     <taxon idref="easterngorillamkubwa"/>
     <taxon idref="human"/>
     <taxon idref="neanderthal"/>
     <taxon idref="westerngorilla011120"/>
     <taxon idref="westerngorillachipua"/>
</taxa>
<taxa id="Chimp-Bonobo">
     <taxon idref="P.t.ellioti"/>
     <taxon idref="P.t.schweinfurthii"/>
     <taxon idref="P.t.troglodytes"/>
     <taxon idref="P.t.verus"/>
     <taxon idref="bonoboisolatePP30"/>
     <taxon idref="bonoboisolatePP75"/>
</taxa>
<taxa id="Gorilla">
     <taxon idref="easterngorillamkubwa"/>
     <taxon idref="westerngorilla011120"/>
     <taxon idref="westerngorillachipua"/>
</taxa>
<taxa id="Hominid">
     <taxon idref="P.t.ellioti"/>
     <taxon idref="P.t.schweinfurthii"/>
     <taxon idref="P.t.troglodytes"/>
     <taxon idref="P.t.verus"/>
     <taxon idref="Pongo_abelii"/>
     <taxon idref="Pongo_pygmaeus"/>
     <taxon idref="bonoboisolatePP30"/>
     <taxon idref="bonoboisolatePP75"/>
     <taxon idref="easterngorillamkubwa"/>
     <taxon idref="human"/>
     <taxon idref="neanderthal"/>
     <taxon idref="westerngorilla011120"/>
     <taxon idref="westerngorillachipua"/>
</taxa>
<taxa id="Hominoid">
     <taxon idref="P.t.ellioti"/>
```

```
        <taxon idref="P.t.schweinfurthii"/>
        <taxon idref="P.t.troglodytes"/>
        <taxon idref="P.t.verus"/>
        <taxon idref="Pongo_abelii"/>
        <taxon idref="Pongo_pygmaeus"/>
        <taxon idref="bonoboisolatePP30"/>
        <taxon idref="bonoboisolatePP75"/>
        <taxon idref="easterngorillamkubwa"/>
        <taxon idref="gibbon"/>
        <taxon idref="human"/>
        <taxon idref="neanderthal"/>
        <taxon idref="westerngorilla011120"/>
        <taxon idref="westerngorillachipua"/>
    </taxa>
    <taxa id="Human-Chimp">
        <taxon idref="P.t.ellioti"/>
        <taxon idref="P.t.schweinfurthii"/>
        <taxon idref="P.t.troglodytes"/>
        <taxon idref="P.t.verus"/>
        <taxon idref="bonoboisolatePP30"/>
        <taxon idref="bonoboisolatePP75"/>
        <taxon idref="human"/>
        <taxon idref="neanderthal"/>
    </taxa>
    <taxa id="Pongo">
        <taxon idref="Pongo_abelii"/>
        <taxon idref="Pongo_pygmaeus"/>
    </taxa>
    <taxa id="Western Gorilla">
        <taxon idref="westerngorilla011120"/>
        <taxon idref="westerngorillachipua"/>
    </taxa>
    <taxa id="Bonobo">
        <taxon idref="bonoboisolatePP30"/>
        <taxon idref="bonoboisolatePP75"/>
    </taxa>
    <taxa id="Chimp">
        <taxon idref="P.t.ellioti"/>
        <taxon idref="P.t.schweinfurthii"/>
        <taxon idref="P.t.troglodytes"/>
        <taxon idref="P.t.verus"/>
    </taxa>
    <taxa id="Human">
        <taxon idref="human"/>
```

```
          <taxon idref="neanderthal"/>
     </taxa>

     <!-- The sequence alignment (each sequence refers to a
taxon above).         -->
     <!-- ntax=15 nchar=10907
-->
     <alignment id="alignment" dataType="nucleotide">
          <sequence>
               <taxon idref="…"/>
</sequence>
     </alignment>

     <!-- The unique patterns for codon positions 1 & 2
-->
     <mergePatterns id="patterns1+2">
          <!-- The unique patterns for codon position 1
-->
          <!-- npatterns=623
-->
          <patterns id="patterns1" from="1" every="3">
               <alignment idref="alignment"/>
          </patterns>
          <!-- The unique patterns for codon position 2
-->
          <!-- npatterns=301
-->
          <patterns id="patterns2" from="2" every="3">
               <alignment idref="alignment"/>
          </patterns>
     </mergePatterns>
     <!-- The unique patterns for codon position 3
-->
     <!-- npatterns=1513
-->
     <patterns id="patterns3" from="3" every="3">
          <alignment idref="alignment"/>
     </patterns>

     <!-- A prior on the distribution node heights defined
given                  -->
     <!-- a Yule speciation process (a pure birth process).
-->
     <yuleModel id="yule" units="substitutions">
```

```
          <birthRate>
               <parameter id="yule.birthRate" value="1.0"
lower="0.0" upper="Infinity"/>
          </birthRate>
     </yuleModel>
     <!-- This is a simple constant population size
coalescent model              -->
     <!-- that is used to generate an initial tree for the
chain.                 -->
     <constantSize id="initialDemo" units="substitutions">
          <populationSize>
               <parameter id="initialDemo.popSize"
value="100.0"/>
          </populationSize>
     </constantSize>

     <!-- Generate a random starting tree under the
coalescent process            -->
     <coalescentTree id="startingTree">
          <constrainedTaxa>
               <taxa idref="taxa"/>
               <tmrca monophyletic="false">
                    <taxa idref="Western Gorilla"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Gorilla"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Hominid"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Hominoid"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Bonobo"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Human-Chimp"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="African Apes"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Chimp-Bonobo"/>
```

```
                </tmrca>
                <tmrca monophyletic="false">
                     <taxa idref="Pongo"/>
                </tmrca>
                <tmrca monophyletic="false">
                     <taxa idref="Bonobo"/>
                </tmrca>
                <tmrca monophyletic="false">
                     <taxa idref="Human"/>
                </tmrca>
           </constrainedTaxa>
           <constantSize idref="initialDemo"/>
     </coalescentTree>

     <treeModel id="treeModel">
           <coalescentTree idref="startingTree"/>
           <rootHeight>
                <parameter id="treeModel.rootHeight"/>
           </rootHeight>
           <nodeHeights internalNodes="true">
                <parameter
id="treeModel.internalNodeHeights"/>
           </nodeHeights>
           <nodeHeights internalNodes="true"
rootNode="true">
                <parameter
id="treeModel.allInternalNodeHeights"/>
           </nodeHeights>
     </treeModel>

     <speciationLikelihood id="speciation">
           <model>
                <yuleModel idref="yule"/>
           </model>
           <speciesTree>
                <treeModel idref="treeModel"/>
           </speciesTree>
     </speciationLikelihood>

     <!-- The uncorrelated relaxed clock (Drummond, Ho,
Phillips & Rambaut, 2006) -->
     <discretizedBranchRates id="branchRates">
           <treeModel idref="treeModel"/>
           <distribution>
```

```
                <logNormalDistributionModel
meanInRealSpace="true">
                    <mean>
                        <parameter id="ucld.mean"
value="0.038" lower="0.0" upper="100.0"/>
                    </mean>
                    <stdev>
                        <parameter id="ucld.stdev"
value="0.1" lower="0.0" upper="10.0"/>
                    </stdev>
                </logNormalDistributionModel>
            </distribution>
            <rateCategories>
                <parameter id="branchRates.categories"
dimension="26"/>
            </rateCategories>
        </discretizedBranchRates>

        <rateStatistic id="meanRate" name="meanRate"
mode="mean" internal="true" external="true">
            <treeModel idref="treeModel"/>
            <discretizedBranchRates idref="branchRates"/>
        </rateStatistic>

        <rateStatistic id="coefficientOfVariation"
name="coefficientOfVariation" mode="coefficientOfVariation"
internal="true" external="true">
            <treeModel idref="treeModel"/>
            <discretizedBranchRates idref="branchRates"/>
        </rateStatistic>

        <rateCovarianceStatistic id="covariance"
name="covariance">
            <treeModel idref="treeModel"/>
            <discretizedBranchRates idref="branchRates"/>
        </rateCovarianceStatistic>

        <!-- The HKY substitution model (Hasegawa, Kishino &
Yano, 1985)             -->
        <hkyModel id="hky1">
            <frequencies>
                <frequencyModel dataType="nucleotide">
                    <alignment idref="alignment"/>
                    <frequencies>
```

```
                        <parameter id="hky1.frequencies"
dimension="4"/>
                    </frequencies>
               </frequencyModel>
          </frequencies>
          <kappa>
               <parameter id="hky1.kappa" value="1.0"
lower="1.0E-8" upper="Infinity"/>
          </kappa>
     </hkyModel>
     <!-- The HKY substitution model (Hasegawa, Kishino &
Yano, 1985)              -->
     <hkyModel id="hky2">
          <frequencies>
               <frequencyModel dataType="nucleotide">
                    <alignment idref="alignment"/>
                    <frequencies>
                         <parameter id="hky2.frequencies"
dimension="4"/>
                    </frequencies>
               </frequencyModel>
          </frequencies>
          <kappa>
               <parameter id="hky2.kappa" value="1.0"
lower="1.0E-8" upper="Infinity"/>
          </kappa>
     </hkyModel>

     <!-- site model
-->
     <siteModel id="siteModel1">
          <substitutionModel>
               <hkyModel idref="hky1"/>
          </substitutionModel>
          <relativeRate>
               <parameter id="siteModel1.mu" value="1.0"
lower="0.0" upper="Infinity"/>
          </relativeRate>
          <gammaShape gammaCategories="4">
               <parameter id="siteModel1.alpha" value="0.5"
lower="0.0" upper="Infinity"/>
          </gammaShape>
     </siteModel>
```

```
     <!-- site model
-->
     <siteModel id="siteModel2">
          <substitutionModel>
               <hkyModel idref="hky2"/>
          </substitutionModel>
          <relativeRate>
               <parameter id="siteModel2.mu" value="1.0"
lower="0.0" upper="Infinity"/>
          </relativeRate>
          <gammaShape gammaCategories="4">
               <parameter id="siteModel2.alpha" value="0.5"
lower="0.0" upper="Infinity"/>
          </gammaShape>
     </siteModel>

     <compoundParameter id="allMus">
          <parameter idref="siteModel1.mu"/>
          <parameter idref="siteModel2.mu"/>
     </compoundParameter>

     <treeLikelihood id="treeLikelihood1">
          <patterns idref="patterns1+2"/>
          <treeModel idref="treeModel"/>
          <siteModel idref="siteModel1"/>
          <discretizedBranchRates idref="branchRates"/>
     </treeLikelihood>
     <treeLikelihood id="treeLikelihood2">
          <patterns idref="patterns3"/>
          <treeModel idref="treeModel"/>
          <siteModel idref="siteModel2"/>
          <discretizedBranchRates idref="branchRates"/>
     </treeLikelihood>


     <tmrcaStatistic id="tmrca(Western Gorilla)">
          <mrca>
               <taxa idref="Western Gorilla"/>
          </mrca>
          <treeModel idref="treeModel"/>
     </tmrcaStatistic>
     <tmrcaStatistic id="tmrca(Gorilla)">
          <mrca>
               <taxa idref="Gorilla"/>
```

```
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Hominid)">
     <mrca>
          <taxa idref="Hominid"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Hominoid)">
     <mrca>
          <taxa idref="Hominoid"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Chimp)">
     <mrca>
          <taxa idref="Chimp"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Human-Chimp)">
     <mrca>
          <taxa idref="Human-Chimp"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(African Apes)">
     <mrca>
          <taxa idref="African Apes"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Chimp-Bonobo)">
     <mrca>
          <taxa idref="Chimp-Bonobo"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Pongo)">
     <mrca>
          <taxa idref="Pongo"/>
     </mrca>
     <treeModel idref="treeModel"/>
```

```
    </tmrcaStatistic>
    <tmrcaStatistic id="tmrca(Bonobo)">
        <mrca>
            <taxa idref="Bonobo"/>
        </mrca>
        <treeModel idref="treeModel"/>
    </tmrcaStatistic>
    <tmrcaStatistic id="tmrca(Human)">
        <mrca>
            <taxa idref="Human"/>
        </mrca>
        <treeModel idref="treeModel"/>
    </tmrcaStatistic>
    <operators id="operators">
        <scaleOperator scaleFactor="0.75" weight="1">
            <parameter idref="hky1.kappa"/>
        </scaleOperator>
        <scaleOperator scaleFactor="0.75" weight="1">
            <parameter idref="hky2.kappa"/>
        </scaleOperator>
        <scaleOperator scaleFactor="0.75" weight="1">
            <parameter idref="siteModel1.alpha"/>
        </scaleOperator>
        <scaleOperator scaleFactor="0.75" weight="1">
            <parameter idref="siteModel2.alpha"/>
        </scaleOperator>
        <deltaExchange delta="0.75" parameterWeights="2
1" weight="1">
            <parameter idref="allMus"/>
        </deltaExchange>
        <scaleOperator scaleFactor="0.75" weight="3">
            <parameter idref="ucld.mean"/>
        </scaleOperator>
        <scaleOperator scaleFactor="0.75" weight="3">
            <parameter idref="ucld.stdev"/>
        </scaleOperator>
        <upDownOperator scaleFactor="0.75" weight="3">
            <up>
                <parameter idref="ucld.mean"/>
            </up>
            <down>
                <parameter
idref="treeModel.allInternalNodeHeights"/>
            </down>
```

```
          </upDownOperator>
          <swapOperator size="1" weight="10"
autoOptimize="false">
               <parameter idref="branchRates.categories"/>
          </swapOperator>
          <randomWalkIntegerOperator windowSize="1.0"
weight="10">
               <parameter idref="branchRates.categories"/>
          </randomWalkIntegerOperator>
          <uniformIntegerOperator weight="10">
               <parameter idref="branchRates.categories"/>
          </uniformIntegerOperator>
          <scaleOperator scaleFactor="0.75" weight="3">
               <parameter idref="yule.birthRate"/>
          </scaleOperator>
          <scaleOperator scaleFactor="0.75" weight="3">
               <parameter idref="treeModel.rootHeight"/>
          </scaleOperator>
          <uniformOperator weight="30">
               <parameter
idref="treeModel.internalNodeHeights"/>
          </uniformOperator>
          <subtreeSlide size="0.1" gaussian="true"
weight="15">
               <treeModel idref="treeModel"/>
          </subtreeSlide>
          <narrowExchange weight="15">
               <treeModel idref="treeModel"/>
          </narrowExchange>
          <wideExchange weight="3">
               <treeModel idref="treeModel"/>
          </wideExchange>
          <wilsonBalding weight="3">
               <treeModel idref="treeModel"/>
          </wilsonBalding>
     </operators>

     <mcmc id="mcmc" chainLength="10000000"
autoOptimize="true">
          <posterior id="posterior">
               <prior id="prior">
                    <logNormalPrior mean="0.0" stdev="0.56"
offset="5.0" meanInRealSpace="false">
```

```
                        <statistic idref="tmrca(Human-
Chimp)"/>
                    </logNormalPrior>
                    <speciationLikelihood
idref="speciation"/>
               </prior>
               <likelihood id="likelihood">
                    <treeLikelihood
idref="treeLikelihood1"/>
                    <treeLikelihood
idref="treeLikelihood2"/>
               </likelihood>
          </posterior>
          <operators idref="operators"/>
          <log id="screenLog" logEvery="1000">
               <column label="Posterior" dp="4" width="12">
                    <posterior idref="posterior"/>
               </column>
               <column label="Prior" dp="4" width="12">
                    <prior idref="prior"/>
               </column>
               <column label="Likelihood" dp="4"
width="12">
                    <likelihood idref="likelihood"/>
               </column>
               <column label="Root Height" sf="6"
width="12">
                    <parameter
idref="treeModel.rootHeight"/>
               </column>
               <column label="Rate" sf="6" width="12">
                    <rateStatistic idref="meanRate"/>
               </column>
          </log>
          <log id="fileLog" logEvery="1000"
fileName="12_heavy_strand_genes.log">
               <posterior idref="posterior"/>
               <prior idref="prior"/>
               <likelihood idref="likelihood"/>
               <rateStatistic idref="meanRate"/>
               <parameter idref="treeModel.rootHeight"/>
               <tmrcaStatistic idref="tmrca(Western
Gorilla)"/>
               <tmrcaStatistic idref="tmrca(Gorilla)"/>
```

```
                <tmrcaStatistic idref="tmrca(Hominid)"/>
                <tmrcaStatistic idref="tmrca(Hominoid)"/>
                <tmrcaStatistic idref="tmrca(Chimp)"/>
                <tmrcaStatistic idref="tmrca(Human-Chimp)"/>
                <tmrcaStatistic idref="tmrca(African
Apes)"/>
                <tmrcaStatistic idref="tmrca(Chimp-
Bonobo)"/>
                <tmrcaStatistic idref="tmrca(Pongo)"/>
                <tmrcaStatistic idref="tmrca(Bonobo)"/>
                <tmrcaStatistic idref="tmrca(Human)"/>
                <parameter idref="yule.birthRate"/>
                <parameter idref="siteModel1.mu"/>
                <parameter idref="siteModel2.mu"/>
                <parameter idref="hky1.kappa"/>
                <parameter idref="hky2.kappa"/>
                <parameter idref="siteModel1.alpha"/>
                <parameter idref="siteModel2.alpha"/>
                <parameter idref="ucld.mean"/>
                <parameter idref="ucld.stdev"/>
                <rateStatistic
idref="coefficientOfVariation"/>
                <rateCovarianceStatistic
idref="covariance"/>
                <treeLikelihood idref="treeLikelihood1"/>
                <treeLikelihood idref="treeLikelihood2"/>
                <speciationLikelihood idref="speciation"/>
          </log>
          <logTree id="treeFileLog" logEvery="500"
nexusFormat="true" fileName="12_heavy_strand_genes.trees"
sortTranslationTable="true">
                <treeModel idref="treeModel"/>
                <discretizedBranchRates
idref="branchRates"/>
                <posterior idref="posterior"/>
          </logTree>
     </mcmc>

     <report>
          <property name="timer">
                <object idref="mcmc"/>
          </property>
     </report>
```

```
</beast>
```

1.1.2 Whole mtDNA without D-loop file (With GAGP gorillas)

```
<?xml version="1.0" standalone="yes"?>

<!-- Generated by BEAUTi v1.7.5
-->
<!--       by Alexei J. Drummond and Andrew Rambaut
-->
<!--       Department of Computer Science, University of
Auckland and        -->
<!--       Institute of Evolutionary Biology, University of
Edinburgh        -->
<!--       http: //beast.bio.ed.ac.uk/
-->
<beast>

     <!-- The list of taxa analyse (can also include
dates/ages).                 -->
     <!-- ntax=23
-->
     <taxa id="taxa">
          <taxon id="easterngorillmkubwa"/>
          <taxon id="easterngorillakaisi"/>
          <taxon id="westerngorilladiehlinyango"/>
          <taxon id="westerngorillaoko"/>
          <taxon id="westerngorillachoomba"/>
          <taxon id="westerngorillatzambo"/>
          <taxon id="westerngorillasuzie"/>
          <taxon id="westerngorillakokamo"/>
          <taxon id="westerngorillasandra"/>
          <taxon id="westerngorillaanthal"/>
          <taxon id="chipua"/>
          <taxon id="bonobo30"/>
          <taxon id="bonobo75"/>
          <taxon id="p.t.verus"/>
          <taxon id="p.t.schweinfurthii"/>
          <taxon id="p.t.troglodytes"/>
          <taxon id="p.t.ellioti"/>
          <taxon id="human"/>
          <taxon id="neanderthal"/>
          <taxon id="Pongo_abelii"/>
```

```
        <taxon id="Pongo_pygmaeus"/>
        <taxon id="gibbon"/>
        <taxon id="macaque"/>
    </taxa>
    <taxa id="Eastern Gorilla">
        <taxon idref="easterngorillmkubwa"/>
        <taxon idref="easterngorillakaisi"/>
    </taxa>
    <taxa id="African Apes">
        <taxon idref="easterngorillmkubwa"/>
        <taxon idref="easterngorillakaisi"/>
        <taxon idref="westerngorilladiehlinyango"/>
        <taxon idref="westerngorillaoko"/>
        <taxon idref="westerngorillachoomba"/>
        <taxon idref="westerngorillatzambo"/>
        <taxon idref="westerngorillasuzie"/>
        <taxon idref="westerngorillakokamo"/>
        <taxon idref="westerngorillasandra"/>
        <taxon idref="westerngorillaanthal"/>
        <taxon idref="chipua"/>
        <taxon idref="bonobo30"/>
        <taxon idref="bonobo75"/>
        <taxon idref="p.t.verus"/>
        <taxon idref="p.t.schweinfurthii"/>
        <taxon idref="p.t.troglodytes"/>
        <taxon idref="p.t.ellioti"/>
        <taxon idref="human"/>
        <taxon idref="neanderthal"/>
    </taxa>
    <taxa id="Gorilla">
        <taxon idref="easterngorillmkubwa"/>
        <taxon idref="easterngorillakaisi"/>
        <taxon idref="westerngorilladiehlinyango"/>
        <taxon idref="westerngorillaoko"/>
        <taxon idref="westerngorillachoomba"/>
        <taxon idref="westerngorillatzambo"/>
        <taxon idref="westerngorillasuzie"/>
        <taxon idref="westerngorillakokamo"/>
        <taxon idref="westerngorillasandra"/>
        <taxon idref="westerngorillaanthal"/>
        <taxon idref="chipua"/>
    </taxa>
    <taxa id="Human-Chimp">
        <taxon idref="bonobo30"/>
```

```
        <taxon idref="bonobo75"/>
        <taxon idref="p.t.verus"/>
        <taxon idref="p.t.schweinfurthii"/>
        <taxon idref="p.t.troglodytes"/>
        <taxon idref="p.t.ellioti"/>
        <taxon idref="human"/>
        <taxon idref="neanderthal"/>
    </taxa>
    <taxa id="Chimp-Bonobo">
        <taxon idref="bonobo30"/>
        <taxon idref="bonobo75"/>
        <taxon idref="p.t.verus"/>
        <taxon idref="p.t.schweinfurthii"/>
        <taxon idref="p.t.troglodytes"/>
        <taxon idref="p.t.ellioti"/>
    </taxa>
    <taxa id="Western Gorilla">
        <taxon idref="westerngorilladiehlinyango"/>
        <taxon idref="westerngorillaoko"/>
        <taxon idref="westerngorillachoomba"/>
        <taxon idref="westerngorillatzambo"/>
        <taxon idref="westerngorillasuzie"/>
        <taxon idref="westerngorillakokamo"/>
        <taxon idref="westerngorillasandra"/>
        <taxon idref="westerngorillaanthal"/>
        <taxon idref="chipua"/>
    </taxa>
    <taxa id="Bonobo">
        <taxon idref="bonobo30"/>
        <taxon idref="bonobo75"/>
    </taxa>
    <taxa id="Human">
        <taxon idref="human"/>
        <taxon idref="neanderthal"/>
    </taxa>
    <taxa id="Chimp">
        <taxon idref="p.t.verus"/>
        <taxon idref="p.t.schweinfurthii"/>
        <taxon idref="p.t.troglodytes"/>
        <taxon idref="p.t.ellioti"/>
    </taxa>
    <taxa id="Pongo">
        <taxon idref="Pongo_abelii"/>
        <taxon idref="Pongo_pygmaeus"/>
```

```
</taxa>
<taxa id="Hominid">
     <taxon idref="p.t.ellioti"/>
     <taxon idref="p.t.schweinfurthii"/>
     <taxon idref="p.t.troglodytes"/>
     <taxon idref="p.t.verus"/>
     <taxon idref="Pongo_abelii"/>
     <taxon idref="Pongo_pygmaeus"/>
     <taxon idref="bonobo30"/>
     <taxon idref="bonobo75"/>
     <taxon idref="easterngorillmkubwa"/>
     <taxon idref="easterngorillakaisi"/>
     <taxon idref="human"/>
     <taxon idref="neanderthal"/>
     <taxon idref="westerngorilladiehlinyango"/>
     <taxon idref="westerngorillaoko"/>
     <taxon idref="westerngorillachoomba"/>
     <taxon idref="westerngorillatzambo"/>
     <taxon idref="westerngorillasuzie"/>
     <taxon idref="westerngorillakokamo"/>
     <taxon idref="westerngorillasandra"/>
     <taxon idref="westerngorillaanthal"/>
     <taxon idref="chipua"/>
</taxa>
<taxa id="Hominoid">
     <taxon idref="p.t.ellioti"/>
     <taxon idref="p.t.schweinfurthii"/>
     <taxon idref="p.t.troglodytes"/>
     <taxon idref="p.t.verus"/>
     <taxon idref="Pongo_abelii"/>
     <taxon idref="Pongo_pygmaeus"/>
     <taxon idref="bonobo30"/>
     <taxon idref="bonobo75"/>
     <taxon idref="easterngorillmkubwa"/>
     <taxon idref="easterngorillakaisi"/>
     <taxon idref="human"/>
     <taxon idref="neanderthal"/>
     <taxon idref="westerngorilladiehlinyango"/>
     <taxon idref="westerngorillaoko"/>
     <taxon idref="westerngorillachoomba"/>
     <taxon idref="westerngorillatzambo"/>
     <taxon idref="westerngorillasuzie"/>
     <taxon idref="westerngorillakokamo"/>
     <taxon idref="westerngorillasandra"/>
```

```
          <taxon idref="westerngorillaanthal"/>
          <taxon idref="chipua"/>
          <taxon idref="gibbon"/>
     </taxa>

     <!-- The sequence alignment (each sequence refers to a
taxon above).         -->
     <!-- ntax=23 nchar=15599
-->
     <alignment id="alignment" dataType="nucleotide">
          <sequence>
               <taxon idref="…"/>
</sequence>
     </alignment>
<!-- The unique patterns for codon positions 1 & 2
-->
     <mergePatterns id="patterns1+2">
          <!-- The unique patterns for codon position 1
-->
          <!-- npatterns=623
-->
          <patterns id="patterns1" from="1" every="3">
               <alignment idref="alignment"/>
          </patterns>
          <!-- The unique patterns for codon position 2
-->
          <!-- npatterns=301
-->
          <patterns id="patterns2" from="2" every="3">
               <alignment idref="alignment"/>
          </patterns>
     </mergePatterns>
     <!-- The unique patterns for codon position 3
-->
     <!-- npatterns=1513
-->
     <patterns id="patterns3" from="3" every="3">
          <alignment idref="alignment"/>
     </patterns>

     <!-- A prior on the distribution node heights defined
given                  -->
     <!-- a Yule speciation process (a pure birth process).
-->
```

```
     <yuleModel id="yule" units="substitutions">
          <birthRate>
               <parameter id="yule.birthRate" value="1.0"
lower="0.0" upper="Infinity"/>
          </birthRate>
     </yuleModel>
     <!-- This is a simple constant population size
coalescent model              -->
     <!-- that is used to generate an initial tree for the
chain.                 -->
     <constantSize id="initialDemo" units="substitutions">
          <populationSize>
               <parameter id="initialDemo.popSize"
value="100.0"/>
          </populationSize>
     </constantSize>

     <!-- Generate a random starting tree under the
coalescent process            -->
     <coalescentTree id="startingTree">
          <constrainedTaxa>
               <taxa idref="taxa"/>
               <tmrca monophyletic="false">
                    <taxa idref="Western Gorilla"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Gorilla"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Hominid"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Hominoid"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Bonobo"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Human-Chimp"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="African Apes"/>
               </tmrca>
               <tmrca monophyletic="false">
```

```
                    <taxa idref="Chimp-Bonobo"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Pongo"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Bonobo"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Human"/>
               </tmrca>
               <tmrca monophyletic="false">
                    <taxa idref="Eastern Gorilla"/>
               </tmrca>
          </constrainedTaxa>
          <constantSize idref="initialDemo"/>
     </coalescentTree>

     <treeModel id="treeModel">
          <coalescentTree idref="startingTree"/>
          <rootHeight>
               <parameter id="treeModel.rootHeight"/>
          </rootHeight>
          <nodeHeights internalNodes="true">
               <parameter
id="treeModel.internalNodeHeights"/>
          </nodeHeights>
          <nodeHeights internalNodes="true"
rootNode="true">
               <parameter
id="treeModel.allInternalNodeHeights"/>
          </nodeHeights>
     </treeModel>

     <speciationLikelihood id="speciation">
          <model>
               <yuleModel idref="yule"/>
          </model>
          <speciesTree>
               <treeModel idref="treeModel"/>
          </speciesTree>
     </speciationLikelihood>
```

```
     <!-- The uncorrelated relaxed clock (Drummond, Ho,
Phillips & Rambaut, 2006) -->
     <discretizedBranchRates id="branchRates">
          <treeModel idref="treeModel"/>
          <distribution>
               <logNormalDistributionModel
meanInRealSpace="true">
                    <mean>
                         <parameter id="ucld.mean"
value="0.038" lower="0.0" upper="100.0"/>
                    </mean>
                    <stdev>
                         <parameter id="ucld.stdev"
value="0.1" lower="0.0" upper="10.0"/>
                    </stdev>
               </logNormalDistributionModel>
          </distribution>
          <rateCategories>
               <parameter id="branchRates.categories"
dimension="26"/>
          </rateCategories>
     </discretizedBranchRates>

     <rateStatistic id="meanRate" name="meanRate"
mode="mean" internal="true" external="true">
          <treeModel idref="treeModel"/>
          <discretizedBranchRates idref="branchRates"/>
     </rateStatistic>

     <rateStatistic id="coefficientOfVariation"
name="coefficientOfVariation" mode="coefficientOfVariation"
internal="true" external="true">
          <treeModel idref="treeModel"/>
          <discretizedBranchRates idref="branchRates"/>
     </rateStatistic>

     <rateCovarianceStatistic id="covariance"
name="covariance">
          <treeModel idref="treeModel"/>
          <discretizedBranchRates idref="branchRates"/>
     </rateCovarianceStatistic>

     <!-- The HKY substitution model (Hasegawa, Kishino &
Yano, 1985)             -->
```

```
     <hkyModel id="hky1">
          <frequencies>
               <frequencyModel dataType="nucleotide">
                    <alignment idref="alignment"/>
                    <frequencies>
                         <parameter id="hky1.frequencies"
dimension="4"/>
                    </frequencies>
               </frequencyModel>
          </frequencies>
          <kappa>
               <parameter id="hky1.kappa" value="1.0"
lower="1.0E-8" upper="Infinity"/>
          </kappa>
     </hkyModel>
     <!-- The HKY substitution model (Hasegawa, Kishino &
Yano, 1985)              -->
     <hkyModel id="hky2">
          <frequencies>
               <frequencyModel dataType="nucleotide">
                    <alignment idref="alignment"/>
                    <frequencies>
                         <parameter id="hky2.frequencies"
dimension="4"/>
                    </frequencies>
               </frequencyModel>
          </frequencies>
          <kappa>
               <parameter id="hky2.kappa" value="1.0"
lower="1.0E-8" upper="Infinity"/>
          </kappa>
     </hkyModel>

     <!-- site model
-->
     <siteModel id="siteModel1">
          <substitutionModel>
               <hkyModel idref="hky1"/>
          </substitutionModel>
          <relativeRate>
               <parameter id="siteModel1.mu" value="1.0"
lower="0.0" upper="Infinity"/>
          </relativeRate>
          <gammaShape gammaCategories="4">
```

```
                <parameter id="siteModel1.alpha" value="0.5"
lower="0.0" upper="Infinity"/>
            </gammaShape>
        </siteModel>
        <!-- site model
-->
        <siteModel id="siteModel2">
            <substitutionModel>
                <hkyModel idref="hky2"/>
            </substitutionModel>
            <relativeRate>
                <parameter id="siteModel2.mu" value="1.0"
lower="0.0" upper="Infinity"/>
            </relativeRate>
            <gammaShape gammaCategories="4">
                <parameter id="siteModel2.alpha" value="0.5"
lower="0.0" upper="Infinity"/>
            </gammaShape>
        </siteModel>

        <compoundParameter id="allMus">
            <parameter idref="siteModel1.mu"/>
            <parameter idref="siteModel2.mu"/>
        </compoundParameter>

        <treeLikelihood id="treeLikelihood1">
            <patterns idref="patterns1+2"/>
            <treeModel idref="treeModel"/>
            <siteModel idref="siteModel1"/>
            <discretizedBranchRates idref="branchRates"/>
        </treeLikelihood>
        <treeLikelihood id="treeLikelihood2">
            <patterns idref="patterns3"/>
            <treeModel idref="treeModel"/>
            <siteModel idref="siteModel2"/>
            <discretizedBranchRates idref="branchRates"/>
        </treeLikelihood>
        <tmrcaStatistic id="tmrca(Eastern Gorilla)">
            <mrca>
                <taxa idref="Eastern Gorilla"/>
            </mrca>
            <treeModel idref="treeModel"/>
        </tmrcaStatistic>
        <tmrcaStatistic id="tmrca(Western Gorilla)">
```

```
     <mrca>
          <taxa idref="Western Gorilla"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Gorilla)">
     <mrca>
          <taxa idref="Gorilla"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Hominid)">
     <mrca>
          <taxa idref="Hominid"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Hominoid)">
     <mrca>
          <taxa idref="Hominoid"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Chimp)">
     <mrca>
          <taxa idref="Chimp"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Human-Chimp)">
     <mrca>
          <taxa idref="Human-Chimp"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(African Apes)">
     <mrca>
          <taxa idref="African Apes"/>
     </mrca>
     <treeModel idref="treeModel"/>
</tmrcaStatistic>
<tmrcaStatistic id="tmrca(Chimp-Bonobo)">
     <mrca>
          <taxa idref="Chimp-Bonobo"/>
```

```
            </mrca>
            <treeModel idref="treeModel"/>
        </tmrcaStatistic>
        <tmrcaStatistic id="tmrca(Pongo)">
            <mrca>
                <taxa idref="Pongo"/>
            </mrca>
            <treeModel idref="treeModel"/>
        </tmrcaStatistic>
        <tmrcaStatistic id="tmrca(Bonobo)">
            <mrca>
                <taxa idref="Bonobo"/>
            </mrca>
            <treeModel idref="treeModel"/>
        </tmrcaStatistic>
        <tmrcaStatistic id="tmrca(Human)">
            <mrca>
                <taxa idref="Human"/>
            </mrca>
            <treeModel idref="treeModel"/>
        </tmrcaStatistic>
        <operators id="operators">
            <scaleOperator scaleFactor="0.75" weight="1">
                <parameter idref="hky1.kappa"/>
            </scaleOperator>
            <scaleOperator scaleFactor="0.75" weight="1">
                <parameter idref="hky2.kappa"/>
            </scaleOperator>
            <scaleOperator scaleFactor="0.75" weight="1">
                <parameter idref="siteModel1.alpha"/>
            </scaleOperator>
            <scaleOperator scaleFactor="0.75" weight="1">
                <parameter idref="siteModel2.alpha"/>
            </scaleOperator>
            <deltaExchange delta="0.75" parameterWeights="2
1" weight="1">
                <parameter idref="allMus"/>
            </deltaExchange>
            <scaleOperator scaleFactor="0.75" weight="3">
                <parameter idref="ucld.mean"/>
            </scaleOperator>
            <scaleOperator scaleFactor="0.75" weight="3">
                <parameter idref="ucld.stdev"/>
            </scaleOperator>
```

```
          <upDownOperator scaleFactor="0.75" weight="3">
               <up>
                    <parameter idref="ucld.mean"/>
               </up>
               <down>
                    <parameter
idref="treeModel.allInternalNodeHeights"/>
               </down>
          </upDownOperator>
          <swapOperator size="1" weight="10"
autoOptimize="false">
               <parameter idref="branchRates.categories"/>
          </swapOperator>
          <randomWalkIntegerOperator windowSize="1.0"
weight="10">
               <parameter idref="branchRates.categories"/>
          </randomWalkIntegerOperator>
          <uniformIntegerOperator weight="10">
               <parameter idref="branchRates.categories"/>
          </uniformIntegerOperator>
          <scaleOperator scaleFactor="0.75" weight="3">
               <parameter idref="yule.birthRate"/>
          </scaleOperator>
          <scaleOperator scaleFactor="0.75" weight="3">
               <parameter idref="treeModel.rootHeight"/>
          </scaleOperator>
          <uniformOperator weight="30">
               <parameter
idref="treeModel.internalNodeHeights"/>
          </uniformOperator>
          <subtreeSlide size="0.1" gaussian="true"
weight="15">
               <treeModel idref="treeModel"/>
          </subtreeSlide>
          <narrowExchange weight="15">
               <treeModel idref="treeModel"/>
          </narrowExchange>
          <wideExchange weight="3">
               <treeModel idref="treeModel"/>
          </wideExchange>
          <wilsonBalding weight="3">
               <treeModel idref="treeModel"/>
          </wilsonBalding>
     </operators>
```

```
     <mcmc id="mcmc" chainLength="10000000"
autoOptimize="true">
          <posterior id="posterior">
               <prior id="prior">
                    <logNormalPrior mean="0.0" stdev="0.56"
offset="5.0" meanInRealSpace="false">
                         <statistic idref="tmrca(Human-
Chimp)"/>
                    </logNormalPrior>
                    <speciationLikelihood
idref="speciation"/>
               </prior>
               <likelihood id="likelihood">
                    <treeLikelihood
idref="treeLikelihood1"/>
                    <treeLikelihood
idref="treeLikelihood2"/>
               </likelihood>
          </posterior>
          <operators idref="operators"/>
          <log id="screenLog" logEvery="1000">
               <column label="Posterior" dp="4" width="12">
                    <posterior idref="posterior"/>
               </column>
               <column label="Prior" dp="4" width="12">
                    <prior idref="prior"/>
               </column>
               <column label="Likelihood" dp="4"
width="12">
                    <likelihood idref="likelihood"/>
               </column>
               <column label="Root Height" sf="6"
width="12">
                    <parameter
idref="treeModel.rootHeight"/>
               </column>
               <column label="Rate" sf="6" width="12">
                    <rateStatistic idref="meanRate"/>
               </column>
          </log>
          <log id="fileLog" logEvery="1000"
fileName="all_primates_no_D-loop.log">
               <posterior idref="posterior"/>
```

```
                <prior idref="prior"/>
                <likelihood idref="likelihood"/>
                <rateStatistic idref="meanRate"/>
                <parameter idref="treeModel.rootHeight"/>
                <tmrcaStatistic idref="tmrca(Western
Gorilla)"/>
                <tmrcaStatistic idref="tmrca(Eastern
Gorilla)"/>
                <tmrcaStatistic idref="tmrca(Gorilla)"/>
                <tmrcaStatistic idref="tmrca(Hominid)"/>
                <tmrcaStatistic idref="tmrca(Hominoid)"/>
                <tmrcaStatistic idref="tmrca(Chimp)"/>
                <tmrcaStatistic idref="tmrca(Human-Chimp)"/>
                <tmrcaStatistic idref="tmrca(African
Apes)"/>
                <tmrcaStatistic idref="tmrca(Chimp-
Bonobo)"/>
                <tmrcaStatistic idref="tmrca(Pongo)"/>
                <tmrcaStatistic idref="tmrca(Bonobo)"/>
                <tmrcaStatistic idref="tmrca(Human)"/>
                <parameter idref="yule.birthRate"/>
                <parameter idref="siteModel1.mu"/>
                <parameter idref="siteModel2.mu"/>
                <parameter idref="hky1.kappa"/>
                <parameter idref="hky2.kappa"/>
                <parameter idref="siteModel1.alpha"/>
                <parameter idref="siteModel2.alpha"/>
                <parameter idref="ucld.mean"/>
                <parameter idref="ucld.stdev"/>
                <rateStatistic
idref="coefficientOfVariation"/>
                <rateCovarianceStatistic
idref="covariance"/>
                <treeLikelihood idref="treeLikelihood1"/>
                <treeLikelihood idref="treeLikelihood2"/>
                <speciationLikelihood idref="speciation"/>
            </log>
            <logTree id="treeFileLog" logEvery="500"
nexusFormat="true" fileName="all_primates_no_D-loop.trees"
sortTranslationTable="true">
                <treeModel idref="treeModel"/>
                <discretizedBranchRates
idref="branchRates"/>
                <posterior idref="posterior"/>
```

```
        </logTree>
    </mcmc>

    <report>
        <property name="timer">
            <object idref="mcmc"/>
        </property>
    </report>

</beast>
```

1.2 Model Test

```
-------------------------------------------------------------
* *
* CORRECTED AKAIKE INFORMATION CRITERION (AICc) *
* *
-------------------------------------------------------------
Sample size: 10887.0
Model selected:
Model = TIM2+I+G
partition = 010232
-lnL = 46679.1559
K = 38
freqA = 0.3168
freqC = 0.3573
freqG = 0.0963
freqT = 0.2296
R(a) [AC] = 1.7204
R(b) [AG] = 44.6253
R(c) [AT] = 1.7204
R(d) [CG] = 1.0000
R(e) [CT] = 27.0073
R(f) [GT] = 1.0000
p-inv = 0.4860
gamma shape = 1.5150
Tree for the best AICc model = (((((((P.t.schweinfurthii:
0.00278285,P.t.troglodytes:
0.00413033): 0.00777519,(P.t.ellioti: 0.00585703,P.t.verus:
0.00597321):
0.00716770): 0.01295540,
(bonoboisolatePP75: 0.00449164,bonoboisolatePP30: 0.00447993):
0.01967346):
0.04302055,(neanderthal: 0.00536956,human: 0.00823402): 0.06945606):
0.03064082,(easterngorillmkubwa: 0.02296721,(westerngorillachipua:
0.00407187,
(westerngorilla011120: 0.00083084,westerngorilla001645: 0.00144142):
0.00348912): 0.01690757): 0.07811057): 0.08234293,(macaque:
0.63545366,gibbon: 0.20635170): 0.06556222): 0.15511283,Pongo_pygmaeus:
0.05112502,Pongo_abelii: 0.04814319);
* AICc MODEL SELECTION : Selection uncertainty
Model -lnL K AICc delta weight cumWeight
-----------------------------------------------------------------------
-
TIM2+I+G 46679.1559 38 93434.5851 0.0000 0.8375 0.8375
GTR+I+G 46678.8366 40 93437.9755 3.3904 0.1537 0.9912
TrN+I+G 46685.2711 37 93444.8014 10.2163 0.0051 0.9963
TIM3+I+G 46685.2610 38 93446.7952 12.2101 0.0019 0.9981
TIM1+I+G 46685.2677 38 93446.8086 12.2235 0.0019 1.0000
TIM2+G 46696.4676 37 93467.1944 32.6093 6.95e-008 1.0000
GTR+G 46696.0604 39 93470.4084 35.8233 1.39e-008 1.0000
TrN+G 46703.4712 36 93479.1879 44.6028 1.73e-010 1.0000
TIM3+G 46703.4692 37 93481.1977 46.6126 6.33e-011 1.0000
TIM1+G 46703.4712 37 93481.2016 46.6165 6.31e-011 1.0000
```

TPM2uf+I+G 46711.4320 37 93497.1232 62.5381 2.20e-014 1.0000
TVM+I+G 46711.0335 39 93500.3547 65.7696 4.38e-015 1.0000
TPM2uf+G 46723.9798 36 93520.2051 85.6200 2.14e-019 1.0000
TVM+G 46723.5950 38 93523.4632 88.8781 4.20e-020 1.0000
HKY+I+G 46726.2932 36 93524.8320 90.2469 2.12e-020 1.0000
TPM3uf+I+G 46726.1220 37 93526.5031 91.9180 9.19e-021 1.0000
TPM1uf+I+G 46726.2312 37 93526.7216 92.1365 8.24e-021 1.0000
HKY+G 46737.9734 35 93546.1790 111.5939 4.91e-025 1.0000
TPM3uf+G 46737.9226 36 93548.0908 113.5057 1.89e-025 1.0000
TPM1uf+G 46737.9270 36 93548.0994 113.5144 1.88e-025 1.0000
TIM2+I 46777.5799 37 93629.4190 194.8339 4.12e-043 1.0000
GTR+I 46777.0982 39 93632.4840 197.8989 8.91e-044 1.0000
TIM1+I 46794.3619 37 93662.9831 228.3980 2.12e-050 1.0000
TrN+I 46795.4669 36 93663.1792 228.5942 1.92e-050 1.0000
TIM3+I 46795.2324 37 93664.7240 230.1389 8.89e-051 1.0000
TPM2uf+I 46812.6483 36 93697.5421 262.9570 6.65e-058 1.0000
TVM+I 46812.3680 38 93701.0093 266.4242 1.17e-058 1.0000
TPM1uf+I 46836.7490 36 93745.7435 311.1584 2.27e-068 1.0000
HKY+I 46838.0070 35 93746.2462 311.6611 1.76e-068 1.0000
TPM3uf+I 46837.9876 36 93748.2208 313.6357 6.57e-069 1.0000
SYM+I+G 47774.1211 37 95622.5014 2187.9163 0.00e+000 1.0000
SYM+G 47781.6557 36 95635.5569 2200.9718 0.00e+000 1.0000
TIM2ef+I+G 47799.5439 35 95669.3200 2234.7349 0.00e+000 1.0000
TIM2ef+G 47807.2614 34 95682.7420 2248.1569 0.00e+000 1.0000
SYM+I 47838.3258 36 95748.8970 2314.3120 0.00e+000 1.0000
TIM2ef+I 47861.2382 34 95790.6958 2356.1107 0.00e+000 1.0000
TVMef+I+G 47892.2084 36 95856.6623 2422.0772 0.00e+000 1.0000
TVMef+G 47895.1932 35 95860.6186 2426.0335 0.00e+000 1.0000
TPM2+I+G 47916.4409 34 95901.1010 2466.5159 0.00e+000 1.0000
TPM2+G 47918.6856 33 95903.5781 2468.9930 0.00e+000 1.0000
TVMef+I 47960.1328 35 95990.4979 2555.9128 0.00e+000 1.0000
TPM2+I 47985.1966 33 96036.5999 2602.0148 0.00e+000 1.0000
TIM1ef+I+G 47994.9708 35 96060.1738 2625.5887 0.00e+000 1.0000
TIM1ef+G 48003.6780 34 96075.5753 2640.9902 0.00e+000 1.0000
TIM3ef+I+G 48017.4259 35 96105.0841 2670.4990 0.00e+000 1.0000
TIM3ef+G 48023.8699 34 96115.9591 2681.3740 0.00e+000 1.0000
TrNef+I+G 48035.0668 34 96138.3529 2703.7678 0.00e+000 1.0000
TrNef+G 48041.8734 33 96149.9536 2715.3686 0.00e+000 1.0000
TIM1ef+I 48059.8587 34 96187.9366 2753.3515 0.00e+000 1.0000
TIM3ef+I 48095.3677 34 96258.9548 2824.3697 0.00e+000 1.0000
TrNef+I 48108.8890 33 96283.9847 2849.3996 0.00e+000 1.0000
TPM1+I+G 48115.4669 34 96299.1532 2864.5681 0.00e+000 1.0000
TPM1+G 48119.6108 33 96305.4284 2870.8433 0.00e+000 1.0000
TPM3+I+G 48137.7309 34 96343.6811 2909.0960 0.00e+000 1.0000
TPM3+G 48140.6540 33 96347.5148 2912.9297 0.00e+000 1.0000
K80+I+G 48154.2472 33 96374.7011 2940.1160 0.00e+000 1.0000
K80+G 48156.7201 32 96377.6348 2943.0497 0.00e+000 1.0000
TPM1+I 48175.5037 33 96417.2141 2982.6290 0.00e+000 1.0000
TPM3+I 48209.3891 33 96484.9849 3050.3999 0.00e+000 1.0000
K80+I 48224.2308 32 96512.6562 3078.0711 0.00e+000 1.0000
TVM 48411.3642 37 96896.9877 3462.4026 0.00e+000 1.0000
GTR 48410.7754 38 96897.8241 3463.2390 0.00e+000 1.0000
TPM2uf 48469.3274 35 97008.8870 3574.3019 0.00e+000 1.0000
TIM2 48468.9757 36 97010.1969 3575.6118 0.00e+000 1.0000

```
TIM3 48534.0416 36 97140.3288 3705.7437 0.00e+000 1.0000
TPM3uf 48535.9138 35 97142.0599 3707.4748 0.00e+000 1.0000
TIM1 48550.8492 36 97173.9439 3739.3588 0.00e+000 1.0000
TPM1uf 48552.3314 35 97174.8951 3740.3100 0.00e+000 1.0000
TrN 48591.4110 35 97253.0543 3818.4692 0.00e+000 1.0000
HKY 48592.7780 34 97253.7752 3819.1901 0.00e+000 1.0000
SYM 48969.3667 35 98008.9655 4574.3804 0.00e+000 1.0000
TIM2ef 49039.8000 33 98145.8067 4711.2216 0.00e+000 1.0000
TVMef 49123.0205 34 98314.2603 4879.6752 0.00e+000 1.0000
TPM2 49194.4351 32 98453.0648 5018.4797 0.00e+000 1.0000
TIM3ef 49267.1116 33 98600.4301 5165.8450 0.00e+000 1.0000
TIM1ef 49271.3735 33 98608.9538 5174.3687 0.00e+000 1.0000
TrNef 49331.5505 32 98727.2956 5292.7106 0.00e+000 1.0000
TPM3 49418.4853 32 98901.1652 5466.5801 0.00e+000 1.0000
TPM1 49423.4215 32 98911.0376 5476.4525 0.00e+000 1.0000
K80 49483.8454 31 99029.8737 5595.2886 0.00e+000 1.0000
F81+I+G 51017.8270 35 102105.8862 8671.3011 0.00e+000 1.0000
F81+G 51030.9864 34 102130.1921 8695.6070 0.00e+000 1.0000
F81+I 51033.8040 34 102135.8273 8701.2422 0.00e+000 1.0000
JC+I+G 52042.8693 32 104149.9331 10715.3481 0.00e+000 1.0000
JC+I 52050.7687 31 104163.7202 10729.1351 0.00e+000 1.0000
JC+G 52064.0168 31 104190.2164 10755.6313 0.00e+000 1.0000
F81 52171.2148 33 104408.6364 10974.0513 0.00e+000 1.0000
JC 53062.8996 30 106185.9705 12751.3854 0.00e+000 1.0000
---------------------------------------------------------------------
-
-lnL: negative log likelihod
K: number of estimated parameters
AICc: Corrected Akaike Information Criterion
delta: AICc difference
weight: AICc weight
cumWeight: cumulative AICc weight
Model selection results also available at the "Model > Show model
table" menu
* AICc MODEL SELECTION : Confidence interval
There are 88 models in the 100% confidence interval: [ TIM2+I+G GTR+I+G
TrN+I
+G TIM3+I+G TIM1+I+G TIM2+G GTR+G TrN+G TIM3+G TIM1+G TPM2uf+I+G TVM
+I+G TPM2uf+G TVM+G HKY+I+G TPM3uf+I+G TPM1uf+I+G HKY+G TPM3uf+G
TPM1uf+G TIM2+I GTR+I TIM1+I TrN+I TIM3+I TPM2uf+I TVM+I TPM1uf+I HKY+I
TPM3uf+I SYM+I+G SYM+G TIM2ef+I+G TIM2ef+G SYM+I TIM2ef+I TVMef+I+G
TVMef+G TPM2+I+G TPM2+G TVMef+I TPM2+I TIM1ef+I+G TIM1ef+G TIM3ef+I+G
TIM3ef+G TrNef+I+G TrNef+G TIM1ef+I TIM3ef+I TrNef+I TPM1+I+G TPM1+G
TPM3+I+G TPM3+G K80+I+G K80+G TPM1+I TPM3+I K80+I TVM GTR TPM2uf
TIM2 TIM3 TPM3uf TIM1 TPM1uf TrN HKY SYM TIM2ef TVMef TPM2 TIM3ef
TIM1ef
TrNef TPM3 TPM1 K80 F81+I+G F81+G F81+I JC+I+G JC+I JC+G F81 JC ]
* AICc MODEL SELECTION : Parameter importance
Parameter Importance
----------------------
fA 1.0000
fC 1.0000
fG 1.0000
fT 1.0000
```

```
kappa 0.0000
titv 0.0000
rAC 0.9931
rAG 1.0000
rAT 0.9931
rCG 0.1575
rCT 1.0000
rGT 1.0000
pinv(I) 0.0000
alpha(G) 0.0000
pinv(IG) 1.0000
alpha(IG) 1.0000
----------------------
Values have been rounded.
(I): considers only +I models.
(G): considers only +G models.
(IG): considers only +I+G models.
* AICc MODEL SELECTION : Model averaged estimates
Model-averaged
Parameter estimates
-------------------------
fA 0.3169
fC 0.3573
fG 0.0963
fT 0.2296
kappa 20.5968
titv 9.3695
rAC 1.7643
rAG 45.8249
rAT 1.7819
rCG 1.3091
rCT 27.7375
rGT 1.0000
pinv(I) 0.5400
alpha(G) 0.2530
pinv(IG) 0.4858
alpha(IG) 1.5134
-------------------------
Numbers have been rounded.
(I): considers only +I models.
(G): considers only +G models.
(IG): considers only +I+G models.
* AICc MODEL SELECTION : Best Model's command line
phyml -i /var/folders/2N/2N4+4DpyG0WNn1FygOWtsU+++TI/-Tmp-/
jmodeltest2871601046970155456.phy -d nt -n 1 -b 0 --run_id TIM2+I+G -m
010232 -f m -v e -c 4 -a e -s NNI --no_memory_check -o tlr
---------------------------------------------------------------
* *
* BAYESIAN INFORMATION CRITERION (BIC) *
* *
---------------------------------------------------------------
Sample size: 10887.0
Model selected:
Model = TIM2+I+G
```

```
partition = 010232
-lnL = 46679.1559
K = 38
freqA = 0.3168
freqC = 0.3573
freqG = 0.0963
freqT = 0.2296
R(a) [AC] = 1.7204
R(b) [AG] = 44.6253
R(c) [AT] = 1.7204
R(d) [CG] = 1.0000
R(e) [CT] = 27.0073
R(f) [GT] = 1.0000
p-inv = 0.4860
gamma shape = 1.5150
Tree for the best BIC model = (((((((P.t.schweinfurthii:
0.00278285,P.t.troglodytes:
0.00413033): 0.00777519,(P.t.ellioti: 0.00585703,P.t.verus:
0.00597321):
0.00716770): 0.01295540,
(bonoboisolatePP75: 0.00449164,bonoboisolatePP30: 0.00447993):
0.01967346):
0.04302055,(neanderthal: 0.00536956,human: 0.00823402): 0.06945606):
0.03064082,(easterngorillmkubwa: 0.02296721,(westerngorillachipua:
0.00407187,
(westerngorilla011120: 0.00083084,westerngorilla001645: 0.00144142):
0.00348912): 0.01690757): 0.07811057): 0.08234293,(macaque:
0.63545366,gibbon: 0.20635170): 0.06556222): 0.15511283,Pongo_pygmaeus:
0.05112502,Pongo_abelii: 0.04814319);
* BIC MODEL SELECTION : Selection uncertainty
Model -lnL K BIC delta weight cumWeight
------------------------------------------------------------------------
-
TIM2+I+G 46679.1559 38 93711.5342 0.0000 0.8097 0.8097
TrN+I+G 46685.2711 37 93714.4693 2.9351 0.1866 0.9963
TIM3+I+G 46685.2610 38 93723.7443 12.2101 0.0018 0.9981
TIM1+I+G 46685.2677 38 93723.7577 12.2235 0.0018 0.9999
GTR+I+G 46678.8366 40 93729.4861 17.9519 0.0001 1.0000
TIM2+G 46696.4676 37 93736.8622 25.3280 2.56e-006 1.0000
TrN+G 46703.4712 36 93741.5741 30.0399 2.43e-007 1.0000
TIM3+G 46703.4692 37 93750.8655 39.3313 2.33e-009 1.0000
TIM1+G 46703.4712 37 93750.8695 39.3353 2.33e-009 1.0000
GTR+G 46696.0604 39 93754.6384 43.1042 3.53e-010 1.0000
TPM2uf+I+G 46711.4320 37 93766.7910 55.2568 8.12e-013 1.0000
TPM2uf+G 46723.9798 36 93782.5913 71.0571 3.01e-016 1.0000
TVM+I+G 46711.0335 39 93784.5847 73.0505 1.11e-016 1.0000
HKY+I+G 46726.2932 36 93787.2181 75.6840 2.98e-017 1.0000
TPM3uf+I+G 46726.1220 37 93796.1709 84.6367 3.39e-019 1.0000
TPM1uf+I+G 46726.2312 37 93796.3895 84.8553 3.04e-019 1.0000
TVM+G 46723.5950 38 93800.4123 88.8781 4.06e-020 1.0000
HKY+G 46737.9734 35 93801.2832 89.7490 2.63e-020 1.0000
TPM3uf+G 46737.9226 36 93810.4770 98.9428 2.65e-022 1.0000
TPM1uf+G 46737.9270 36 93810.4856 98.9514 2.64e-022 1.0000
TIM2+I 46777.5799 37 93899.0868 187.5526 1.52e-041 1.0000
```

GTR+I 46777.0982 39 93916.7140 205.1798 2.26e-045 1.0000
TrN+I 46795.4669 36 93925.5654 214.0312 2.70e-047 1.0000
TIM1+I 46794.3619 37 93932.6509 221.1167 7.82e-049 1.0000
TIM3+I 46795.2324 37 93934.3918 222.8576 3.28e-049 1.0000
TPM2uf+I 46812.6483 36 93959.9283 248.3941 9.34e-055 1.0000
TVM+I 46812.3680 38 93977.9584 266.4242 1.14e-058 1.0000
HKY+I 46838.0070 35 94001.3503 289.8161 9.45e-064 1.0000
TPM1uf+I 46836.7490 36 94008.1297 296.5955 3.19e-065 1.0000
TPM3uf+I 46837.9876 36 94010.6069 299.0728 9.24e-066 1.0000
SYM+I+G 47774.1211 37 95892.1693 2180.6351 0.00e+000 1.0000
SYM+G 47781.6557 36 95897.9430 2186.4088 0.00e+000 1.0000
TIM2ef+I+G 47799.5439 35 95924.4242 2212.8900 0.00e+000 1.0000
TIM2ef+G 47807.2614 34 95930.5638 2219.0296 0.00e+000 1.0000
SYM+I 47838.3258 36 96011.2832 2299.7490 0.00e+000 1.0000
TIM2ef+I 47861.2382 34 96038.5175 2326.9833 0.00e+000 1.0000
TVMef+G 47895.1932 35 96115.7227 2404.1885 0.00e+000 1.0000
TVMef+I+G 47892.2084 36 96119.0485 2407.5143 0.00e+000 1.0000
TPM2+G 47918.6856 33 96144.1170 2432.5828 0.00e+000 1.0000
TPM2+I+G 47916.4409 34 96148.9228 2437.3886 0.00e+000 1.0000
TVMef+I 47960.1328 35 96245.6020 2534.0678 0.00e+000 1.0000
TPM2+I 47985.1966 33 96277.1389 2565.6047 0.00e+000 1.0000
TIM1ef+I+G 47994.9708 35 96315.2779 2603.7437 0.00e+000 1.0000
TIM1ef+G 48003.6780 34 96323.3970 2611.8628 0.00e+000 1.0000
TIM3ef+I+G 48017.4259 35 96360.1882 2648.6540 0.00e+000 1.0000
TIM3ef+G 48023.8699 34 96363.7808 2652.2466 0.00e+000 1.0000
TrNef+I+G 48035.0668 34 96386.1746 2674.6404 0.00e+000 1.0000
TrNef+G 48041.8734 33 96390.4926 2678.9584 0.00e+000 1.0000
TIM1ef+I 48059.8587 34 96435.7583 2724.2241 0.00e+000 1.0000
TIM3ef+I 48095.3677 34 96506.7765 2795.2423 0.00e+000 1.0000
TrNef+I 48108.8890 33 96524.5237 2812.9895 0.00e+000 1.0000
TPM1+G 48119.6108 33 96545.9674 2834.4332 0.00e+000 1.0000
TPM1+I+G 48115.4669 34 96546.9749 2835.4407 0.00e+000 1.0000
TPM3+G 48140.6540 33 96588.0538 2876.5196 0.00e+000 1.0000
TPM3+I+G 48137.7309 34 96591.5028 2879.9686 0.00e+000 1.0000
K80+G 48156.7201 32 96610.8906 2899.3564 0.00e+000 1.0000
K80+I+G 48154.2472 33 96615.2400 2903.7058 0.00e+000 1.0000
TPM1+I 48175.5037 33 96657.7531 2946.2189 0.00e+000 1.0000
TPM3+I 48209.3891 33 96725.5239 3013.9897 0.00e+000 1.0000
K80+I 48224.2308 32 96745.9120 3034.3778 0.00e+000 1.0000
TVM 48411.3642 37 97166.6555 3455.1213 0.00e+000 1.0000
GTR 48410.7754 38 97174.7732 3463.2390 0.00e+000 1.0000
TPM2uf 48469.3274 35 97263.9911 3552.4569 0.00e+000 1.0000
TIM2 48468.9757 36 97272.5831 3561.0489 0.00e+000 1.0000
TPM3uf 48535.9138 35 97397.1640 3685.6298 0.00e+000 1.0000
TIM3 48534.0416 36 97402.7150 3691.1808 0.00e+000 1.0000
TPM1uf 48552.3314 35 97429.9993 3718.4651 0.00e+000 1.0000
TIM1 48550.8492 36 97436.3300 3724.7959 0.00e+000 1.0000
HKY 48592.7780 34 97501.5969 3790.0627 0.00e+000 1.0000
TrN 48591.4110 35 97508.1584 3796.6242 0.00e+000 1.0000
SYM 48969.3667 35 98264.0697 4552.5355 0.00e+000 1.0000
TIM2ef 49039.8000 33 98386.3456 4674.8114 0.00e+000 1.0000
TVMef 49123.0205 34 98562.0821 4850.5479 0.00e+000 1.0000
TPM2 49194.4351 32 98686.3206 4974.7864 0.00e+000 1.0000
TIM3ef 49267.1116 33 98840.9690 5129.4348 0.00e+000 1.0000

```
TIM1ef 49271.3735 33 98849.4928 5137.9586 0.00e+000 1.0000
TrNef 49331.5505 32 98960.5515 5249.0173 0.00e+000 1.0000
TPM3 49418.4853 32 99134.4210 5422.8868 0.00e+000 1.0000
TPM1 49423.4215 32 99144.2934 5432.7592 0.00e+000 1.0000
K80 49483.8454 31 99255.8459 5544.3117 0.00e+000 1.0000
F81+I+G 51017.8270 35 102360.9903 8649.4561 0.00e+000 1.0000
F81+G 51030.9864 34 102378.0139 8666.4797 0.00e+000 1.0000
F81+I 51033.8040 34 102383.6490 8672.1148 0.00e+000 1.0000
JC+I+G 52042.8693 32 104383.1890 10671.6548 0.00e+000 1.0000
JC+I 52050.7687 31 104389.6925 10678.1583 0.00e+000 1.0000
JC+G 52064.0168 31 104416.1886 10704.6544 0.00e+000 1.0000
F81 52171.2148 33 104649.1754 10937.6412 0.00e+000 1.0000
JC 53062.8996 30 106404.6589 12693.1247 0.00e+000 1.0000
-------------------------------------------------------------------------
-
-lnL: negative log likelihod
K: number of estimated parameters
BIC: Bayesian Information Criterion
delta: BIC difference
weight: BIC weight
cumWeight: cumulative BIC weight
Model selection results also available at the "Model > Show model
table" menu
* BIC MODEL SELECTION : Confidence interval
There are 88 models in the 100% confidence interval: [ TIM2+I+G TrN+I+G
TIM3+I
+G TIM1+I+G GTR+I+G TIM2+G TrN+G TIM3+G TIM1+G GTR+G TPM2uf+I+G
TPM2uf+G TVM+I+G HKY+I+G TPM3uf+I+G TPM1uf+I+G TVM+G HKY+G TPM3uf
+G TPM1uf+G TIM2+I GTR+I TrN+I TIM1+I TIM3+I TPM2uf+I TVM+I HKY+I
TPM1uf
+I TPM3uf+I SYM+I+G SYM+G TIM2ef+I+G TIM2ef+G SYM+I TIM2ef+I TVMef+G
TVMef+I+G TPM2+G TPM2+I+G TVMef+I TPM2+I TIM1ef+I+G TIM1ef+G TIM3ef+I
+G TIM3ef+G TrNef+I+G TrNef+G TIM1ef+I TIM3ef+I TrNef+I TPM1+G TPM1+I+G
TPM3+G TPM3+I+G K80+G K80+I+G TPM1+I TPM3+I K80+I TVM GTR TPM2uf
TIM2 TPM3uf TIM3 TPM1uf TIM1 HKY TrN SYM TIM2ef TVMef TPM2 TIM3ef
TIM1ef
TrNef TPM3 TPM1 K80 F81+I+G F81+G F81+I JC+I+G JC+I JC+G F81 JC ]
* BIC MODEL SELECTION : Parameter importance
Parameter Importance
----------------------
fA 1.0000
fC 1.0000
fG 1.0000
fT 1.0000
kappa 0.0000
titv 0.0000
rAC 0.8116
rAG 1.0000
rAT 0.8116
rCG 0.0037
rCT 1.0000
rGT 1.0000
pinv(I) 0.0000
alpha(G) 0.0000
```

```
pinv(IG) 1.0000
alpha(IG) 1.0000
----------------------
Values have been rounded.
(I): considers only +I models.
(G): considers only +G models.
(IG): considers only +I+G models.
* BIC MODEL SELECTION : Model averaged estimates
Model-averaged
Parameter estimates
-------------------------
fA 0.3172
fC 0.3574
fG 0.0958
fT 0.2296
kappa 20.5970
titv 9.3696
rAC 1.7188
rAG 41.9140
rAT 1.7188
rCG 0.9969
rCT 25.1960
rGT 1.0000
pinv(I) 0.5400
alpha(G) 0.2525
pinv(IG) 0.4860
alpha(IG) 1.5070
-------------------------
Numbers have been rounded.
(I): considers only +I models.
(G): considers only +G models.
(IG): considers only +I+G models.
```

1.4 R Codes and Results for K_a-K_s Test

```
> library(seqinr)

> atpase6 = read.alignment(file = "prank-atpase6.fas.txt", format = "fasta")
> kaks (atpase6, verbose = TRUE)
$ka
                          bonobo      chimp      human westerngorilla011120
chipua
chimp                 0.02109728
human                 0.03301850 0.03432890
westerngorilla011120  0.05591217 0.05726049 0.03949691
chipua                0.06312457 0.05403634 0.04141727           0.00654429
easterngorillmakabuwa 0.05920123 0.06056555 0.04421372           0.02741823
0.02930448

$ks
                          bonobo      chimp      human westerngorilla011120
chipua
chimp                 0.06664418
human                 0.19315065 0.23126684
westerngorilla011120  0.39178300 0.36996234 0.35468859
chipua                0.39062559 0.36868184 0.33690464           0.01593767
easterngorillmakabuwa 0.41688639 0.42852901 0.30492608           0.09150999
0.09113255

> atpase8 = read.alignment(file = "prank-atpase8.fas.txt", format = "fasta")
> kaks (atpase8, verbose = TRUE)
$ka
          bonobo chimp human westerngorilla001645 chipua westerngorilla011120
chimp 0.022645380
human 0.028507273 0.036437088
westerngorilla001645 0.088331824 0.079976312 0.056641611
chipua 0.088331823 0.079976312 0.056641611 0.000000000
westerngorilla011120 0.097065670 0.088597216 0.064872249 0.007462825
0.007462825
easterngorillmakabuwa 0.097111149 0.088618712 0.064905710 0.030543846
0.030543846 0.022731187

$ks
       bonobo chimp human westerngorilla001645 chipua westerngorilla011120
chimp 0.07529756
human 0.23015564 0.24874243
westerngorilla001645 0.20559424 0.22200310 0.29819284
chipua 0.24791344 0.26542923 0.29819285 0.02926681
westerngorilla011120 0.20559424 0.22200310 0.29819284 0.00000000 0.02926681
easterngorillmakabuwa 0.17843581 0.19703200 0.27227572 0.04352543 0.07483508
0.04352543

> COI = read.alignment(file = "prank-COI.fas.txt", format = "fasta")
> kaks (COI, verbose = TRUE)
$ka
       bonobo chimp human westerngorilla001645 westerngorilla011120 chipua
chimp 0.008645159
human 0.011707437 0.012494412
westerngorilla001645 0.016428044 0.016970494 0.018217622
westerngorilla011120 0.014370875 0.014919189 0.016155389 0.002006021
```

```
chipua 0.013448351 0.013995494 0.015231114 0.002902743 0.000895873
easterngorillmakabuwa 0.019404465 0.017882791 0.017062862 0.006938151
0.004916061 0.005815523

$ks
       bonobo chimp human westerngorilla001645 westerngorilla011120 chipua
chimp 0.11006319
human 0.33955034 0.30635836
westerngorilla001645 0.34893358 0.37795925 0.42194936
westerngorilla011120 0.34883969 0.37793416 0.42192071 0.00000000
chipua 0.35413952 0.37602941 0.41831638 0.01499078 0.01498376
easterngorillmakabuwa 0.36356385 0.36747113 0.42240657 0.07729085 0.07727766
0.07701687

> COII = read.alignment(file = "prank-COII.fas.txt", format = "fasta")
> kaks (COII, verbose = TRUE)
$ka
       bonobo chimp human westerngorilla001645 westerngorilla011120 chipua
chimp 0.000000000
human 0.015934307 0.015951732
westerngorilla001645 0.015125486 0.015138879 0.022481533
westerngorilla011120 0.013337019 0.013350348 0.020675933 0.001755557
chipua 0.020100116 0.020120257 0.022731966 0.008398048 0.006628495
easterngorillmakabuwa 0.015396685 0.015411282 0.022733584 0.013056989
0.011275077 0.009216851

$ks
       bonobo chimp human westerngorilla001645 westerngorilla011120 chipua
chimp 0.083096779
human 0.306463372 0.329313768
westerngorilla001645 0.364522815 0.431475559 0.479942075
westerngorilla011120 0.356769917 0.422920081 0.470310215 0.004084678
chipua 0.380146678 0.447719138 0.481921664 0.016822750 0.012498921
easterngorillmakabuwa 0.423237995 0.454121303 0.474879522 0.078827278
0.083594204 0.078842463

> COIII = read.alignment(file = "prank-COIII.fas.txt", format = "fasta")
> kaks (COIII, verbose = TRUE)
$ka
                          bonobo       chimp       human westerngorilla011120
chipua
chimp                 0.011473431
human                 0.021522300 0.021080547
westerngorilla011120  0.023172147 0.026826497 0.027512416
chipua                0.021135424 0.024779718 0.025466800          0.001964639
easterngorillmakabuwa 0.021367980 0.025014967 0.029812587          0.009676727
0.011677337

$ks
                         bonobo      chimp      human westerngorilla011120
chipua
chimp                 0.09696149
human                 0.32175556 0.33142771
westerngorilla011120  0.39157439 0.44261274 0.39372906
chipua                0.39782569 0.43593204 0.39369835           0.02706461
easterngorillmakabuwa 0.35941472 0.39942165 0.36268148           0.06416918
0.06872249


> cytb = read.alignment(file = "PRANK-cytb.fas.txt", format = "fasta")
```

```
> kaks (cytb, verbose = TRUE)
$ka
       bonobo chimp human westerngorilla001645 chiupa westerngorilla011120
chimp 0.020749865
human 0.035013831 0.030695821
westerngorilla001645 0.041299487 0.045699013 0.040571673
chiupa 0.036923092 0.041283974 0.036233251 0.004065063
westerngorilla011120 0.036923092 0.041283974 0.036233251 0.004065063
0.008163447
easterngorillmakabuwa 0.036293832 0.040618198 0.035648743 0.011822245
0.007715445 0.015963055

$ks
       bonobo chimp human westerngorilla001645 chiupa westerngorilla011120
chimp 0.10064792
human 0.30272151 0.32164775
westerngorilla001645 0.36468950 0.39937949 0.35693593
chiupa 0.37658997 0.39985728 0.37746644 0.02287873
westerngorilla011120 0.34318175 0.36247364 0.34887742 0.02287873 0.01576513
easterngorillmakabuwa 0.38111399 0.40161549 0.38759772 0.14554721 0.13795317
0.11961111

> NADH1 = read.alignment(file = "prank-NADH1.fas.txt", format = "fasta")
> kaks (NADH1, verbose = TRUE)
$ka
       bonobo chimp human westerngorilla001645 westerngorilla011120 chipua
chimp 0.011513097
human 0.037199002 0.032157901
westerngorilla001645 0.046225497 0.041083898 0.032342237
westerngorilla011120 0.044422747 0.039297663 0.030588568 0.001658376
chipua 0.044422747 0.039297663 0.030588568 0.001658376 0.000000000
easterngorillmakabuwa 0.047956567 0.042808655 0.033653267 0.014717241
0.013020018 0.013020018

$ks
       bonobo chimp human westerngorilla001645 westerngorilla011120 chipua
chimp 0.126646734
human 0.271687182 0.263155637
westerngorilla001645 0.368932590 0.352400499 0.311442465
westerngorilla011120 0.375121992 0.369430032 0.326502450 0.008652176
chipua 0.368907114 0.386573074 0.331407096 0.017444209 0.020425539
easterngorillmakabuwa 0.411461462 0.425439391 0.357092564 0.089720325
0.092972064 0.090189565

> NADH2 = read.alignment(file = "prank-NADH2.fas.txt", format = "fasta")
> kaks (NADH2, verbose = TRUE)
$ka
       bonobo chimp human westerngorilla001645 westerngorilla011120 chipua
chimp 0.016317884
human 0.028387465 0.024115023
westerngorilla001645 0.045677883 0.044521462 0.043425706
westerngorilla011120 0.044071365 0.042911062 0.041816336 0.001490314
chipua 0.042464662 0.041301404 0.040208023 0.002985083 0.001490314
easterngorillmakabuwa 0.042929473 0.038565908 0.047215743 0.017822204
0.016287283 0.014756334

$ks
       bonobo chimp human westerngorilla001645 westerngorilla011120 chipua
chimp 0.107884398
human 0.274059871 0.306883294
```

```
westerngorilla001645 0.354997756 0.352076915 0.431036826
westerngorilla011120 0.350407979 0.347640211 0.425064146 0.002701703
chipua 0.360777900 0.358088095 0.436579562 0.002703877 0.005405614
easterngorillmakabuwa 0.370705345 0.367190634 0.432669156 0.074630198
0.077724244 0.077811574

> NADH3 = read.alignment(file = "prank-NADH3.fas.txt", format = "fasta")
> kaks (NADH3, verbose = TRUE)
$ka
       human westerngorilla001645 westerngorilla011120 chipua
easterngorillmakabuwa
bonobo
westerngorilla001645 0.036420411
westerngorilla011120 0.041509513 0.004761941
chipua 0.036420410 0.009569670 0.004761941
easterngorillmakabuwa 0.056360339 0.038712064 0.033622962 0.028585136
bonobo 0.037031898 0.047393849 0.042239164 0.037137079 0.058941391
chimp 0.030429198 0.048073675 0.053299426 0.048073675 0.073602830 0.013074734

$ks
       human westerngorilla001645 westerngorilla011120 chipua
easterngorillmakabuwa
bonobo
westerngorilla001645 0.36862614
westerngorilla011120 0.36862614 0.00000000
chipua 0.39118659 0.02368062 0.02368062
easterngorillmakabuwa 0.36358120 0.08166063 0.08166063 0.09127782
bonobo 0.33313465 0.28073129 0.28073129 0.29452357 0.20924564
chimp 0.37478646 0.29953820 0.29953820 0.31502263 0.24108085 0.09122324

> ND4 = read.alignment(file = "prank-ND4.fas.txt", format = "fasta")
> kaks (ND4, verbose = TRUE)
$ka
                          bonobo       chimp       human westerngorilla011120
chipua
chimp                 0.010978149
human                 0.027511784 0.035351284
westerngorilla011120  0.037921178 0.046380133 0.034300238
chipua                0.039842626 0.048309509 0.036227905          0.006500819
easterngorillmakabuwa 0.045289931 0.053838371 0.037052222          0.017639175
0.015404184

$ks
                         bonobo      chimp      human westerngorilla011120
chipua
chimp                 0.09693760
human                 0.24631290 0.27239892
westerngorilla011120  0.32752641 0.37138290 0.35850595
chipua                0.32194944 0.37267813 0.34548788           0.01219652
easterngorillmakabuwa 0.32430840 0.39331762 0.34658366           0.06812317
0.07285115

> kaks (NADH4L, verbose = TRUE)
$ka
       bonobo chimp westerngorilla001645 westerngorilla011120 chipua
easterngorillmakabuwa
chimp 0.005714348
westerngorilla001645 0.015831469 0.011399082
westerngorilla011120 0.010077827 0.005692206 0.005681879
chipua 0.010077827 0.005692206 0.005681879 0.000000000
```

```
easterngorillmakabuwa 0.015179748 0.010806606 0.010693262 0.005000056
0.005000056
human 0.011461820 0.005698067 0.015819737 0.010082788 0.010082788 0.015189358

$ks
       bonobo chimp westerngorilla001645 westerngorilla011120 chipua
easterngorillmakabuwa
chimp 0.109791327
westerngorilla001645 0.331789608 0.279045640
westerngorilla011120 0.331091987 0.278418760 0.000000000
chipua 0.313619627 0.262416550 0.009175342 0.009173820
easterngorillmakabuwa 0.316499502 0.264103870 0.090197921 0.090144408
0.078880613
human 0.378353727 0.268724859 0.308264422 0.307245896 0.288728312 0.288973082

> nd5 = read.alignment(file = "prank-ND5.fas.txt", format = "fasta")
> kaks (nd5, verbose = TRUE)
$ka
           bonobo       chimp       human westerngorilla001645
westerngorilla011120      chipua
chimp                  0.030252277
human                  0.044514855 0.046962992
westerngorilla001645   0.058511363 0.059200342 0.065357150
westerngorilla011120   0.058220383 0.058908692 0.065057578          0.001533572
chipua                 0.057263465 0.056037420 0.064088260        0.009380739
0.007826247
easterngorillmakabuwa 0.057255145 0.062273990 0.063627186        0.020509588
0.018924861 0.021617561

$ks
              bonobo       chimp       human westerngorilla001645
westerngorilla011120      chipua
chimp                  0.119817866
human                  0.270773907 0.294950236
westerngorilla001645   0.370278367 0.351735775 0.439949434
westerngorilla011120   0.368105517 0.349358565 0.434361923          0.001534793
chipua                 0.357002197 0.341511432 0.421161342         0.026711662
0.028328781
easterngorillmakabuwa 0.359411955 0.346282533 0.410705692         0.130432665
0.132447418
0.124747035


> ND6 = read.alignment(file = "prank-ND6.fas.txt", format = "fasta")
> kaks (ND6, verbose = TRUE)
$ka
                        bonobo      chimp      human westerngorilla011120
chipua
chimp                  0.01270588
human                  0.03818915 0.03814574
westerngorilla011120   0.02875145 0.03491732 0.02329407
chipua                 0.02875725 0.03492312 0.02320297           0.00000000
easterngorillmakabuwa 0.02875145 0.03491732 0.02329407           0.00000000
0.00000000

$ks
                          bonobo      chimp      human westerngorilla011120
chipua
chimp                  0.12019329
human                  0.26865379 0.26640541
```

```
westerngorilla011120  0.33882884 0.39800981 0.30393235
chipua                0.36001087 0.42505452 0.32740976           0.02279626
easterngorillmakabuwa 0.38190200 0.48387969 0.49486620           0.16397797
0.16287392
```

1.5 R Codes and Results for Chimp-Bonobo and EL-WL distances

```
> a <- read.dna("Chimp-Bonobo.phy", format = "interleaved")
> a
2 DNA sequences in binary format stored in a matrix.
All sequences of same length: 10887
Labels: P.t.troglo bonoboisol
Base composition:
a c g t
0.296 0.325 0.117 0.262
> dist.dna(a, model = "TN93", variance = TRUE, as.matrix =
TRUE)
P.t.troglo bonoboisol
P.t.troglo 0.00000000 0.04283056
bonoboisol 0.04283056 0.00000000
attr(,"variance")
[1] 4.456362e-06

> b <- read.dna("EL-WL.phy", format = "interleaved")
> b
2 DNA sequences in binary format stored in a matrix.
All sequences of same length: 10887
Labels: westerngor easterngor
Base composition:
a c g t
0.294 0.324 0.119 0.263

> dist.dna(b, model = "TN93", variance = TRUE, as.matrix =
TRUE)
westerngor easterngor
westerngor 0.00000000 0.04016323
easterngor 0.04016323 0.00000000
attr(,"variance")
[1] 4.162352e-06

> t = function(K1,K2,V1,V2){(K1-K2)/(V1+V2)^0.5}
> t(0.04283056,0.04016323,4.456362e-06,4.162352e-06)
[1] 0.908564
> qt(0.05/2,Inf)
[1] -1.959964

> library (ape)
> a <- read.dna("Chimp-Bonobo.phy", format = "interleaved")
> a
```

```
2 DNA sequences in binary format stored in a matrix.
All sequences of same length: 10887
Labels: P.t.troglo bonoboisol
Base composition:
a c g t
0.296 0.325 0.117 0.262

> dist.dna(a, model = "F81", variance = TRUE, as.matrix =
TRUE)
P.t.troglo bonoboisol
P.t.troglo 0.00000000 0.04219243
bonoboisol 0.04219243 0.00000000
attr(,"variance")
[1] 4.067689e-06

> b <- read.dna("EL-WL.phy", format = "interleaved")
> dist.dna(b, model = "F81", variance = TRUE, as.matrix =
TRUE)
westerngor easterngor
westerngor 0.00000000 0.03960885
easterngor 0.03960885 0.00000000
attr(,"variance")
[1] 3.803036e-06

> t = function(K1,K2,V1,V2){(K1-K2)/(V1+V2)^0.5}
> t (0.04219243,0.03960885,4.067689e-06,3.803036e-06)
[1] 0.9209044
```

Appendix 2: Data from Chapter 3

2.1 CRTAC1 coding region sequence (CDS)

2.1.1 cDNA Alignment

```
chimp           ATGGCTCCGAGCGCTGACCCCGGCATGTCCAGGATGTTACTGTTCCTGCTGCTGCTCTGG 60
gorilla         ATGGCTCCGAGCGCTGACCCCGGCATGTCCAGGATGTTACTGTTCCTGCTGCTGCTCTGG 60
human           ATGGCTCCGAGCGCTGACCCCGGCATGTCCAGGATGTTACTGTTCCTGCTGCTGCTCTGG 60
orang           ATGGCTCCGAGCGCTGACCCCGGCATGTCCAGGATGTTACTGTTCCTGCTGCTGCTGTGG 60
                ************************************************************ ***

chimp           TTTCTGCCCATCACTGAGGGGTCCCAGCGGGCTGAACCCATGTTCACTGCAGTCACCAAC 120
gorilla         TTTCTGCCCATCACTGAGGGGTCCCAGCGGGCTGAACCCATGTTCACTGCAGTCACCAAC 120
human           TTTCTGCCCATCACTGAGGGGTCCCAGCGGGCTGAACCCATGTTCACTGCAGTCACCAAC 120
orang           TTTCTGCCCATCACTGAGGGGTCCCAGCGGGCTGAACCCATGTTCACTGCAGTCACCAAC 120
                ************************************************************

chimp           TCAGTTCTGCCTCCTGACTATGACAGTAATCCCACCCAGCTCAACTATGGTGTGGCAGTT 180
gorilla         TCAGTTCTGCCTCCTGACTATGACAGTAATCCCACCCAGCTCAACTATGGTGTGGCAGTT 180
human           TCAGTTCTGCCTCCTGACTATGACAGTAATCCCACCCAGCTCAACTATGGTGTGGCAGTT 180
orang           TCAGTTCTGCCCCCTGACTATGACAGTAATCCCACCCAGCTCAACTATGGTGTGGCAGTT 180
                *********** ************************************************

chimp           ACTGACGTGGACCATGATGGGGACTTTGAGATCGTCGTGGCGGGGTACAATGGACCCAAC 240
gorilla         ACTGACGTGGACCATGATGGGGACTTTGAGATCGTCGTGGCGGGGTACAATGGACCCAAC 240
human           ACTGACGTGGACCATGATGGGGACTTTGAGATCGTCGTGGCGGGGTACAATGGACCCAAC 240
orang           ACTGATGTGGACCATGATGGGGACTTTGAGATCGTCGTGGCGGGGTACAATGGCCCCAAC 240
                ***** ********************************************** ******

chimp           CTGGTTCTGAAGTATGACCGGGCCCAGAAGCGGCTGGTGAACATCGCGGTCGATGAGCGC 300
gorilla         CTGGTTCTGAAGTATGACCGGGCCCAGAAGCGGCTGGTGAACATCGCGGTCGATGAGCGC 300
human           CTGGTTCTGAAGTATGACCGGGCCCAGAAGCGGCTGGTGAACATCGCGGTCGATGAGCGC 300
orang           CTGGTTCTGAAGTATGACCGGGCCCAGAAGCGGCTGGTGAACATCGCGGTCGATGAGCGC 300
                ************************************************************

chimp           AGCTCACCCTACTACGCGCTGCGCGACCGGCAGGGGAACGCCATCGGGGTCACAGCCTGC 360
gorilla         AGCTCACCCTACTATGCGCTGCGGGACCGGCAGGGGAACGCCATCGGGGTCACAGCCTGC 360
human           AGCTCACCCTACTACGCGCTGCGGGACCGGCAGGGGAACGCCATTGGGGTCACAGCCTGC 360
orang           AGCTCACCCTACTACGCACTGCGGGACCGGCAGGGGAACGCCATCGGGGTCACAGCCTGC 360
                ************** ** ***** ******************** ***************

chimp           GACATCGACGGGGATGGCCGGGAGGAGATCTACTTCCTCAACACCAATAATGCCTTCTCG 420
gorilla         GACATCGACGGGGATGGCCGGGAGGAGATCTACTTCCTCAACACCAATAATGCCTTCTCG 420
human           GACATCGACGGGGACGGCCGGGAGGAGATCTACTTCCTCAACACCAATAATGCCTTCTCG 420
orang           GACATCGATGGAGATGGCCGGGAGGAGATCTACTTCCTCAACACCAATAACGCCTTCTCG 420
                ******** ** ** ********************************** *********

chimp           GGGGTGGCCACGTACACCGACAAGTTGTTCAAGTTCCGCAATAACCGGTGGGAAGACATC 480
```

```
gorilla         GGGGTGGCCACGTACACCGACAAGTTGTTCAAGTTCCGCAATAACCGGTGGGAAGACATC 480
human           GGGGTGGCCACGTACACCGACAAGTTGTTCAAGTTCCGCAATAACCGGTGGGAAGACATC 480
orang           GGGGTGGCCACGTACACCGACAAGTTGTTCAAGTTCCGCAATAACCGGTGGGAAGACATC 480
                ************************************************************

chimp           CTGAGCGATGAGGTCAACGTGGCCCGTGGTGTGGCCAGCCTCTTTGCCGGACGCTCTGTG 540
gorilla         CTGAGCGATGAGGTCAACGTGGCCCGTGGTGTGGCCAGCCTCTTTGCCGGACGCTCTGTG 540
human           CTGAGCGATGAGGTCAACGTGGCCCGTGGTGTGGCCAGCCTCTTTGCCGGACGCTCTGTG 540
orang           CTGAGCGATGAGGTCAACGTGGCCCGTGGTGTGGCCAGCCTCTTTGCCGGACGCTCTGTG 540
                ************************************************************

chimp           GCCTGTGTGGACAGAAAGGGCTCTGGACGCTACTCTATCTACATTGCCAATTATGCCTAC 600
gorilla         GCCTGTGTGGACAGAAAGGGCTCTGGACGCTACTCTATCTACATTGCCAATTACGCCTAC 600
human           GCCTGTGTGGACAGAAAGGGCTCTGGACGCTACTCTATCTACATTGCCAATTACGCCTAC 600
orang           GCCTGTGTGGACAGAAAGGGCTCTGGACGCTACTCTATCTACATTGCCAATTACGCCTAC 600
                ************************************************** ******


chimp           GGTAATGTGGGCCCTGATGCCCTCATTGAAATGGACCCTGAGGCCAGTGACCTCTCCCGG 660
gorilla         GGTAATGTGGGCCCTGATGCCCTCATTGAAATGGACCCTGAGGCCAGTGACCTCTCCCGG 660
human           GGTAATGTGGGCCCTGATGCCCTCATTGAAATGGACCCTGAGGCCAGTGACCTCTCCCGG 660
orang           GGTAATGTGGGCCCTGATGCCCTCATTGAAATGGACCCTGAGGCCAGTGACCTCTCCCAG 660
                ************************************************************ *

chimp           GGCATTCTGGCGCTCAGAGATGTGGCTGCTGAGGCTGGGGTCAGCAAATATACAGGGGGC 720
gorilla         GGCATTCTGGCGCTCAGAGATGTGGCTGCTGAGGCTGGGGTCAGCAAATATACAGGGGGC 720
human           GGCATTCTGGCGCTCAGAGATGTGGCTGCTGAGGCTGGGGTCAGCAAATATACAGGGGGC 720
orang           GGCATTCTGGCGCTCAGAGATGTGGCTGCTGAGGCTGGGGTCAGCAAATATACAGGGGGC 720
                ************************************************************

chimp           CGAGGCGTCAGCGTGGGCCCCATCCTCAGCAGCAGTGCCTCGGATATCTTCTGCGACAAC 780
gorilla         CGAGGCGTCAGCGTGGGCCCCATCCTCAGCAGCAGTGCCTCGGATATCTTCTGCGACAAC 780
human           CGAGGCGTCAGCGTGGGCCCCATCCTCAGCAGCAGTGCCTCGGATATCTTCTGCGACAAT 780
orang           CGAGGCGTCAGCGTGGGCCCCATCCTCAGCAGCAGTGCCTCGGATATCTTCTGCGACAAC 780
                ************************************************************

chimp           GAGAATGGGCCTAACTTCCTTTTCCACAACCGGGGCGATGGCACCTTTGTGGACGCTGCG 840
gorilla         GAGAATGGGCCTAACTTCCTTTTCCACAACCGGGGCGATGGCACCTTTGTGGACGCTGCG 840
human           GAGAATGGGCCTAACTTCCTTTTCCACAACCGGGGCGATGGCACCTTTGTGGACGCTGCG 840
orang           GAGAATGGGCCTAACTTCCTTTTCCACAACCGGGGCGATGGCACCTTTGTGGACGCTGCG 840
                ************************************************************

chimp           GCCAGTGCTGGTGTGGACGACCCCCACCAGCATGGGCGAGGTGTCGCCCTGGCTGACTTC 900
gorilla         GCCAGTGCTGGTGTGGACGACCCCCACCAGCATGGGCGAGGTGTCGCCCTGGCTGACTTC 900
human           GCCAGTGCTGGTGTGGACGACCCCCACCAGCATGGGCGAGGTGTCGCCCTGGCTGACTTC 900
orang           GCCAGTGCTGGTGTGGACGACCCCCACCAGCATGGGCGAGGTGTCGCCCTGGCTGACTTC 900
                ************************************************************

chimp           AACCGTGATGGCAAAGTGGACATCGTCTATGGCAACTGGAATGGCCCCCACCGCCTCTAT 960
gorilla         AACCGTGATGGCAAAGTGGACATCGTCTATGGCAACTGGAATGGCCCCCACCGCCTCTAT 960
human           AACCGTGATGGCAAAGTGGACATCGTCTATGGCAACTGGAATGGCCCCCACCGCCTCTAT 960
orang           AACCGTGATGGCAAAGTGGACATCGTCTATGGCAACTGGAATGGCCCCCACCGCCTCTAT 960
                ************************************************************

chimp           CTGCAGATGAGCGCCCATGGGAAGTTCCGCTTCCGGGACATCGCCTCACCCAAGTTCTCC 1020
gorilla         CTGCAGATGAGCACCCATGGGAAGGTCCGCTTCCGGGACATCGCCTCACCCAAGTTCTCC 1020
human           CTGCAAATGAGCACCCATGGGAAGGTCCGCTTCCGGGACATCGCCTCACCCAAGTTCTCC 1020
orang           CTGCAGATGAGCGCCCATGGGAAGGTCCGCTTCCGGGACATCGCCTCGCCCAAGTTCTCC 1020
                ***** ****** *********** ********************** ************

chimp           ATGCCCTCCCCTGTCCGCACGGTCATCACCGCCGACTTTGACAATGACCAGGAGCTGGAG 1080
gorilla         ATGCCCTCCCCTGTCCGCACGGTCATCACCGCCGACTTTGACAATGACCAGGAGCTGGAG 1080
human           ATGCCCTCCCCTGTCCGCACGGTCATCACCGCCGACTTTGACAATGACCAGGAGCTGGAG 1080
orang           ATGCCCTCCCCTGTCCGCACGGTCATCACCGCTGACTTTGACAATGACCAGGAGCTGGAG 1080
                ******************************* *************************

chimp           ATCTTCTTCAACAACATCGCCTACCGCAGCTCCTCAGCCAACCGCCTCTTCCGCGTCATC 1140
```

```
gorilla         ATCTTCTTCAACAACATCGCCTACCGCAGCTCCTCAGCCAACCGCCTCTTCCGCGTCATC 1140
human           ATCTTCTTCAACAACATTGCCTACCGCAGCTCCTCAGCCAACCGCCTCTTCCGCGTCATC 1140
orang           ATCTTCTTCAACAACATCGCCTACCGCAGCTCCTCAGCCAACCGCCTCTTCCGCGTCATC 1140
                ***************** ****************************************

chimp           CGTAGGGAGCACGGAGACCCCCTCATCGAGGAGCTCAATCCCGGCGATGCCTTGGAGCCT 1200
gorilla         CGTAGGGAGCACGGAGACCCCCTCATCGAGGAGCTCAATCCCGGCGATGCCTTGGAGCCT 1200
human           CGTAGAGAGCACGGAGACCCCCTCATCGAGGAGCTCAATCCCGGCGACGCCTTGGAGCCT 1200
orang           CGTAGGGAGCACGGAGATCCCCTCATCGAGGAGCTCAATCCCGGTGACGCCTTGGAACCT 1200
                ***** *********** ************************** ** ******** ***

chimp           GAGGGCCGGGGCACAGGGGGTGTGGTGACCGACTTCGACGGAGATGGGATGCTGGACCTC 1260
gorilla         GAGGGCCGGGGCACAGGGGGTGTGGTGACCGACTTCGACGGAGACGGGATGCTGGACCTC 1260
human           GAGGGCCGGGGCACAGGGGGTGTGGTGACCGACTTCGACGGAGACGGGATGCTGGACCTC 1260
orang           GAGGGCCGGGGCACAGGGGGTGTGGTGACCGACTTCGATGGAGACGGGATGCTGGACCTC 1260
                ************************************ ***** ***************

chimp           ATCTTGTCCCATGGAGAGTCCATGGCTCAGCCGCTGTCCGTCTTCCGGGGCAACCAGGGC 1320
gorilla         ATCTTGTCCCATGGAGAGTCCATGGCTCAGCCGCTGTCCGTCTTCCGGGGCAATCAGGGC 1320
human           ATCTTGTCCCATGGAGAGTCCATGGCTCAGCCGCTGTCCGTCTTCCGGGGCAATCAGGGC 1320
orang           ATCTTGTCCCATGGAGAGTCCATGGCTCAGCCGCTGTCCGTCTTCCGGGGCAATCAGGGC 1320
                ************************************************** ******


chimp           TTCAACAACAACTGGCTGCGAGTGGTGCCACGCACCCGGTTTGGGGCCTTTGCCAGGGGG 1380
gorilla         TTCAACAACAACTGGCTGCGAGTGGTGCCACGCACCCGGTTTGGGGCCTTTGCCAGGGGG 1380
human           TTCAACAACAACTGGCTGCGAGTGGTGCCACGCACCCGGTTTGGGGCCTTTGCCAGGGGA 1380
orang           TTCAACAACAACTGGCTGCGAGTGGTGCCACGCACCCGGTTTGGGGCCTTTGCCAGGGGG 1380
                ************************************************************

chimp           GCTAAGGTCGTGCTCTACACCAAGAAGAGCGGGGCCCACCTGAGGATCATCGACGGGGGC 1440
gorilla         GCTAAGGTCGTGCTCTACACCAAGAAGAGCGGGGCCCACCTAAGGATCATCGACGGGGGC 1440
human           GCTAAGGTCGTGCTCTACACCAAGAAGAGTGGGGCCCACCTGAGGATCATCGACGGGGGC 1440
orang           GCTAAGGTCGTGCTCTACACCAAGAAGAGCGGGGCCCACTTGAGGATCATCGACGGGGGC 1440
                **************************** ********* * ******************

chimp           TCAGGCTACCTGTGTGAGATGGAGCCCGTGGCACACTTTGGCCTGGGGAAGGATGAAGCC 1500
gorilla         TCAGGCTACCTGTGTGAGATGGAGCCTGTGGCACACTTTGGCCTGGGGAAGGATGAAGCC 1500
human           TCAGGCTACCTGTGTGAGATGGAGCCCGTGGCACACTTTGGCCTGGGGAAGGATGAAGCC 1500
orang           TCAGGCTACCTGTGTGAGATGGAGCCCGTGGCACACTTTGGCCTGGGGAAGGATGAAGCC 1500
                ************************** ********************************

chimp           AGCAGTGTGGAGGTGACGTGGCCAGATGGCAAGATGGTGAGCCGGAACGTGGCCAGCGGG 1560
gorilla         AGCAGTGTGGAGGTGACGTGGCCAGATGGCAAGATGGTGAGCCGGAATGTGGCCAGCGGG 1560
human           AGCAGTGTGGAGGTGACGTGGCCAGATGGCAAGATGGTGAGCCGGAACGTGGCCAGCGGG 1560
orang           AGCAGTGTGGAGGTGACGTGGCCAGATGGCAAGATGGTGAGCCGGAACGTGGCCAGCGGG 1560
                ********************************************* ************

chimp           GAGATGAACTCAGTGCTGGAGATCCTCTACCCCCGGGATGAGGACACACTTCAGGACCCA 1620
gorilla         GAGATGAACTCAGTGCTGGAGATCCTCTACCCCCGGGATGAGGACACACTTCAGGACCCA 1620
human           GAGATGAACTCAGTGCTGGAGATCCTCTACCCCCGGGATGAGGACACACTTCAGGACCCA 1620
orang           GAGATGAACTCAGTGCTGGAGATCCTCTATCCCCGGGATGAGGACACACTTCAGGACCCA 1620
                **************************** ******************************

chimp           GCCCCACTGGAGTGTGGCCAAGGATTCTCCCAGCAGGAAAATGGCCATTGCATGGACACC 1680
gorilla         GCCCCACTGGAGTGTGGCCAAGGATTCTCCCAGCAGGAAAATGGCCATTGCATGGACACC 1680
human           GCCCCACTGGAGTGTGGCCAAGGATTCTCCCAGCAGGAAAATGGCCATTGCATGGACACC 1680
orang           GCCCCACTGGAGTGTGGCCAAGGATTCTCCCAGCAGGAAAATGGCCATTGCATGGACACC 1680
                ************************************************************

chimp           AATGAATGCATCCAGTTCCCATTCGTGTGCCCTCGAGACAAGCCCGTATGTGTCAACACC 1740
gorilla         AATGAATGCATCCAGTTCCCATTCGTGTGCCCTCGAGACAAGCCCGTATGTGTCAACACC 1740
human           AATGAATGCATCCAGTTCCCATTCGTGTGCCCTCGAGACAAGCCCGTATGTGTCAACACC 1740
orang           AATGAATGCATCCAGTTCCCATTCGTGTGCCCTCGAGACAAGCCTGTATGTGTCAACACC 1740
                ****************************************** ***************

chimp           TATGGAAGCTACAGGTGCCGGACCAACAAGAAGTGCAGTCGGGGCTACGAGCCCAACGAG 1800
```

```
gorilla         TATGGAAGCTACAGGTGCCGGACCAACAAGAAGTGCAGTCGGGGCTACGAGCCCAACGAG 1800
human           TATGGAAGCTACAGGTGCCGGACCAACAAGAAGTGCAGTCGGGGCTACGAGCCCAACGAG 1800
orang           TATGGAAGCTACAGGTGCCGGACCAACAAGAAGTGCAGTCGGGGCTATGAGCCCAACGAG 1800
                ********************************************* ************

chimp           GATGGCACAGCCTGTGTGGTGCCGCTGCTGGAGCTGCCACTGCTGCACCGGTCCTCGTAG 1860
gorilla         GATGGCACAGCCTGTGTGGTGACGCTGCTGGAGCTGCCACTGCTGCACCGGTCCTCGTAG 1860
human           GATGGCACAGCCTGCGTGGGGCCGCTGCTGGAGCTGCCACTGCTGCACCGGTCCTCGTAG 1860
orang           GATGGCACAGCCTGTGTGGTGCCACTGCTGGAGCTGCCACTGCTGCACCGGTCCTCGTAG 1860
                ************** **** * * ************************************

chimp           ATGGAGATCTCAATCTGGGGTCGGCGGTTAAGGAGAGCTGCGAGCCCAGCTGCTGA 1916
gorilla         ATGGAGATCTCAATCTGGGGTCGGCGGTTAAGGAGAGCTGCGAGCCCAGCTGCTGA 1916
human           ATGGAGATCTCAATCTGGGGTCGGTGGTTAAGGAGAGCTGCGAGCCCAGCTGCTGA 1916
orang           ATGGAGATCTCAATCTGGAGTCGGCGGTTAAGGAGAGCTGCGAGCCCAGCTGCTGA 1916
                ****************** ***** ******************************
```

2.1.2 Amino Acid Alignment

```
chimp          MAPSADPGMSRMLLFLLLLWFLPITEGSQRAEPMFTAVTNSVLPPDYDSNPTQLNYGVAV 60
orang          MAPSADPGMSRMLLFLLLLWFLPITEGSQRAEPMFTAVTNSVLPPDYDSNPTQLNYGVAV 60
human          MAPSADPGMSRMLLFLLLLWFLPITEGSQRAEPMFTAVTNSVLPPDYDSNPTQLNYGVAV 60
gorilla        MAPSADPGMSRMLLFLLLLWFLPITEGSQRAEPMFTAVTNSVLPPDYDSNPTQLNYGVAV 60
               ************************************************************

chimp          TDVDHDGDFEIVVAGYNGPNLVLKYDRAQKRLVNIAVDERSSPYYALRDRQGNAIGVTAC 120
orang          TDVDHDGDFEIVVAGYNGPNLVLKYDRAQKRLVNIAVDERSSPYYALRDRQGNAIGVTAC 120
human          TDVDHDGDFEIVVAGYNGPNLVLKYDRAQKRLVNIAVDERSSPYYALRDRQGNAIGVTAC 120
gorilla        TDVDHDGDFEIVVAGYNGPNLVLKYDRAQKRLVNIAVDERSSPYYALRDRQGNAIGVTAC 120
               ************************************************************

chimp          DIDGDGREEIYFLNTNNAFSGVATYTDKLFKFRNNRWEDILSDEVNVARGVASLFAGRSV 180
orang          DIDGDGREEIYFLNTNNAFSGVATYTDKLFKFRNNRWEDILSDEVNVARGVASLFAGRSV 180
human          DIDGDGREEIYFLNTNNAFSGVATYTDKLFKFRNNRWEDILSDEVNVARGVASLFAGRSV 180
gorilla        DIDGDGREEIYFLNTNNAFSGVATYTDKLFKFRNNRWEDILSDEVNVARGVASLFAGRSV 180
               ************************************************************

chimp          ACVDRKGSGRYSIYIANYAYGNVGPDALIEMDPEASDLSRGILALRDVAAEAGVSKYTGG 240
orang          ACVDRKGSGRYSIYIANYAYGNVGPDALIEMDPEASDLSQGILALRDVAAEAGVSKYTGG 240
human          ACVDRKGSGRYSIYIANYAYGNVGPDALIEMDPEASDLSRGILALRDVAAEAGVSKYTGG 240
gorilla        ACVDRKGSGRYSIYIANYAYGNVGPDALIEMDPEASDLSRGILALRDVAAEAGVSKYTGG 240
               ************************************: ********************

chimp          RGVSVGPILSSSASDIFCDNENGPNFLFHNRGDGTFVDAAASAGVDDPHQHGRGVALADF 300
orang          RGVSVGPILSSSASDIFCDNENGPNFLFHNRGDGTFVDAAASAGVDDPHQHGRGVALADF 300
human          RGVSVGPILSSSASDIFCDNENGPNFLFHNRGDGTFVDAAASAGVDDPHQHGRGVALADF 300
gorilla        RGVSVGPILSSSASDIFCDNENGPNFLFHNRGDGTFVDAAASAGVDDPHQHGRGVALADF 300
               ************************************************************

chimp          NRDGKVDIVYGNWNGPHRLYLQMSAHGKFRFRDIASPKFSMPSPVRTVITADFDNDQELE 360
orang          NRDGKVDIVYGNWNGPHRLYLQMSAHGKVRFRDIASPKFSMPSPVRTVITADFDNDQELE 360
human          NRDGKVDIVYGNWNGPHRLYLQMSTHGKVRFRDIASPKFSMPSPVRTVITADFDNDQELE 360
gorilla        NRDGKVDIVYGNWNGPHRLYLQMSTHGKVRFRDIASPKFSMPSPVRTVITADFDNDQELE 360
               ************************: ***.******************************

chimp          IFFNNIAYRSSSANRLFRVIRREHGDPLIEELNPGDALEPEGRGTGGVVTDFDGDGMLDL 420
orang          IFFNNIAYRSSSANRLFRVIRREHGDPLIEELNPGDALEPEGRGTGGVVTDFDGDGMLDL 420
human          IFFNNIAYRSSSANRLFRVIRREHGDPLIEELNPGDALEPEGRGTGGVVTDFDGDGMLDL 420
gorilla        IFFNNIAYRSSSANRLFRVIRREHGDPLIEELNPGDALEPEGRGTGGVVTDFDGDGMLDL 420
               ************************************************************

chimp          ILSHGESMAQPLSVFRGNQGFNNNWLRVVPRTRFGAFARGAKVVLYTKKSGAHLRIIDGG 48
orang          ILSHGESMAQPLSVFRGNQGFNNNWLRVVPRTRFGAFARGAKVVLYTKKSGAHLRIIDGG 480
human          ILSHGESMAQPLSVFRGNQGFNNNWLRVVPRTRFGAFARGAKVVLYTKKSGAHLRIIDGG 480
gorilla        ILSHGESMAQPLSVFRGNQGFNNNWLRVVPRTRFGAFARGAKVVLYTKKSGAHLRIIDGG 480
               ************************************************************

chimp          SGYLCEMEPVAHFGLGKDEASSVEVTWPDGKMVSRNVASGEMNSVLEILYPRDEDTLQDP 540
orang          SGYLCEMEPVAHFGLGKDEASSVEVTWPDGKMVSRNVASGEMNSVLEILYPRDEDTLQDP 540
human          SGYLCEMEPVAHFGLGKDEASSVEVTWPDGKMVSRNVASGEMNSVLEILYPRDEDTLQDP 540
gorilla        SGYLCEMEPVAHFGLGKDEASSVEVTWPDGKMVSRNVASGEMNSVLEILYPRDEDTLQDP 540
               ************************************************************

chimp          APLECGQGFSQQENGHCMDTNECIQFPFVCPRDKPVCVNTYGSYRCRTNKKCSRGYEPNE 600
orang          APLECGQGFSQQENGHCMDTNECIQFPFVCPRDKPVCVNTYGSYRCRTNKKCSRGYEPNE 600
human          APLECGQGFSQQENGHCMDTNECIQFPFVCPRDKPVCVNTYGSYRCRTNKKCSRGYEPNE 600
gorilla        APLECGQGFSQQENGHCMDTNECIQFPFVCPRDKPVCVNTYGSYRCRTNKKCSRGYEPNE 600
               ************************************************************

chimp          DGTACVVPLLELPLLHRSS-MEISIWGRRLRRAASPAA 637
orang          DGTACVVPLLELPLLHRSS-MEISIWSRRLRRAASPAA 637
human          DGTACVGPLLELPLLHRSS-MEISIWGRWLRRAASPAA 637
gorilla        DGTACVVTLLELPLLHRSS-MEISIWGRRLRRAASPAA 637
               ****** .*********** ******.* *********
```

2.1.3 PAML Codeml control file

```
seqfile = crtac1.nuc * sequence data filename
treefile = crtac1.trees * tree structure file name
outfile = rj * main result file name
noisy = 9 * 0,1,2,3,9: how much rubbish on the screen
verbose = 0 * 0: concise; 1: detailed, 2: too much
runmode = 0 * 0: user tree; 1: semi-automatic; 2: automatic
* 3: StepwiseAddition; (4,5): PerturbationNNI; -2:
pairwise
seqtype = 1 * 1: codons; 2: AAs; 3: codons-->AAs
CodonFreq = 2 * 0: 1/61 each, 1: F1X4, 2: F3X4, 3: codon
table
* ndata = 10
clock = 0 * 0: no clock, 1: clock; 2: local clock; 3:
CombinedAnalysis
aaDist = 0 * 0: equal, +: geometric; -: linear,
1-6: G1974,Miyata,c,p,v,a
aaRatefile = dat/jones.dat * only used for aa seqs with
model=empirical
(_F)
* dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your
own

model[1] = 0
* models for codons:
* 0: one, 1: b, 2: 2 or more dN/dS ratios for branches
* models for AAs or codon-translated AAs:
* 0: poisson, 1: proportional, 2: Empirical,
3: Empirical+F
* 6: FromCodon, 7: AAClasses, 8: REVaa_0, 9: REVaa
(nr=189)

NSsites[2] = 0 * 0: one w;1: neutral;2: selection; 3:
discrete;4: freqs;
* 5: gamma;6: 2gamma;7: beta;8: beta&w;9: beta&gamma;
* 10: beta&gamma+1; 11: beta&normal>1; 12: 0&2normal>1;
* 13: 3normal>0
icode = 0 * 0: universal code; 1: mammalian mt; 2-10: see
below
Mgene = 0
* codon: 0: rates, 1: separate; 2: diff pi, 3: diff kapa,
4: all diff
```

```
* AA: 0: rates, 1: separate

fix_kappa = 0 * 1: kappa fixed, 0: kappa to be estimated
kappa = 2 * initial or fixed kappa
```
fix_omega[3] = 0 * 1: omega or omega_1 fixed, 0: estimate
```
omega = .4 * initial or fixed omega, for codons or codon-
based AAs

fix_alpha = 1 * 0: estimate gamma shape parameter; 1: fix
it at alpha
alpha = 0. * initial or fixed alpha, 0: infinity (constant
rate)
Malpha = 0 * different alphas for genes
ncatG = 8 * # of categories in dG of NSsites models

getSE = 0 * 0: don't want them, 1: want S.E.s of estimates
RateAncestor = 1 * (0,1,2): rates (alpha>0) or ancestral
states (1 or 2)

Small_Diff = .5e-6
cleandata = 1 * remove sites with ambiguity data (1: yes,
0: no)?
* fix_blength = -1 * 0: ignore, -1: random, 1: initial, 2:
fixed
method = 0 * Optimization method 0: simultaneous; 1: one
branch a
time

* Genetic codes: 0: universal, 1: mammalian mt., 2: yeast
mt., 3: mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm
mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian
mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of
GENEBANK.
```

[1]The **model parameter** was changed to 0 for M0 (Uniform model)
Changed to 1 for M1 (Free ratio model)
Changed to 2 for MA (Branch-site model)
Changed to 0 for M1a, M2a, and M3 (Site models)
Changed to 0 for M7, M8, and M8a (Beta distributed models)

[2]The **NSsites parameter** was changed to 0 for M0,and M1 (Branch models)
Changed to 1 for M1a (Neutral model)
Changed to 2 for M2a (Positive selection model)
Changed to 3 for M3 (Discrete selection model)
Changed to 7 for M7 (Beta distributed neutral model)
Changed to 8 for M8, and M8a (Beta distributed positive selection models)
Changed to 2 for MA (Branch-site model)

[3]**ω** was fixed at 1 for model M8a (Beta distributed positive selection with fixed **ω**)

TREE FILE:

```
4 1
((chimp,human),#1 gorilla,orang);
```

2.1.4 UNIX commands used for studying GAGP based population genetics of hominoids

```
1. zcat Homo.vcf.gz|sed '/^#/d'| grep chr10|awk
'{if($2>=99780120&&$2<=99782105)print;}' >
human_CRTAC1__SNPs.txt

2. cat Gorilla.vcf|sed '/^#/d'| grep chr10|awk
'{if($2>=111411673&&$2<=111413657)print;}' >
gorilla_CRTAC1_SNPs.txt

3. zcat Pan_troglodytes.vcf.gz|sed '/^#/d'| grep chr10|awk
'{if($2>=97518314&&$2<=97520300)print;}' >
chimp_CRTAC1_SNPs.txt

4. zcat Pongo_abelii.vcf.gz|sed '/^#/d'| grep chr10|awk
'{if($2>=97012196&&$2<=97014163)print;}' >
orangutan_CRTAC1_SNPs.txt
```

```
Chimp: First 10 = ellioti
11-16 = schweinfurthii
17-20 = troglodytes
21-25 = verus
```

2.2 *Cis*-regulatory elements data

2.2.1 Promoter alignment

```
Orangutan       TACTGTCCTAGACCCCTGAGGAAACAGGTTGACATTGGACAGAGCCCCAAGAGACTCAGG
Chimpanzee      TACTGTCCTAGACCCCTGAAGAAACAGGTTGACATTGGACAGAGCCCCAAGAGACTCAGG
Bonobo          TACTGTCCTAGACCCCTGAAGAAACAGGTTGACATTGGACAGAGCCCCAAGAGACTCAGG
Human           TACTGTCCTAGACCCCTGAAGAAACAGGTTGACATTGGACCGAGCCCCAAGAGACTCAGG
Gorilla         TACTGTTCTAGACCCCTGAAGAAACAGGTTGACACTGGACAGAGCCCCCAGAGACTCAGG
                ****** ************ ************** ***** ******* ***********

Orangutan       GTTCCTAATGATCTGAAAGTATACTCAAGG------------------TGGTGTGTGTGT
Chimpanzee      GTTCCTAATGATCCTGAAAGTATACCCAAGGTGTGTGTGTGTGTGTG--------TGTGT
Bonobo          GTTCCTAATGATCCTGAAAGTATACTCAAGGTGTGTGTGTGT------------------
Human           GTTCCTAATGATCCTGAAAGTATACTCAAGGTGTGTGTGTGTGTGTGTGTGTGTGTGT
Gorilla         GTTCCTAATGATCCTGAAAGTATATGCAGG-----GGTGTGTGTGTGTGTGTGTGTGTGT
                *************   **         * *

Orangutan       GTGTGTGTAAATGTGTGTGTTTGTATATTTTTCTAGAAAAATGATTCAGTAATTTTTGCC
Chimpanzee      GTGTAAGTAAATGTGTGTGTTTGTATATTTTTCTAGAAAAATGATTCAGTAATTTTTGCC
Bonobo          ------GTAAATGTGTGTGTTTGTATATTTTTCTAGAAAAATGATTCAGTAATTTTTGCC
Human           GTGTGTGTAAATGTGTGTGTTTGTATATTTTTCTAGAAAAATGATTCAGTAATTTTTGCC
Gorilla         GTGTGTGTAAATGTGTGTGTTTGTATATTTTTCTATGAACATGATTCAGTAATTTTTGCC
                      ******************************  ** ********************

Orangutan       AGATTTTCAAAGGGGCCCACAACCCTTAAAGGCTTGATGGGGGCTCAAAAAGCCAGAAAG
Chimpanzee      AGATTTTCAAAGGGGCCCACAA-CCTTACAGGCTTGATGGGGGCTCAAAAAGCCAGAAAG
Bonobo          AGATTTTCAAAGGGGCCCACAA-CCTTACAGGCTTGATGGGGGCTCAAAAAGCCAGAAAG
Human           AGATTTTCAAAGGGGCCCACAA-CCTTACAGGCTTGATGGGGGCTCAAAAAGCCAGAAAG
Gorilla         AGATTTTCAAAGGGGCCCACAA-CCTTAAAGGCTTGATGGGGGCTCAAAAAGCCAGAAAG
                ********************** ***** ******************************

Orangutan       TTTGTCTGTGCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGCC
Chimpanzee      TTTGTCTGTGCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGCC
Bonobo          TTTGTCTGTGCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGCC
Human           TTTGTCTGTGCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGCC
Gorilla         TGTGTCTGTGCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGCC
                * ************************************************************

Orangutan       CAGACTAGGTCCTGGTAGTTCTGCTCAAAATCTCTGGGTAATAGGGGCCAGGGGCTGGAT
Chimpanzee      CAGACTAGGTCCTGATAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGAGGCTGGAT
Bonobo          CAGACTAGGTCCTGATAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGAGGCTGGAT
Human           CAGACTAGGTCCTGGTAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGAGGCTGGAT
Gorilla         CAGAGTAGGTCCTGGTAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGGGGCTGGAT
                **** ********* ***************** *** *** ********* ********

Orangutan       GTGTGGCCCTCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGCC
Chimpanzee      GTGTGGCCCTCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGCC
Bonobo          GTGTGGCCCTCTATAAAAGAACACATTTAAATCATTCTAAGAGTCAAGGGCCCAGAAGCC
Human           GTGTGGCCCTCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGCC
Gorilla         GTGTGGCCCTCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGCC
                *************************** ********************************

Orangutan       AGACTAGTGGGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTTC
Chimpanzee      AGACTAGTGGGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTTC
Bonobo          AGACTAGTGGGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTTC
Human           AGACTAGTGGGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTTC
Gorilla         AGACTAGTGGGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTTC
                ************************************************************

Orangutan       CCTCTCCGATGGTGGATTCCCAAGTCTTTCTCTTCTGTAGCCCCTTTACCCTGCTTTCAG
Chimpanzee      CCTCTCCGATGGTGGATTCCCAAGTCTTTCTCTTCTGTGGCCCCTTTACCTGGCTTTCAG
Bonobo          CCTCTCCGATGGTGGATTCCCAAGTCTTTCTCTTCTGTGGCCCCTTTACCTGGCTTTCAG
Human           CCTCTCCGATGGTGGATTCCCAAGTCTTTCTCTTCTGTGGCCCCTTTACCTGGCTTTCAG
```

```
Gorilla         CCTCTCCGATGGTGGATTCCCAAGTCTTTCTCTTCTGTAGCCCCTTTACCTGGCTTTCAG
                ************************************ ***********  ********

Orangutan       AAACCCAAAGACCTCACTCACAGAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTAC
Chimpanzee      AAACCCAAAGACCTCACTCATGGAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTAC
Bonobo          AAACCCAAAGACCTCACTCATGGAAGCTCAGAATTCATTCAATAAACACTGAGTGCCTAC
Human           AAACCCAAAGACCTCACTCATGGAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTAC
Gorilla         AAACCCAAAGACCTCACTCATGAAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTAC
                ********************   ****************************** ******

Orangutan       TATGTGCCAGGCTCTATTTTAGGCGTAAAATAAAGGATTTGCCTTCAGAGTGTGACCCAC
Chimpanzee      TATGTGCCAGGCTCTATTCTAGGCGTAAAAA-AAGGATTTGCCTTCAGAGTGTGACCCAC
Bonobo          TATGTGCCAGGCTCTATTCTAGGCGTAAAAA-AAGGATTTGCCTTCAGAGTGTGACCCAC
Human           TATGTGCCAGGCTCTATTCTAGGCGTAAAATAAAGGATTTGCCTTCAGAGTGTGACCCAC
Gorilla         TATGAGCCAGGCTCTATTCTAGGCGTAAAATAAAGGATTTGCCTTCAGAGTGTGACCCAC
                **** ************* ***********  ****************************

Orangutan       CAAGGCCAGCTTCCTGCCTGGACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTTT
Chimpanzee      CAAGGCCAGTTTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTTT
Bonobo          CAAGGCCAGTTTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTTT
Human           CAAGGCCAGTTTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTTT
Gorilla         CAAGGCCAGTTTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTTT
                ********* ********** ****************************************

Orangutan       CTGGCGCTTAGCACCCCCATTCCCCTGGTCCAGAGAGGCAGAAGCCATCCCCCCCCCCAC
Chimpanzee      CTGGCGCTTAGCACCCCCATTCCCCTGGTCCAGAGAGGCAGAAACCATCCCC-CC-----
Bonobo          CTGGCGCTTAGCACCCCCATTCCCCTGGTCCAGAGAGGCAGAAACCATCCCC-CC-----
Human           CTGGCGCTTAGCACCCCCATTCCCCTGGTCCAGAGAGGCAGAAACCATCCCC-CC-----
Gorilla         CTGGCGCTTAGCACCCCCATTTCCCTGGTCCAGAGAGGCAGAAACCATCCCCCCC-----
                ********************* ********************* ******** **

Orangutan       CACCACCACCACCAAGGCAGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCA
Chimpanzee      ----ACCACCACCAAGGCAGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCA
Bonobo          ----ACCACCACCAAGGCAGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCA
Human           ----ACCACCACCAAGGCAGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCA
Gorilla         ----ACCACCACCAAGGCAGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCA
                    ********************************************************

Orangutan       AAGCGGTGAGGGAGGCTAGTAGGACGCAGGCGGCAGGGGCGGGAGGGCCAGGCCCGACTC
Chimpanzee      AAGCGGTGAGGGAGGCTAGTAGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTC
Bonobo          AAGCGGTGAGGGAGGCTAGTAGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTC
Human           AAGCGGTGAGGGAGGCTAGTAGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTC
Gorilla         AAGCGGTGAGGGAGGCTAATAGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTC
                ****************** ****************** ***** ****************

Orangutan       GGCCACTGCTGGGGTAGGGACTAGGTGGGACGGGGTGGGGGCACTGCTGGTGGGGACAGG
Chimpanzee      GGCCACTGCTGGGGTAGGGACTAGGTGGGACGGGGTGGGGGCACTGCTGGTGGGGACAGG
Bonobo          GGCCACTGCTGGGGTAGGGACTAGGTGGGACGGGGTGGGGGCACTGCTGGTGGGGACAGG
Human           GGCCACTGCTGGGGTAGGGACTAGGTGGGATGGGGTGGGGGCACTGCTGGTGGGGACG--
Gorilla         GGCCATTGCTGGGGTAGGGACTAGGTGGGACGGGGTGGGGGCACTGCTGGTGGGGACAGG
                ***** ************************ **************************

Orangutan       GGGTGGGGTAGAAGCGGCGCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCGGA
Chimpanzee      GGGTGGGGTAGAAGCGGCGCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCCGA
Bonobo          GGGTGGGGTAGAAGCGGCGCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCCGA
Human           GGGTGGGGTAGAAGCGGCGCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCCGA
Gorilla         GGGTGGGGTAGAAGCGGCGCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCTTGATCTCGGA
                ********************************************** ******* **

Orangutan       GCTGCGCTTGCCCGCTCCCCGCCTACGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGA
Chimpanzee      GCTGCGCTCGCCCGCTCCCCGCCTGCGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGA
Bonobo          GCTGCGCTCGCCCGCTCCCCGCCTGCGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGA
Human           GCTGCGCTCGCCCGCTCCCCGCCTGCGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGA
Gorilla         GCTGCGCTCGCCCGCTCCCCGCCTGCGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGA
                ******** *************** **********************************

Orangutan       GCAGTCCCCAGGAGGTGCTCGGACACGTCCCCAGGCTGGATAAAGATCGGCTCGGCGCTA
```

```
Chimpanzee      GCAGTCCCCGGGAGGTGCTCGGACACGTCCCCAGGCTGGATAAAGATCGGCTCGGCGCTA
Bonobo          GCAGTCCCCGGGAGGTGCTCGGACACGTCCCCAGGCTGGATAAAGATCGGCTCGGCGCTA
Human           GCAGTCCCCGGGAGGTGCTCGGACACGTCCTCAGGCTGGATAAAGATCGGCTCGGCGCTA
Gorilla         GCAGTCCCCGGGAGGTGCTCGGACACGTCCCCAGGCTGGATAAAGATCGGCTCGGCGCTA
                ********* ******************** ******************************

Orangutan       GCTCCGTAGTCGAAATCTCGCCATCAGCGCGGCTCGCGCGGCCGCTTTGGCCCGGCCCGG
Chimpanzee      GCTCCGTAGTCGAAATCTCGCCATCAGCGCGGCTCGCTCGGCCGCTTTGGCCCGGCCCGG
Bonobo          GCTCCGTAGTCGAAATCTCGCCATCAGCGCGGCTCGCTCGGCCGCTTTGGCCCGGCCCGG
Human           GCTCCGTAGTCGAAATCTCGCCATCAGCGCGGCTCGCTTGGCCGCTTTGGCCCGGCCCGG
Gorilla         GCTCCGTAGTCGAAATCTCGCCATCAGCGCGGCTCGCTCGGGCGATTTGGCCCGGCTCGG
                ************************************  ** ** ************ ***

Orangutan       CGACGCCAGATCGCTATCCTGGAGTGAAATGGGAAGGCAGTGCCGCCGCTCCCGCCACCA
Chimpanzee      CGACGCCAGATCGCTATCCTGGGGGGAAATGGGAAGGGAGTGCCGCCGCTCCCGCCTCCA
Bonobo          CGACGCCAGATCGCTATCCTGGGGGGAAATGGGAAAGCAGTGCCGCCGCTCCCGCCTCCA
Human           CGACGCCAGATCGCTATCCTGGGGGGAAATGGGAAGGCAGTGCAGCCGCTCCCGCCTCCA
Gorilla         CGACGCCAGATCCCTATCCTGGGGGGAAATGGAAAGGCAGTGCCGCCGCTACCGCCTCCA
                ************ ********* * ******* ** * ***** ****** ***** ***

Orangutan       CCCTCGGTCCTGCGCGCAGGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCC
Chimpanzee      CCCTCGGTCCTGCGCGCATGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCC
Bonobo          CCCTCGGTCCTGCGCGCATGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCC
Human           CCCTCGGTCCTGCGCGCAGGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCC
Gorilla         CCCTCGGTCCTGCGCGCAGGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCC
                ****************** ****************************************

Orangutan       CCTTCGCGTTCCTTCCCTGCGCTGCCTCCCGAGGGACCCCCGCTTCCCTCCGGCCTGGGG
Chimpanzee      CCTTCGCGTTCCTTCCCTGCGCTGCCTCCCGAGGGACCCTCGCTC-CCTCCGGCCTGGGG
Bonobo          CCTTCGCGTTCCTTCCCTGCGCTGCCTCCCGAGGGACCCTCGCTC-CCTCCGGCCTGGGG
Human           CCTTCGCGTTCCTTCCCTGCGCTGCCTCCCGAGGGACCCTCGCTTCCCTCCGGCCTGGGG
Gorilla         CCTTGGCGTTCCTTCCCTGCGCAGCATCCCGAGGGACCCTCGCTTCCCTCCGGCCTGGGG
                **** ****************** ** ************* ****  **************

Orangutan       CCCCTAGCGCCCAGCCAGGCTAGGGAGCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCC
Chimpanzee      CCCCCAGCGCCCAGCCAGG-----CGCCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCC
Bonobo          CCCCCAGCGCCCAGCCAGG-----CGCCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCC
Human           CCCCCAGCGCCCAGCCAGG-----CGCCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCC
Gorilla         CCCCCAGCGCCCAGCCAGG-----CGCCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCC
                **** **************        **********************************

Orangutan       TGAAGGTTACGCGACGCAGTGGCGGGGCGCGGGAGGCGCCCGCCCTCG------------
Chimpanzee      TGAAGGTTACGCGACGCAGTGGCGGGGCGCGGGGGGCGCCCGCCCTCGCCGCCGCCTTCG
Bonobo          TGAAGGTTACGCGACGCAGTGGCGGGGCGCGGGGGGCGCCCGCCCTCGCCGCCGCCTTCG
Human           TGAAGGTTACGCGACGCAGTGGCGGGGCGCGGGGGGCGCCCGCCCTCGCCGCCGCCTTCG
Gorilla         TGAGGTTACGCGACGCAGTGTGCGGGGCGCGGGGGGCGCCCGCCCTCGCCGCCGCCCACC
                *** * *              ************ **************

Orangutan       CCGCCGCCCACCCCAGTGCCCGCCGCGCCGCGCCGGGCCAGCCTGGCTGCCGGCTGCTGC
Chimpanzee      CCGCCGCCCACCCCAGTGGCCGCCGCGCCGCGCCGGGCCAGCCTGGCTGCCGGCTGCTGC
Bonobo          CCGCCGCCCACCCCAGTGGCCGCCGCGCCGCGCCGGGCCAGCCTGGCTGCCGGCTGCTGC
Human           CCGCCGCCCACCCCAGTGGCCGCCGCGCCGCGCCGGGCCAGCCTGGCTGCCGGCTGCTGC
Gorilla         CC-------AGTGCCCGCCGCGCCGCGCCGCGCCGGGCCAGCCTGGCTGCCGGCTGCTGC
                **       *   *      ****************************************

Orangutan       CACCGCAATCCCGGCTCCTAAATCAGCGCCGGGAGGCGCTCCCTCCCCCCGCCCGGCTCT
Chimpanzee      CACCGCAATCCCGGCTCCTAAATCAGCGCGGGGAGGCGCTCCCTCCCCCAGCCCGGCTCT
Bonobo          CACCGCAATCCCGGCTCCTAAATCAGCGCGGGGAGGCGCTCCCTCCCCCAGCCCGGCTCT
Human           CACCGCAATCCCGGCTCCTAAATCAGCGCGGGGAGGCGCTCCCTCCCCCACGCCCGGCTCT
Gorilla         CACCGCAATCCCGGCTCCTAAATCAGCGCGGGGAGGCGCTCCCTCCCCCCGCCCTGCTCT
                **************************** ******************  **** *****

Orangutan       CCGGGCTCCCCGGGCCGCGATTGGCCGCGCCGGCGCCCCCCACCCCGGGCCCCCGGCTCC
Chimpanzee      GGGGGCTCTCCGGGCAGCGATTGGCCGCGCCGGCGCCCCCCACCCCGGGCCCCCGGCTCC
Bonobo          GGGGGCTCTCCGGGCCGCGATTGGCCGCGCCGGCGCCCCCCACCCCGGGCCCCCGGCTCC
Human           CCGGGCTCTCGGGGCCGCGATTGGCCGCGCCGGCGCCCCCCACCCCGGGCCCCCGGCTCC
Gorilla         CCGGGCTCTCCGGGCCGCGATTGGCCGCGCCGGCGCCCCCCACCCCGGGCCCCAGGCTCC
```

```
                 ****** * **** ************************************ ******

Orangutan        AGCTGCCGCGCCATTGGCTGCAGGCCTCCGCCAGCCTTTACATAAGACCGGGCGCGCTCG
Chimpanzee       AGCTGCCGCGCCATTGGCTGCGGGCCTCCGCCAGCCTTTACATAAGACCGGGCGCGCTCG
Bonobo           AGCTGCCGCGCCATTGGCTGCGGGCCTCCGCCAGCCTTTACATAAGACCGGGAGCGCTCG
Human            AGCTGCCGCGCCATTGGCTGCGGGCCTCCGCCAGCCTTTACATAAGACCGGGCGCGCTCG
Gorilla          AGCTGCCGCGCCATTGGCTGCGGGCCTCCGCCAGCCTTTACATAAGACCGGGCGCGCTCG
                 ********************* ***************************** *******

Orangutan        AGTGGAGTTGTATAAAGCGAGCGCGCGGCGTCGGGGCGGGAGGCTCGAGGCCAGCCCGGG
Chimpanzee       AGTGGAGTTGTATAAAGCGAGCGCGCGGCGTCGGGGCGGGAGGCTCGAGGCCAGCCCGGG
Bonobo           AGTGGAGTTGTATAAAGCGAGCGCGCGGCGTCGGGGCGGGAGGCTCGAGGCCAGCCCGGG
Human            AGTGGAGTTGTATAAAGCGAGCGCGCGGCGTCGGGGCGGGAGGCTCGAGGCCAGCCCGGG
Gorilla          AGTGGAGTTGTATAAAGCGAGCGCGCGGCGTCGGGGCGGGAGGCTCGAGGCCAGCCCGGG
                 ************************************************************

Orangutan        ACCGGGGCTGGGAGCAAGCAGGCGGCGGCGCCGGCGGCAGAGGCGGCAGCGAGCGCCAGC
Chimpanzee       ACCGGGGCTGAGAGCAAGCAGGCGGCGGCGCCGGCGGCAGAGGCGGCAGCGAGCGCCCGC
Bonobo           ACCGGGGCTGAGAGCAAGCAGGCGGAGGCGCCGGCGGCAGAGGCGGCAGCGAGCGCCCGC
Human            ACCGGGGCTGGGAGCAAGCAGGCGGCGGCGCCGGCGGCAGAGGCGGCAGCGAGCGCCCGC
Gorilla          ACCGGGGCTGGGAGATAGCAGGCGGCGGCGCCGGCGGCAGAGGCGGCAGCGAGCTCCCGC
                 ********** ***  ********* *************************** ** **

Orangutan        TTCCCACGCCCCTAGGCGGCGGGG-CCGAGAGCGGGAGGACGGCTAAGCTT-
Chimpanzee       TCCCACGCCCCCTAAGGCGGCGGGGCCCACAGCGGGAGGACGGGCTAAGCTT
Bonobo           TTCCACGCCCCT----------------------------------------
Human            TTCCCACGCCCCTAGGCGGCGG-GGCCGAGAGCGGGAGGACGGCTAAGCTT-
Gorilla          --------------------------------------------------
```

2.2.2 Silencer Alignment

```
Human            AAGAGTTTGTGAGGCGGAGAAGGGAAGGAAGGAGCGTTCCAGGGAGAGGATACAGCGTGG
Chimpanzee       AAGAGTTTGTGAGGCGGAGAAGGGAAGGAAGGAGCGTTCCAGGGAGAGGATACAGTGTGG
                 ************************************************************ ****

Human            GCCAAGGTACAGCAGCGTGAAAGGGTAAGTTGTAGCTGAAGATCGGGAAGAAGTTCAGAG
Chimpanzee       GCAAAGGTACAGCAGCGTGAAAGGGTAAGTTGTAGCTGAAGATCGGGAAGAAGTTCAGAG
                 ** ************************************************************

Human            CGGGAAAGAGTGTCAGAATGTGTATCAGGTGGGAATGTGTATCAGAGGAGACTGGGGAGG
Chimpanzee       CGGGAAAGAGTGTCAGAATGTGAATCAGGTGGGAATGTGTATCAGAGGAGACTGGGGAGG
                 ********************** ************************************

Human            GGCTTCAGAGGCTTCACACTTACGAGGTGGGTATTTATCCCCATTCTCCAGATACTGACA
Chimpanzee       GGCTTCAGAGGCTTCACACTTACGAGGTGGGTATTTATCCCCATTCTCCAGATACCGACA
                 ************************************************************ ****

Human            TTGAGGCTTGACAAGGGGAAGTGACTTGCAGGTGTCTGAGCTAGGACTGGAAACCAGTCT
Chimpanzee       TTGAGGCTTGACAAGGGGAAGTGACTTGCAGGTGTCTGAGCTAGGACTGGAAACCAGTCT
                 ************************************************************

Human            CCTGTCCTTAGGCGCAGCTCATCCCTCTGCCCAGGAAATCTGCGTAACATCTTAAGTCTT
Chimpanzee       CCTGTCCTTAGGCGCAGCTCATCCCTCTGCCCAGGAAATCTGCATAACATCTTAAGTCTT
                 ****************************************** ****************

Human            TCCGGGGCTGCACGGGCTGAGGGGGACTGGCTCCCAGACACCGCACACCAGCAGCCTCGA
Chimpanzee       CCCGGGGCTGCGCGGGGTGAGGGGGACTGGCTCCCAGACACCGCACACCAGCAGCCTCGA
                  ********** **** ******************************************

Human            AGCCAGCTGGTCCTGCCAGGCCTGTGTATAGGGGAGGGCTGAGCAGAAGAGCGCCCCCCA
Chimpanzee       AGTCAGCCAGTCCTGCCAGGCCTGTGTATAGGGGAGGGCTGAGCAGAAGAGCACCCCCCA
                 ** ****  ****************************************** *******

Human            TTAACCAGCAGAGACATGAGGTCAGGAGCAGCAGTGAGTCACCTCTGGCAGCTTTTAAAG
```

```
Chimpanzee      TTAACCAGCAGAGACATGAGGTCAGGAGCAGCAGTGAGTCACCTCTGGCAGCTTTTAAAG
                ************************************************************

Human           GACAGAGGCCAAGGAGGCAGAGAGAATTGCACTTTTCAGAAGAATTACAAACAAAACAGA
Chimpanzee      GACAGAGGCCAAGGAGGCAGAGAGAATTGCACTTTTCAGAAGAATTACAAACAAAACAGA
                ************************************************************

Human           TGGTGGGAATACAGTTGGTAGAATATTTTTGCACTTTAATAAAGCAATTTTTATGTGGGC
Chimpanzee      TGGTGGGAATACAGTTGGTAGAATATTTTTGCACTTTAATAAAGCAATTTTTATGTGGGC
                ************************************************************

Human           CACTGAATTTCCCTTCCCTAAATGAGCATCAGCCGGCTCGGAGAGGCAGCTCTGAGTCAC
Chimpanzee      CACTGAATTTCCCTTCCCTAAATGAGCATCAGCTGGCTCGGAGAGGCAGCTCTGAGTCAC
                ********************************** *************************

Human           CTGCAAGCAATTAGCTGAAAGGC-GGTGGCATGTGTGGGTGGGGCTGGGGCACCAGCAAT
Chimpanzee      CTGCAAGCAATTAGCTGAAAGGACGGTGGCATGTGTGGGTGGGGCTGGGGCACCGGCAAT
                ********************** ***************************** *****

Human           ACCAGGGGCAGTGGAGGCGGGAGGGAAGGGGGGGAGGGGGAAGGGGGGAGGGAGAATGGT
Chimpanzee      ACCAGGGGCAGCCGGG--GGGAGGGAAGACGGGGAGGGGGAAGAGGGGAGGGATAATGGT
                *********** * *  **********  ************* ********* ******

Human           TGGGGTGGCTGCCTGAGCCAGTGACTCCCAGTGGAGCTGGCGGAGGCAAATAGATGATTG
Chimpanzee      TGGGGTGGCTGCCTGAGCCAGTGACTCCCAGTGGAGCTGGCGGAGGCAAATAGATGATTG
                ************************************************************

Human           GTGTTGTTTTTATTTTCCGGGTTTGGGAGTGAGCTGCCTCAGCCCGCTGCCAGCTGGGAA
Chimpanzee      GTGTTGTTTTTATTTTCCGGGTTTGGGAGTGAGCTGCCTCGGCCCGCTGCCAGCTGGGAA
                ************************************** *******************

Human           GTGGAGGGTGGGCATGTGGGGCTTGGGGGTGGGTTCCGAGCAACCAGGTTGGCGGTGGGG
Chimpanzee      GTGGAGGGTGGGCATGTGGGGCTTGGGGGTGGGTTCCGAGCAACCAGGCTGGCAGTGGGG
                ********************************************** **** *******

Human           GAGCCCTCGACCAGGGCCATAAGCAGAAATGCTTATTTTCAGCTGGAGGTCCGTCTGCAC
Chimpanzee      GAGCCCTCGACAAGGGCTATAAGCAGAAATGCTTATTTTCAGCTGGAGGTCCGTCTGCAC
                *********** ***** ******************************************

Human           CGAGGAGCAGCAGGGACAGGCCAGGGTGAAGAGGGGCTGTGGAACTTGGTGGGGGGGGGC
Chimpanzee      CGAGGAGCAGCAGGGACAGGCCAGGGTGAAGAGGGGCTGTGGAACTTGGTGGGG---GGG
                ******************************************************   **

Human           GGTCCACTGAGCAAGGCAGGGTCCACTGAGAAGGGACCCCACAGCAAGATGTCGCTATGA
Chimpanzee      GGTCCACTGAGCAAGGCAGGGTCCACTGAGAAGGGACCCCACAGCAAGATGTCGCTATGA
                ************************************************************

Human           GAGCTCTGCACAACACAGTGGGCCCTCTGGGAACCCCGCATGCACTCATTACCCATGGAC
Chimpanzee      GAGCTCTGCACAACACAGTGGGCCCTCTGGGAAACCCCCATGCACTCATTACCCATGGAC
                ********************************** *** **********************

Human           GGAGGTGGGGAGTACAGGTGAGGGCCTGCGTCCTCTGGGTGTGCTCCCGCGGGAGCAGAC
Chimpanzee      GGAGGTGGGGAGTACAGGTGAGGGCCTGCGTCCTCTGGGTGTGCTCCCGCGGGAGCAGAC
                ************************************************************

Human           ACGATGTACATGGGGCAAGGCTGAGGCTGTCTGAGAGATGCACACGTGTACACACTCACA
Chimpanzee      ACGATGTACATGGGGCAAGGCTGAGGCTGTCTGAGAGAGGCACACGTGTACACACTCACA
                ************************************ ********************

Human           TACATGTCCGTGTACTCACACATGCGCACTGCACCTTGCTGCAAAGTCATGGGGAGGCAG
Chimpanzee      TACATGTCCGTGTACTCACACATGCGCACCGCACCTTGCTGCAAAGTCATAGGGAGGCAG
                ***************************** ******************** *********

Human           CGCGTGACTCATGAAGGCAGGAAGGGGGCAGGGGCCCTGGGGGCAGGCAAGCTGAAGCCT
Chimpanzee      CGCGTGACTCATGGAGGCAGGAAGGGGGCAGGGGCCCTGGGGGCAGGCAAGCTGAAGCCT
                ************* ************************************************
```

```
Human           TGTGAGTGCTGGCAGGGGAGCTTCCCGCCGGCCCCTCCCCTCTACTCCTCAGTCATGCGG
Chimpanzee      TGTGAGTGCTGGCAGGGGAGCTTCCCGCCGGCCCCTCCCCTCTACTCCTCAGTCATGCGG
                ************************************************************

Human           GACCTCAAGGCAGGCCTTTCTCAGATTCATGTTGTTGACATTCCCGCTTCTCACCCCAAC
Chimpanzee      GACCTCAAGGCAGGCCTTTCTCAGATTCATGTTGTTGACATTCCCGCTTCTCACCCCAAC
                ************************************************************

Human           TGCTCAGAATTGGAAACAATCTAACTGTCCTTCCAGAGGGGACTGGGTAAATAGATGATG
Chimpanzee      TGCTCAGAATTGGAAACAATCTAACTGTCCTTCCAGAGGGGACTGGGTAAATAGATGATG
                ************************************************************

Human           AGTGTGAGTGCTGTCTCACAGTGAAATACTCCACAGCACTGAAAGGAAACAAGATATATG
Chimpanzee      AGTGTGAGTGCTGCCTCACAGTGAAATACTCCACAGCACTGAAAGGAAACAAGAGATATG
                ************* ****************************************** *****

Human           CACAGACTTGGTATGCTGTCCAGGATGTGTACACACACACACACACACACACACACAC
Chimpanzee      CACAGACTTGGTATGCTGTCCAGGATGTGTACACACAC------------------AC
                ****************************************                **

Human           ACACACTCTCACACGTGTATGTAAGACAGGCTCTCACTCTGTTACCCAAGCTGGAGTACA
Chimpanzee      ACACACTCTCACACGTGTATGTAAGACAGGCTCTCACTCTGTTACCCAAGCTGGAGTACA
                ************************************************************

Human           GTGTTGTGATCATAGCTCACTACAGCCTCAAACTCCTGGCCTTGAGCGATCTTCCTGCCT
Chimpanzee      GTGTTGTGATCATAGCTCACTACAGCCTCAAACTCCTGGCCTTGAGTGATCTTCCTGCCT
                ********************************************** *************

Human           CATCCTCCTGAGTAACTGGGACTATAGGCTTGCACCACCACACCGGGCTAATTTTTGTAG
Chimpanzee      CATCCTCCTGAGTAGCTGGGACTATAGGCTCGCATCACCACACCGGGCTAATTTTTGTAG
                ************** *************** *** *************************

Human           AAAGGGGGGTCTCGCCATGTTGCCTAGGCTGGTCTTGAACTCCTGGCCTCAAGCAATCCT
Chimpanzee      AAAGGGGGGTCTTGCCATGTTGCCTAGGCTGGTCTTGAACTCCTGGCCTCAAGCAATCCT
                ************ ********************************************

Human           CCCTCCTTGGCCTCCCAAAGCATTGGGATTACAGGTATGAGCCGCTGTACCCAGCCTTGG
Chimpanzee      CCCTCCTTGGCCTCCCAAAGCATTGGGATTACAGGTATGAGCCGCTGTACCCAGCCTTGG
                ************************************************************

Human           ACATAGG
Chimpanzee      ACATAGG
                *******
```

2.2.3 Comparison of promoter sequences my_sequence vs. UCSC_sequence

Human

```
My_Human        TACTGTCCTAGACCCCTGAAGAAACAGGTTGACATTGGACCGAGCCCCAAGAGACTCAGG
UCSC_Human      TACTGTCCTAGACCCCTGAAGAAACAGGTTGACATTGGACCGAGCCCCAAGAGACTCAGG
                ************************************************************

My_Human        GTTCCTAATGATCCTGAAAGTATACTCAAGGTGTGTGTGTGTGTGTGTGTG------TGT
UCSC_Human      GTTCCTAATGATCCTGAAAGTATACTCAAGGTGTGTGTGTGTGTGTGTGTGTGTGTGT
                ****************************************************      ***

My_Human        GTGTGTGTGTGTGTAAATGTGTGTGTTTGTATATTTTTCTAGAAAAATGATTCAGTAATT
UCSC_Human      GTGTGTGTgtgtgtAAATGTGTGTGTTTGTATATTTTTCTAGAAAAATGATTCAGTAATT
                ************************************************************

My_Human        TTTGCCAGATTTTCAAAGGGGCCCACAACCTTACAGGCTTGATGGGGGCTCAAAAAGCCA
UCSC_Human      TTTGCCAGATTTTCAAAGGGGCCCACAACCTTACAGGCTTGATGGGGGCTCAAAAAGCCA
                ************************************************************
```

```
My_Human        GAAAGTTTGTCTGTGCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACA
UCSC_Human      GAAAGTTTGTCTGTGCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACA
                ************************************************************

My_Human        GAGCCCAGACTAGGTCCTGGTAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGAGGC
UCSC_Human      GAGCCCAGACTAGGTCCTGGTAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGAGGC
                ************************************************************

My_Human        TGGATGTGTGGCCCTCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAG
UCSC_Human      TGGATGTGTGGCCCTCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAG
                ************************************************************

My_Human        AAGCCAGACTAGTGGGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCT
UCSC_Human      AAGCCAGACTAGTGGGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCT
                ************************************************************

My_Human        CCTTCCCTCTCCGATGGTGGATTCCCAAGTCTTTCTCTTCTGTGGCCCCTTTACCTGGCT
UCSC_Human      CCTTCCCTCTCCGATGGTGGATTCCCAAGTCTTTCTCTTCTGTGGCCCCTTTACCTGGCT
                ************************************************************

My_Human        TTCAGAAACCCAAAGACCTCACTCATGGAAGCTCAGAATTCATTCAATAAACACTGAGCG
UCSC_Human      TTCAGAAACCCAAAGACCTCACTCATGGAAGCTCAGAATTCATTCAATAAACACTGAGCG
                ************************************************************

My_Human        CCTACTATGTGCCAGGCTCTATTCTAGGCGTAAAATAAAGGATTTGCCTTCAGAGTGTGA
UCSC_Human      CCTACTATGTGCCAGGCTCTATTCTAGGCGTAAAATAAAGGATTTGCCTTCAGAGTGTGA
                ************************************************************

My_Human        CCCACCAAGGCCAGTTTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTG
UCSC_Human      CCCACCAAGGCCAGTTTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTG
                ************************************************************

My_Human        TTTTTCTGGCGCTTAGCACCCCCATTCCCCTGGTCCAGAGAGGCAGAAACCATCCCCCCA
UCSC_Human      TTTTTCTGGCGCTTAGCACCCCCATTCCCCTGGTCCAGAGAGGCAGAAACCATCCCCCCA
                ************************************************************

My_Human        CCACCACCAAGGCAGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCAAAGCG
UCSC_Human      CCACCACCAAGGCAGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCAAAGCG
                ************************************************************

My_Human        GTGAGGGAGGCTAGTAGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTCGGCCA
UCSC_Human      GTGAGGGAGGCTAGTAGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTCGGCCA
                ************************************************************

My_Human        CTGCTGGGGTAGGGACTAGGTGGGATGGGGTGGGGGCACTGCTGGTGGGGACGGGGTGGG
UCSC_Human      CTGCTGGGGTAGGGACTAGGTGGGATGGGGTGGGGGCACTGCTGGTGGGGACGGGGTGGG
                ************************************************************

My_Human        GTAGAAGCGGCGCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCCGAGCTGCGC
UCSC_Human      GTAGAAGCGGCGCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCCGAGCTGCGC
                ************************************************************

My_Human        TCGCCCGCTCCCCGCCTGCGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGAGCAGTCC
UCSC_Human      TCGCCCGCTCCCCGCCTGCGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGAGCAGTCC
                ************************************************************

My_Human        CCGGGAGGTGCTCGGACACGTCCTCAGGCTGGATAAAGATCGGCTCGGCGCTAGCTCCGT
UCSC_Human      CCGGGAGGTGCTCGGACACGTCCcCAGGCTGGATAAAGATCGGCTCGGCtCTAGCTCCGT
                ********************** ************************ **********

My_Human        AGTCGAAATCTCGCCATCAGCGCGGCTCGCTTGGCCGCTTTGGCCCGGCCCGGCGACGCC
UCSC_Human      AGTCGAAATCTCGCCATCAGCGCGGCTCGCTcGGCCGCTTTGGCCCGGCCCGGCGACGCC
                ******************************* ****************************

My_Human        AGATCGCTATCCTGGGGGGAAATGGGAAGGCAGTGCAGCCGCTCCCGCCTCCACCCTCGG
UCSC_Human      AGATCGCTATCCTGGGGGGAAATGGGAAGGCAGTGCAGCCGCTCCCGCCTCCACCCTCGG
                ************************************************************
```

```
My_Human        TCCTGCGCGCAGGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCCCCTTCGC
UCSC_Human      TCCTGCGCGCAGGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCCCCTTCGC
                ************************************************************

My_Human        GTTCCTTCCCTGCGCTGCCTCCCGAGGGACCCTCGCTTCCCTCCGGCCTGGGGCCCCCAG
UCSC_Human      GTTCCTTCCCTGCGCTGCCTCCCGAGGGACCCTCGCTTCCCTCCGGCCTGGGGCCCCCAG
                ************************************************************

My_Human        CGCCCAGCCAGGCGCCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCCTGAAGGTTACGC
UCSC_Human      CGCCCAGCCAGGCGCCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCCTGAAGGTTACGC
                ************************************************************

My_Human        GACGCAGTGGCGGGGCGCGGGGGGCGCCCGCCCTCGCCGCCGCCTTCGCCGCCGCCCACC
UCSC_Human      GACGCAGTGGCGGGGCGCGGGGGGCGCCCGCCCTCGCCGCCGCCTTCGCCGCCGCCCACC
                ************************************************************

My_Human        CCAGTGGCCGCCGCGCCGCGCCGGGCCAGCCTGGCTGCCGGCTGCTGCCACCGCAATCCC
UCSC_Human      CCAGTGGCCGCCGCGCCGCGCCGGGCCAGCCTGGCTGCCGGCTGCTGCCACCGCAATCCC
                ************************************************************

My_Human        GGCTCCTAAATCAGCGCGGGGAGGCGCTCCCTCCCCACGCCCGGCTCTCCGGGCTCTCGG
UCSC_Human      GGCTCCTAAATCAGCGCGGGGAGGCGCTCCCTCCCCACGCCCGGCTCTCCGGGCTCTCGG
                ************************************************************

My_Human        GGCCGCGATTGGCCGCGCCGGCGCCCCCCACCCCGGGCCCCCGGCTCCAGCTGCCGCGCC
UCSC_Human      GGCCGCGATTGGCCGCGCCGGCGCCCCCCACCCCGGGCCCCCGGCTCCAGCTGCCGCGCC
                ************************************************************

My_Human        ATTGGCTGCGGGCCTCCGCCAGCCTTTACATAAGACCGGGCGCGCTCGAGTGGAGTTGTA
UCSC_Human      ATTGGCTGCGGGCCTCCGCCAGCCTTTACATAAGACCGGGCGCGCTCGAGTGGAGTTGTA
                ************************************************************

My_Human        TAAAGCGAGCGCGCGGCGTCGGGGCGGGAGGCTCGAGGCCAGCCCGGGACCGGGGCTGGG
UCSC_Human      TAAAGCGAGCGCGCGGCGTCGGGGCGGGAGGCTCGAGGCCAGCCCGGGACCGGGGCTGGG
                ************************************************************

My_Human        AGCAAGCAGGCGGCGGCGCCGGCGGCAGAGGCGGCAGCGAGCGCCCGCTTCCCACGCCCC
UCSC_Human      AGCAAGCAGGCGGCGGCGCCGGCGGCAGAGGCGGCAGCGAGCGCCCGCTTCCCACGCCCC
                ************************************************************

My_Human        TAGGCGGCGGGGCCGAGAGCGGGAGGACGGCTAAGCTT
UCSC_Human      TAGGCGGCGGGGCCGAGAGCGGGAGGAtGGCT------
                *************************** ****
```

Chimpanzee

```
My_Chimp        TACTGTCCTAGACCCCTGAAGAAACAGGTTGACATTGGACAGAGCCCCAAGAGACTCAGG
UCSC_Chimp      TACTGTCCTAGACCCCTGAAGAAACAGGTTGACATTGGACaGAGCCCCAAGAGACTCAGG
                ************************************************************

My_Chimp        GTTCCTAATGATCCTGAAAGTATACCCAAGGTGTGTGTGTGTGTGTGTGTGTGTGTAAGT
UCSC_Chimp      GTTCCTAATGATCCTGAAAGTATACTCAAGGTGTGTGTGTGTGTGTGTGT----GTGTGT
                ************************ ************************    **  **

My_Chimp        AAATGTGTGTGTTTGTATATTTTTCTAGAAAAATGATTCAGTAATTTTTGCCAGATTTTC
UCSC_Chimp      AAATGTGTGTGTTTGTATAcTTTTCTAGAAAAATGATTCAGTAATTTTTGCCAGATTTTC
                ******************* ****************************************

My_Chimp        AAAGGGGCCCACAACCTTACAGGCTTGATGGGGGCTCAAAAAGCCAGAAAGTTTGTCTGT
UCSC_Chimp      AAAGGGGCCCACAACCTTACAGGCTTGATGGGGGCTCAAAAAGCCAGAAAGTTTGTCTGT
                ************************************************************

My_Chimp        GCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGCCCAGACTAGG
UCSC_Chimp      GCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGCCCAGACTAGG
                ************************************************************
```

```
My_Chimp        TCCTGATAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGAGGCTGGATGTGTGGCCC
UCSC_Chimp      TCCTGaTAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGAGGCTGGATGTGTGGCCC
                ************************************************************

My_Chimp        TCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGCCAGACTAGTG
UCSC_Chimp      TCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGCCAGACTAGTG
                ************************************************************

My_Chimp        GGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTTCCCTCTCCGA
UCSC_Chimp      GGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTTCCCTCTCCGA
                ************************************************************

My_Chimp        TGGTGGATTCCCAAGTCTTTCTCTTCTGTGGCCCCTTTACCTGGCTTTCAGAAACCCAAA
UCSC_Chimp      TGGTGGATTCCCAAGTCTTTCTCTTCTGTGGCCCCTTTACCTGGCTTTCAGAAACCCAAA
                ************************************************************

My_Chimp        GACCTCACTCATGGAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTACTATGTGCCA
UCSC_Chimp      GACCTCACTCATGGAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTACTATGTGCCA
                ************************************************************

My_Chimp        GGCTCTATTCTAGGCGTAAAAAAAGGATTTGCCTTCAGAGTGTGACCCACCAAGGCCAGT
UCSC_Chimp      GGCTCTATTCTAGGCGTAAAAAAAGGATTTGCCTTCAGAGTGTGACCCACCAAGGCCAGT
                ************************************************************

My_Chimp        TTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTTTCTGGCGCTTA
UCSC_Chimp      TTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTTTCTGGCGCTTA
                ************************************************************

My_Chimp        GCACCCCCATTCCCCTGGTCCAGAGAGGCAGAAACCATCCCCCCACCACCACCAAGGCAG
UCSC_Chimp      GCACCCCCATTCCCCTGGTCCAGAGAGGCAGAAACCATCCCCCCACCACCACCAAGGCAG
                ************************************************************

My_Chimp        CGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCAAAGCGGTGAGGGAGGCTAGT
UCSC_Chimp      CGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCAAAGCGGTGAGGGAGGCTAGT
                ************************************************************

My_Chimp        AGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTCGGCCACTGCTGGGGTAGGGA
UCSC_Chimp      AGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTCGGCCACTGCTGGGGTAGGGA
                ************************************************************

My_Chimp        CTAGGTGGGACGGGGTGGGGGCACTGCTGGTGGGGACAGGGGGTGGGGTAGAAGCGGCGC
UCSC_Chimp      CTAGGTGGGAcGGGGTGGGGGCACTGCTGGTGGGGACagGGGGTGGGGTAGAAGCGGCGC
                ************************************************************

My_Chimp        TGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCCGAGCTGCGCTCGCCCGCTCCCC
UCSC_Chimp      TGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCCGAGCTGCGCTCGCCCGCTCCCC
                ************************************************************

My_Chimp        GCCTGCGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGAGCAGTCCCCGGGAGGTGCTC
UCSC_Chimp      GCCTGCGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGAGCAGTCCCCGGGAGGTGCTC
                ************************************************************

My_Chimp        GGACACGTCCCCAGGCTGGATAAAGATCGGCTCGGCGCTAGCTCCGTAGTCGAAATCTCG
UCSC_Chimp      GcACACGTCCcCAGGCTGGATAAAGATCGGCTCGGCGCTAGCTCCGTAGTCGA-ATCTCG
                * ********************************************** *******

My_Chimp        CCATCAGCGCGGCTCGCTCGGCCGCTTTGGCCCGGCCCGGCGACGCCAGATCGCTATCCT
UCSC_Chimp      CCATCAGCGCGGCTCGCTcGGCCGCTTTGGCCCGGCCCGGCGACGCCAGATCGCTATCCT
                ************************************************************

My_Chimp        GGGGGGAAATGGGAAGGGAGTGCCGCCGCTCCCGCCTCCACCCTCGGTCCTGCGCGCATG
UCSC_Chimp      GGGGGGAAATGGGAAGGCAGTGCcGCCGCTCCCGCCTCCACCCTCGGTCCTGCGCGCAtG
                ***************** ****************************************

My_Chimp        GGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCCCCTTCGCGTTCCTTCCCTGC
UCSC_Chimp      GGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCCCCTTCGCGTTCCTTCCCTGC
```

```
                ************************************************************

My_Chimp        GCTGCCTCCCGAGGGACCCTCGCTC-CCTCCGGCCTGGGGCCCCCAGCGCCCAGCCAGGC
UCSC_Chimp      GCTGCCTCCCGAGGGACCCTCGCTTCCCTCCGGCCTGGGGCCCCCAGCGCCCAGCCAGGC
                ************************  *********************************

My_Chimp        GCCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCCTGAAGGTTACGCGACGCAGTGGCGG
UCSC_Chimp      GCCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCCTGAAGGTTACGCGACGCAGTGGCGG
                ************************************************************

My_Chimp        GGCGCGGGGGGCGCCCGCCCTCGCCGCCGCCTTCGCCGCCGCCCACCCCAGTGGCCGCCG
UCSC_Chimp      GGCGCGGGGGGnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
                ***********

My_Chimp        CGCCGCGCCGGGCCAGCCTGGCTGCCGGCTGCTGCCACCGCAATCCCGGCTCCTAAATCA
UCSC_Chimp      nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn-------nnnnnn--

My_Chimp        GCGCGGGGAGGCGCTCCCTCCCCCAGCCCGGCTCTGGGGGCTCTCCGGGCAGCGATTGGC
UCSC_Chimp      ------------------------------------------------------------

My_Chimp        CGCGCCGGCGCCCCCCACCCCGGGCCCCCGGCTCCAGCTGCCGCGCCATTGGCTGCGGGC
UCSC_Chimp      ------------------------------------------------------------

My_Chimp        CTCCGCCAGCCTTTACATAAGACCGGGCGCGCTCGAGTGGAGTTGTATAAAGCGAGCGCG
UCSC_Chimp      ------------------------------------------------------------

My_Chimp        CGGCGTCGGGGCGGGAGGCTCGAGGCCAGCCCGGGACCGGGGCTGAGAGCAAGCAGGCGG
UCSC_Chimp      ------------------------------------------------------------

My_Chimp        CGGCGCCGGCGGCAGAGGCGGCAGCGAGCGCCCGCTCCCACGCCCCCTAAGGCGGCGGGG
UCSC_Chimp      ------------------------------------------------------------

My_Chimp        CCCACAGCGGGAGGACGGGCTAAGCTT
UCSC_Chimp      ---------------------------
```

Gorilla

```
My_Gorilla        TACTGTTCTAGACCCCTGAAGAAACAGGTTGACACTGGACAGAGCCCCCAGAGACTCAGG
UCSC_Gorilla      TACTGTtCTAGACCCCTGAAGAAACAGGTTGACAcTGGACaGAGCCCCgAGAGACTCAGG
                  ********************************************** ***********

My_Gorilla        GTTCCTAATGATCCTGAAAGTATATGCAGGGG-------TGTGTGTGTGTGTGTGTGTGT
UCSC_Gorilla      GTTCCTAATGATCCTGAAAGTATAtgcaaggGGTGTGTGTGTGTGTGTGTGTGTGTGT
                  **************************** ***       **********************

My_Gorilla        GTGTGTGTGTAAATGTGTGTGTTTGTATATTTTTCTATGAACATGATTCAGTAATTTTTG
UCSC_Gorilla      GTGTGTGTGTAAATGTGTGTGTTTGTATATTTTTCTAGAAAAATGATTCAGTAATTTTTG
                  ************************************  ** ******************

My_Gorilla        CCAGATTTTCAAAGGGGCCCACAACCTTAAAGGCTTGATGGGGGCTCAAAAAGCCAGAAA
UCSC_Gorilla      CCAGATTTTCAAAGGGGCCCACAACCTTAaAGGCTTGATGGGGGCTCAAAAAGCCAGAAA
                  ************************************************************

My_Gorilla        GTGTGTCTGTGCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGC
UCSC_Gorilla      GTgTGTCTGTGCCTAGCTATTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGC
                  ************************************************************

My_Gorilla        CCAGAGTAGGTCCTGGTAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGGGGCTGGA
```

```
UCSC_Gorilla      CCAGAgTAGGTCCTGGTAGTTCTGCTCAAAATCTGTGGATAACAGGGGCCAGgGGCTGGA
                  ************************************************************

My_Gorilla        TGTGTGGCCCTCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGC
UCSC_Gorilla      TGTGTGGCCCTCTATAAAAGAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGC
                  ************************************************************

My_Gorilla        CAGACTAGTGGGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTT
UCSC_Gorilla      CAGACTAGTGGGTGATGGTCACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTT
                  ************************************************************

My_Gorilla        CCCTCTCCGATGGTGGATTCCCAAGTCTTTCTCTTCTGTAGCCCCTTTACCTGGCTTTCA
UCSC_Gorilla      CCCTCTCCGATGGTGGATTCCCAAGTCTTTCTCTTCTGTaGCCCCTTTACCTGGCTTTCA
                  ************************************************************

My_Gorilla        GAAACCCAAAGACCTCACTCATGAAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTA
UCSC_Gorilla      GAAACCCAAAGACCTCACTCATGaAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTA
                  ************************************************************

My_Gorilla        CTATGAGCCAGGCTCTATTCTAGGCGTAAAATAAAGGATTTGCCTTCAGAGTGTGACCCA
UCSC_Gorilla      CTATGaGCCAGGCTCTATTCTAGGCGTAAAATAAAGGATTTGCCTTCAGAGTGTGACCCA
                  ************************************************************

My_Gorilla        CCAAGGCCAGTTTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTT
UCSC_Gorilla      CCAAGGCCAGTTTCCTGCCTGAACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTT
                  ************************************************************

My_Gorilla        TCTGGCGCTTAGCACCCCCATTTCCCTGGTCCAGAGAGGCAGAAACCATCCCCCCCACCA
UCSC_Gorilla      TCTGGCGCTTAGCACCCCCATTtCCCTGGTCCAGAGAGGCAGAAACCATcCCCCCCACCA
                  ************************************************************

My_Gorilla        CCACCAAGGCAGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCAAAGCGGTG
UCSC_Gorilla      CCACCAAGGCAGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCAAAGCGGTG
                  ************************************************************

My_Gorilla        AGGGAGGCTAATAGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTCGGCCATTG
UCSC_Gorilla      AGGGAGGCTAGTAGGACGCAGGCGGCAGGAGCGGGTGGGCCAGGCCCGACTCGGCCAtTG
                  ********** ***************************************************

My_Gorilla        CTGGGGTAGGGACTAGGTGGGACGGGGTGGGGGCACTGCTGGTGGGGACAGGGGGTGGGG
UCSC_Gorilla      CTGGGGTAGGGACTAGGTGGGAcGGGGTGGGGnnnnnnnn----------------------
                  ********************************

My_Gorilla        TAGAAGCGGCGCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCTTGATCTCGGAGCTGCGCT
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        CGCCCGCTCCCCGCCTGCGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGAGCAGTCCC
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        CGGGAGGTGCTCGGACACGTCCCCAGGCTGGATAAAGATCGGCTCGGCGCTAGCTCCGTA
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        GTCGAAATCTCGCCATCAGCGCGGCTCGCTCGGGCGATTTGGCCCGGCTCGGCGACGCCA
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        GATCCCTATCCTGGGGGGAAATGGAAAGGCAGTGCCGCCGCTACCGCCTCCACCCTCGGT
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        CCTGCGCGCAGGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCCCCTTGGCG
UCSC_Gorilla      ------------------------------------------------------------
```

```
My_Gorilla        TTCCTTCCCTGCGCAGCATCCCGAGGGACCCTCGCTTCCCTCCGGCCTGGGGCCCCCAGC
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        GCCCAGCCAGGCGCCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCCTGAGGTTACGCGA
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        CGCAGTGTGCGGGGCGCGGGGGGCGCCCGCCCTCGCCGCCGCCCACCCCAGTGCCCGCCG
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        CGCCGCGCCGCGCCGGGCCAGCCTGGCTGCCGGCTGCTGCCACCGCAATCCCGGCTCCTA
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        AATCAGCGCGGGGAGGCGCTCCCTCCCCCCGCCCTGCTCTCCGGGCTCTCCGGGCCGCGA
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        TTGGCCGCGCCGGCGCCCCCCACCCCGGGCCCCAGGCTCCAGCTGCCGCGCCATTGGCTG
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        CGGGCCTCCGCCAGCCTTTACATAAGACCGGGCGCGCTCGAGTGGAGTTGTATAAAGCGA
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        GCGCGCGGCGTCGGGGCGGGAGGCTCGAGGCCAGCCCGGGACCGGGGCTGGGAGATAGCA
UCSC_Gorilla      ------------------------------------------------------------

My_Gorilla        GGCGGCGGCGCCGGCGGCAGAGGCGGCAGCGAGCTCCCGC
UCSC_Gorilla      ----------------------------------------
```

Orangutan

```
My_Orangutan        TACTGTCCTAGACCCCTGAGGAAACAGGTTGACATTGGACAGAGCCCCAAGAGACTCAGG
UCSC_Orangutan      TACTGTCCTAGACCCCTGAgGAAACAGGTTGACATTGGACaGAGCCCCAAGAGACTCAGG
                    ************************************************************

My_Orangutan        GTTCCTAATGATCTGAAAGTATACTCAAGGTGGTGTGTGTGTGTGTGTGTAAATGTGTGT
UCSC_Orangutan      GTTCCTAATGATCTGAAAGTATACTCAAGGTGGTGTGTGTGTGTGTGTGTAAATGTGTGT
                    ************************************************************

My_Orangutan        GTTTGTATATTTTTCTAGAAAAATGATTCAGTAATTTTTGCCAGATTTTCAAAGGGGCCC
UCSC_Orangutan      GTTTGTATATTTTTCTAGAAAAATGATTCAGTAATTTTTGCCAGATTTTCAAAGGGGCCC
                    ************************************************************

My_Orangutan        ACAACCCTTAAAGGCTTGATGGGGGCTCAAAAAGCCAGAAAGTTTGTCTGTGCCTAGCTA
UCSC_Orangutan      ACAACCcTTAaAGGCTTGATGGGGGCTCAAAAAGCCAGAAAGTTTGTCTGTGCCTAGCTA
                    ************************************************************

My_Orangutan        TTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGCCCAGACTAGGTCCTGGTAG
UCSC_Orangutan      TTTCTTTTCAGAGGAAGATCACAGAGACCTGAAAACAGAGCCCAGAgTAGGTCCTGGTAG
                    ******************************************** *************

My_Orangutan        TTCTGCTCAAAATCTCTGGGTAATAGGGGCCAGGGGCTGGATGTGTGGCCCTCTATAAAA
UCSC_Orangutan      TTCTGCTCAAAATCTcTGGgTAAtAGGGGCCAGgGGCTGGATGTGTGGCCCTCTATAAAA
                    ************************************************************

My_Orangutan        GAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGCCAGACTAGTGGGTGATGGT
UCSC_Orangutan      GAACACATTCAAATCATTCTAAGAGTCAAGGGCCCAGAAGCCAGACTAGTGGGTGATGGT
                    ************************************************************
```

```
My_Orangutan        CACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTTCCCTCTCCGATGGTGGATT
UCSC_Orangutan      CACAAAAGCGTCAACGGATGGGAAAACAGTCCTAGCTCCTTCCCTCTCCGATGGTGGATT
                    ************************************************************

My_Orangutan        CCCAAGTCTTTCTCTTCTGTAGCCCCTTTACCCTGCTTTCAGAAACCCAAAGACCTCACT
UCSC_Orangutan      CCCAAGTCTTTCTCTTCTGTaGCCCCTTTACCctGCTTTCAGAAACCCAAAGACCTCACT
                    ************************************************************

My_Orangutan        CACAGAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTACTATGTGCCAGGCTCTATT
UCSC_Orangutan      CAcaGAAGCTCAGAATTCATTCAATAAACACTGAGCGCCTACTATGTGCCAGGCTCTATT
                    ************************************************************

My_Orangutan        TTAGGCGTAAAATAAAGGATTTGCCTTCAGAGTGTGACCCACCAAGGCCAGCTTCCTGCC
UCSC_Orangutan      tTAGGCGTAAAATAAAGGATTTGCCTTCAGAGTGTGACCCACCAAGGCCAGcTTCCTGCC
                    ************************************************************

My_Orangutan        TGGACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTTTCTGGCGCTTAGCACCCCC
UCSC_Orangutan      TGgACAGGGGTATTTTGAGCATTGCTGATCAAGTGTGTTTTTCTGGCGCTTAGCACCCCC
                    ************************************************************

My_Orangutan        ATTCCCCTGGTCCAGAGAGGCAGAAGCCATCCCCCCCCCCACCACCACCACCACCAAGGC
UCSC_Orangutan      ATTCCCCTGGTCCAGAGAGGCAGAAgCCATCCCCCcccccaccacCACCACCACCAAGGC
                    ************************************************************

My_Orangutan        AGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCAAAGCGGTGAGGGAGGCTA
UCSC_Orangutan      AGCGGGTTGGGGGCGAGGAGAGGCGGGCAGTACCTGTGATCAAAGCGGTGAGGGAGGCTA
                    ************************************************************

My_Orangutan        GTAGGACGCAGGCGGCAGGGGCGGGAGGGCCAGGCCCGACTCGGCCACTGCTGGGGTAGG
UCSC_Orangutan      GTAGGACGCAGGCGGCAGGgGCGGGaGGGCCAGGCCCGACTCGGCCACTGCTGGGGTAGG
                    ************************************************************

My_Orangutan        GACTAGGTGGGACGGGGTGGGGGCACTGCTGGTGGGGACAGGGGGTGGGGTAGAAGCGGC
UCSC_Orangutan      GACTAGGTGGGAcGGGGTGGGGGCACTGCTGGTGGGGACagGGGGTGGGGTAGAAGCGGC
                    ************************************************************

My_Orangutan        GCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCGGAGCTGCGCTTGCCCGCTCC
UCSC_Orangutan      GCTGCCCGCAGCCGCCTGGGCCTCTGCGCGCCTGATCTCgGAGCTGCGCTtGCCCGCTCC
                    ************************************************************

My_Orangutan        CCGCCTACGGGGGCCGTCGCGGAGCCCTGCTCTCCATACTGAGCAGTCCCCAGGAGGTGC
UCSC_Orangutan      CCGCCTaCGGGGGCCaTCGCGGAGCCCTGCTCTCCATACTGAGCAGTCCCCaGGAGGTGC
                    *************** ********************************************

My_Orangutan        TCGGACACGTCCCCAGGCTGGATAAAGATCGGCTCGGCGCTAGCTCCGTAGTCGAAATCT
UCSC_Orangutan      TCGGACACGTCCcCAGGCTGGATAAAGATCGGCTCGGCGCTAGCTCCGTAGTCGAAATCT
                    ************************************************************

My_Orangutan        CGCCATCAGCGCGGCTCGCGCGGCCGCTTTGGCCCGGCCCGGCGACGCCAGATCGCTATC
UCSC_Orangutan      CGCCATCAGCGCGGCTCGCgcGGCCGCTTTGGCCCGGCCCGGCGACGCCAGATCGCTATC
                    ************************************************************

My_Orangutan        CTGGAGTGAAATGGGAAGGCAGTGCCGCCGCTCCCGCCACCACCCTCGGTCCTGCGCGCA
UCSC_Orangutan      CTGGaGtGAAATGGGAAGGCAGTGCcGCCGCTCCCGCCaCCACCCTCGGTCCTGCGCGCA
                    ************************************************************

My_Orangutan        GGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCCCCTTCGCGTTCCTTCCCT
UCSC_Orangutan      GGGGTGGCCGCGGGGTCCTGGGGCTCGCCGCCCTCCCCTCCCCCTTCGCGTTCCTTCCCT
                    ************************************************************

My_Orangutan        GCGCTGCCTCCCGAGGGACCCCCGCTTCCCTCCGGCCTGGGGCCCCTAGCGCCCAGCCAG
UCSC_Orangutan      GCGCTGCCTCCCGAGGGACCCcCGCTTCCCTCCGGCCTGGGGCCCCtAGCGCCCAGCCAG
                    ************************************************************

My_Orangutan        GCTAGGGAGCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCCTGAAGGTTACGCGACGCA
UCSC_Orangutan      GCtagggagCCTCTCCCCTCCTCGCCAGGCCTCGCTGCCGCCTGAAGGTTACGCGACGCA
                    ************************************************************
```

```
My_Orangutan        GTGGCGGGGCGCGGGAGGCGCCCGCCCTCGCCGCCGCCCACCCCAGTGCCCGCCGCGCCG
UCSC_Orangutan      GTGGCGGGGCGCGGGaGGCGCCCGCCcTCGCCGCCGCCCACCCCAGTGcCCGCCGCGCCG
                    ************************************************************

My_Orangutan        CGCCGGGCCAGCCTGGCTGCCGGCTGCTGCCACCGCAATCCCGGCTCCTAAATCAGCGCC
UCSC_Orangutan      CGCCGGGCCAGCCTGGCTGCCGGCTGCTGCCACCGCAATCCCGGCTCCTAAATCAGCGCc
                    ************************************************************

My_Orangutan        GGGAGGCGCTCCCTCCCCCCGCCCGGCTCTCCGGGCTCCCCGGGCCGCGATTGGCCGCGC
UCSC_Orangutan      GGGAGGCGCTCCCTCCCCcCGCCCGGCTCTCCGGGCTCcCcGGGCCGCGATTGGCCGCGC
                    ************************************************************

My_Orangutan        CGGCGCCCCCCACCCCGGGCCCCCGGCTCCAGCTGCCGCGCCATTGGCTGCAGGCCTCCG
UCSC_Orangutan      CGGCGCCCCCCACCCCGGGCCCCCGGCTCCAGCTGCCGCGCCATTGGCTGCaGGCCTCCG
                    ************************************************************

My_Orangutan        CCAGCCTTTACATAAGACCGGGCGCGCTCGAGTGGAGTTGTATAAAGCGAGCGCGCGGCG
UCSC_Orangutan      CCAGCCTTTACATAAGACCGGGCGCGCTCGAGTGGAGTTGTATAAAGCGAGCGCGCGGCG
                    ************************************************************

My_Orangutan        TCGGGGCGGGAGGCTCGAGGCCAGCCCGGGACCGGGGCTGGGAGCAAGCAGGCGGCGGCG
UCSC_Orangutan      TCGGGGCGGGAGGCTCGAGGCCAGCCCGGGACCGGGGCTGGGAGCtAGCAGGCGGCGGCG
                    ******************************************** **************

My_Orangutan        CCGGCGGCAGAGGCGGCAGCGAGCGCCAGCTTCCCACGCCCCTAGGCGGCGGGGCCGAGA
UCSC_Orangutan      CCGGCGGCAGAGGCGGCAGCtAGCGCCCGCTTCCCACGCCCCTAGGCGGCGGGGCCGAGA
                    ******************** ****** ********************************

My_Orangutan        GCGGGAGGACGGCTAAGCTT
UCSC_Orangutan      GCGGGAGGAtGGCT------
                    ********* ****
```

2.2.4 Genotyping data for the 'GT' repeat in CRTAC1 promoter

Human			**Bin Version**	
human1	210.87	215.41	211	215
human2	217.72	228.1	218	228
human3	219.88	219.88	220	220
human4	206.46	222.15	206	222
human5	219.77	228.5	220	228
human6	219.81	219.81	220	220
human7	206.36	219.8	206	220
human8	219.85	219.85	220	220
human9	206.5	222.26	206	222
human10	206.5	221.9	206	222
human11	206.31	221.93	206	222
Chimpanzee				
chimp1	206.34	212.84	206	213
chimp2	206.15	212.72	206	213
chimp3	212.71	212.71	213	213
chimp4	206.24	219.63	206	220
chimp5	195.54	219.73	196	220
chimp6	206.25	219.66	206	220
chimp7	206.37	206.37	206	206
chimp8	206.37	213.05	206	213
chimp9	206.29	221.89	206	222
chimp10	206.27	219.64	206	220
Gorilla				
gorilla1	206.19	221.82	206	222
gorilla2	206.2	221.89	206	222
gorilla3	206.23	221.88	206	222
gorilla4	221.79	221.79	222	222
gorilla5	206.13	221.76	206	222
gorilla6	206.4	222.26	206	222
gorilla7	222.24	222.24	222	222
gorilla8	206.47	222.27	206	222
gorilla9	206.42	222.15	206	222
gorilla10	206.21	221.79	206	222

Human Long PGL4.10 Clone			221.81		222
Human Topo Clone					222
Bonobo Topo clone					196
Chimp Topo clone					213

Bonobo

bonobo1	193.16	210.91	193	211
bonobo2	195.78	210.74	196	211
bonobo3	189.11	206.64	189	206
bonobo5	210.59	210.59	211	211
bonobo6	206.82	215.14	207	215
bonobo7	195.74	210.62	196	211
bonobo8	195.69	210.64	196	211
bonobo9	196.34	210.6	196	211

*The samples used for cloning are highlighted in red

2.2.5 Promoter only transfection raw data

Batch 1

7th September 2012

Human	Fold higher than empty	Chimp	Fold higher than empty	Bonobo	Fold higher than empty
14720241	85.83	4643527	26.98	18806674	109.28
15762741	91.59	4396725	25.55	11109015	64.55
8962319	52.07	1158855	6.73	20086664	116.72

Orang	Fold higher than empty	Empty
1059927	6.16	160419
1310769	7.62	175004
1328816	7.72	180862
		Avg.172095

20th September 2012

Human	Fold higher than empty	Chimp	Fold higher than empty	Bonobo	Fold higher than empty
608070	229.63	80563	30.42	194680	73.52
244185	92.21	63765	24.08	105431	39.81
293532	110.85	47714	18.01	239290	90.37

Orang	Fold higher than empty	Empty
17434	6.58	2648
18240	6.88	
44214	16.7	

4th October 2012

Human	Fold higher than empty	Chimp	Fold higher than empty	Bonobo	Fold higher than empty
25528616	72.68	3777247	10.75	12348091	35.16
27751664	79.01	3716311	10.58	13636062	38.82
32944918	93.79	5049206	14.38	14113665	40.18

Orang	Fold higher than empty	Empty
1575096	4.48	341288
1674156	4.77	381375
1840950	5.24	331058
		Avg. 351240

8th November 2012

Human	Fold higher than empty	Chimp	Fold higher than empty	Bonobo	Fold higher than empty
30478378	200.22	8264132	54.29	16030346	105.31
30430124	199.91	8946729	58.77	20806012	136.68
27377302	179.85	10079410	66.21	23014244	151.18

Empty
173356
141060
142262

Avg Empty: 152226

Batch 2

31st October 2012

Human	Fold higher than empty	Chimp	Fold higher than empty	Bonobo	Fold higher than empty

1629845	2.82	3173665	5.5	2779315	4.82
1438648	2.49	3152543	5.46	1743590	3.02
1473757	2.55	9036864	15.66	1767136	3.06

Empty

713419
576793
440515

Avg.
576909

8th November 2012

Human	Fold higher than empty	Chimp	Fold higher than empty	Bonobo	Fold higher than empty
36937952	242.65	13385293	87.93	18961358	124.56
35086340	230.49	19502056	128.11	15540892	102.09
37162000	244.12	17583874	115.51	21104806	138.64

Empty

173356
141060
142262

Avg Empty: 152226

13th December 2012

Human	Fold higher	Chimp	fold higher	Bonobo	fold higher
3651927	59.24	2341592	37.98	3940274	63.92
3040144	49.32	2205333	35.78	3443331	55.86
2934332	47.6	2092068	33.94	4903805	79.55

Empty

65023
55438
64467

Avg. 61642.7

2.2.6 Androgen stimulation data with promoter only constructs

20th December 2012

Uninduced:

Human	fold higher	Chimp	fold higher	Bonobo	fold higher
12947427	45.26	6939556	24.26	6140085	21.46
12625415	44.13	9937427	34.74	9698301	33.9
12158168	42.5	8867089	30.99	10964034	38.33

Empty

215971
356163
Avg.
286067

Induced:

Human	fold higher	Chimp	fold higher	Bonobo	fold higher
12699198	47.41	8285322	30.93	8073039	30.14
11733022	43.8	8812618	32.9	10155667	37.91
11811277	44.09	8960694	33.45	8590772	32.07

Empty

270722
265019
Avg.
267871

27th December 2012

Uninduced:

Human	fold higher	Chimp	fold higher	Bonobo	fold higher
8921072	32.03510511	5431611	19.50463232	10492568	37.67826543
8345890	29.96965649	5751896	20.65475908	12586352	45.19693477
7925249	28.45915656	5723593	20.55312448	13983686	50.21468841

Empty

271765
233315
330354

Avg. 278478

Induced:

Human	fold higher	Chimp	fold higher	Bonobo	fold higher
5017905	16.83719487	3271439	10.97706233	9204229	30.88408355
3499258	11.74149149	3895840	13.07219193	9520228	31.94439393
3216359	10.79224562	3424356	11.49016358	9692046	32.52091603

Empty

291522
291677
310876

Avg. 298025

3rd January 2013

Uninduced

Human	fold higher	Chimp	fold higher	Bonobo	fold higher
14906595	93.14118706	14319341	89.47183569	10592456	66.18506276
15082548	94.24059784	14452572	90.30430572	13320503	83.23077548
12407246	77.5244528	14208398	88.77862824	11481454	71.73980743

Empty

146439
185499
148192

Avg.
160043

Induced

Human	fold higher	Chimp	fold higher	Bonobo	fold higher
12893792	83.61895497	9815200	63.65363788	8396101	54.45048218
12860521	83.40318554	14237070	92.33039553	11438538	74.18132648
10782164	69.92460294	10662803	69.15052174	11495443	74.55036739

Empty

113735
186222
162634

Avg.
154197

2.2.7 Promoter + Silencer transfection data

11th April 2013

No Androgen

Human Promoter only	Fold Higher	Chimp Promoter only	Fold Higher
6685598	101.04	1423809	21.52
5795329	87.59	1273854	19.25
6121600	92.52	1321899	19.98

Human Promoter + Enhancer	Fold Higher	Chimp Promoter + Enhancer	Fold Higher
1520824	22.99	287695	4.35
1426797	21.56	322751	4.88
2330341	35.22	309616	4.68

Empty

66392
75240
56861

Avg. 66164

Androgen

Human Promoter only	Fold Higher	Chimp Promoter only	Fold Higher
9268027	77.7	3917323	32.84
12868663	107.89	4408154	36.96
12838287	107.63	4508576	37.79

Human Promoter + Enhancer	Fold Higher	Chimp Promoter + Enhancer	Fold Higher
2212239	18.55	177373	1.49
2389463	20.03	238121	1.99
3006833	25.21	260100	2.18

Empty

111490
132260
114077

Avg. 119275

18th April 2013

No Androgen

Human Promoter only	Fold Higher	Chimp Promoter only	Fold Higher
1565460	113.28	141829	10.26
1559010	112.82	119311	8.83
2592900	187.63	75304	5.45

Human Promoter + Enhancer	Fold Higher	Chimp Promoter + Enhancer	Fold Higher
92585	6.69	91214	6.6
127269	9.21	29381	2.13
96459	6.98	63536	4.61

Empty

14342
14426
12690

Avg. 13819

Androgen

Human Promoter only	Fold Higher	Chimp Promoter only	Fold Higher
6229935	117.97	508718	9.63
6825779	129.25	445944	8.44
3950515	74.81	335040	6.34

Human Promoter + Enhancer	Fold Higher	Chimp Promoter + Enhancer	Fold Higher
694025	13.14	112894	2.14
387915	7.34	65000	1.23
397452	7.52	68535	1.31

Empty

45320
53595
59515

Avg. 52810

25th April 2013

No Androgen

Human Promoter only	Fold Higher	Chimp Promoter only	Fold Higher

9129273	237.61		280840	7.31
6657146	173.28		290836	7.57
7458768	194.14		258117	6.72

Human Promoter + Enhancer	Fold Higher	Chimp Promoter + Enhancer	Fold Higher
196337	5.11	53739	1.39
168900	4.39	101657	2.65
257270	6.69	52814	1.37

Empty

33186
43652

Avg. 38419

Androgen

Human Promoter only	Fold Higher	Chimp Promoter only	Fold Higher
8346915	345.52	220633	9.13
5156599	213.46	321998	13.33
3344553	138.45	239004	9.89

Human Promoter + Enhancer	Fold Higher	Chimp Promoter + Enhancer	Fold Higher
519602	21.51	98079	4.06
454936	18.83	83620	3.46
287447	11.89	82999	3.44

Empty

37769
19111
15590

Avg. 24157

4th April 2013 ** (Considered an outlier and not used for data analysis)

No Androgen

Human Promoter only	Fold Higher	Chimp Promoter only	Fold Higher
1080220	22.35	694673	14.37
1158632	23.97	847191	17.53
1305697	27.01	860444	17.81

Human Promoter +Enhancer	Fold Higher	Chimp Promoter +Enhancer	Fold Higher
1906820	39.45	457185	9.46
2183134	45.17	694793	14.38
1714646	35.48	773558	16.01

Empty

31704
66415
46869

Avg. 48329

Androgen

Human Promoter only	Fold Higher	Chimp Promoter only	Fold Higher
2050823	27.34	1581997	21.09
1978190	26.37	1544078	20.59
1833835	24.45	1683025	22.44

Human Promoter +Enhancer	Fold Higher	Chimp Promoter +Enhancer	Fold Higher

2901542	38.69	1684737	22.46
5176413	69.02	1650962	22.01
6484640	86.47	1867906	24.91

Empty

64447
110767
49774

Avg. 74996

2.2.8 Osteoblast (MG63) transfection data

31st January 2013

Human	Fold higher	Chimp	Fold higher
549763	45.96	164994	13.79
467423	39.08	141538	11.83
541244	45.23	169122	14.14

Empty

12252
12629
11005

7th February 2013

Human	Fold higher	Chimp	Fold higher
503768	28.6	169026	9.6
411770	23.38	174107	9.88

335842	19.07	166141	9.43

Empty

24441
16244
12160

9th May 2013

Human Promoter only	Fold higher	Chimp Promoter only	Fold higher
156117	70.45	13348	6.02
218287	98.50	28667	12.94
212114	95.72	16428	7.41

Human Promoter +Enhancer	Fold higher	Chimp Promoter +Enhancer	Fold higher
37157	16.77	3270	1.48
48734	21.99	4543	2.05
47263	21.33	6527	2.95

Empty

1748
2555
2346

2.2.9 Two-way ANOVA design for promoter-only transfection

2.2.9.1 The data file

Observation	Species	Batch	Day	Reading
1	Human	Batch1	7thsep	85.83
2	Human	Batch1	7thsep	91.59
3	Human	Batch1	7thsep	52.07
4	Human	Batch1	20thsep	229.63
5	Human	Batch1	20thsep	92.21
6	Human	Batch1	20thsep	110.85
7	Human	Batch1	4thoct	72.68
8	Human	Batch1	4thoct	79.01
9	Human	Batch1	4thoct	93.79
10	Human	Batch1	8thnov	200.22
11	Human	Batch1	8thnov	199.91
12	Human	Batch1	8thnov	179.85
13	Human	Batch2	31stoct	2.82
14	Human	Batch2	31stoct	2.49
15	Human	Batch2	31stoct	2.55
16	Human	Batch2	7thnov	242.65
17	Human	Batch2	7thnov	230.49
18	Human	Batch2	7thnov	244.12
19	Human	Batch2	13thdec	59.24
20	Human	Batch2	13thdec	49.32
21	Human	Batch2	13thdec	47.6
22	Chimp	Batch1	7thsep	26.98
23	Chimp	Batch1	7thsep	25.55
24	Chimp	Batch1	7thsep	6.73
25	Chimp	Batch1	20thsep	30.42
26	Chimp	Batch1	20thsep	24.08
27	Chimp	Batch1	20thsep	18.01
28	Chimp	Batch1	4thoct	10.75
29	Chimp	Batch1	4thoct	10.58
30	Chimp	Batch1	4thoct	14.38
31	Chimp	Batch1	8thnov	54.29
32	Chimp	Batch1	8thnov	58.77
33	Chimp	Batch1	8thnov	66.21
34	Chimp	Batch2	31stoct	5.5
35	Chimp	Batch2	31stoct	5.46
36	Chimp	Batch2	31stoct	15.66
37	Chimp	Batch2	7thnov	87.93

```
38   Chimp     Batch2    7thnov    128.11
39   Chimp     Batch2    7thnov    115.51
40   Chimp     Batch2    13thdec   37.98
41   Chimp     Batch2    13thdec   35.78
42   Chimp     Batch2    13thdec   33.94
43   Bonobo    Batch1    7thsep    109.28
44   Bonobo    Batch1    7thsep    64.55
45   Bonobo    Batch1    7thsep    116.72
46   Bonobo    Batch1    20thsep   73.52
47   Bonobo    Batch1    20thsep   39.81
48   Bonobo    Batch1    20thsep   90.37
49   Bonobo    Batch1    4thoct    35.16
50   Bonobo    Batch1    4thoct    38.82
51   Bonobo    Batch1    4thoct    40.18
52   Bonobo    Batch1    8thnov    105.31
53   Bonobo    Batch1    8thnov    136.68
54   Bonobo    Batch1    8thnov    151.18
55   Bonobo    Batch2    31stoct   4.82
56   Bonobo    Batch2    31stoct   3.02
57   Bonobo    Batch2    31stoct   3.06
58   Bonobo    Batch2    7thnov    124.56
59   Bonobo    Batch2    7thnov    102.09
60   Bonobo    Batch2    7thnov    138.64
61   Bonobo    Batch2    13thdec   63.92
62   Bonobo    Batch2    13thdec   55.86
63   Bonobo    Batch2    13thdec   79.55
```

2.2.9.2 The R codes and results

```
a <- read.table("two_way_anova.txt", header = T)

k<-aov(Reading~(Species*Batch)+(Species*Day)+(Batch*Day),a)

summary (k)

     Df Sum Sq Mean Sq F value Pr(>F)

Species 2 57674 28837 63.537 1.99e-13 ***

Batch 1 883 883 1.946 0.17031

Day 5 143987 28797 63.449 < 2e-16 ***

Species: Batch 2 7255 3628 7.993 0.00114 **
```

```
Species: Day 10 32983 3298 7.267 1.57e-06 ***

Residuals 42 19062 454
```

```
m <- pairwise.t.test(a$Reading, a$Species, p.adj =
"bonferroni")

head (m)

$method
[1] "t tests with pooled SD"

$data.name
[1] "a$Reading and a$Species"

$p.value

     Bonobo Chimp
Chimp 0.1428680 NA
Human 0.1213403 0.0003576344

$p.adjust.method
[1] "bonferroni"

n <- pairwise.t.test(a$Reading, a$Batch, p.adj = "none")

head (n)

$method
[1] "t tests with pooled SD"

$data.name
[1] "a$Reading and a$Batch"

$p.value

     Batch1
Batch2 0.6511376

$p.adjust.method
[1] "none"
```

```
boxplot(Reading~Batch*Species,data=a)
```

2.2.10 Two-way ANOVA design for Androgen transfection data

2.2.10.1 The data file

Observation	Species	Induction	Day	Reading
1	Human	Uninduced	20thDecember	45.26
2	Human	Uninduced	20thDecember	44.13
3	Human	Uninduced	20thDecember	42.5
4	Human	Induced	20thDecember	47.41
5	Human	Induced	20thDecember	43.8
6	Human	Induced	20thDecember	44.09
7	Human	Uninduced	27thDecember	32.03
8	Human	Uninduced	27thDecember	29.97
9	Human	Uninduced	27thDecember	28.46
10	Human	Induced	27thDecember	16.88
11	Human	Induced	27thDecember	11.74
12	Human	Induced	27thDecember	10.79
13	Human	Uninduced	3rdJanuary	93.14
14	Human	Uninduced	3rdJanuary	94.24
15	Human	Uninduced	3rdJanuary	77.52
16	Human	Induced	3rdJanuary	83.62
17	Human	Induced	3rdJanuary	83.4
18	Human	Induced	3rdJanuary	69.92
19	Chimp	Uninduced	20thDecember	24.26
20	Chimp	Uninduced	20thDecember	34.74
21	Chimp	Uninduced	20thDecember	30.99
22	Chimp	Induced	20thDecember	30.93
23	Chimp	Induced	20thDecember	32.9
24	Chimp	Induced	20thDecember	33.45
25	Chimp	Uninduced	27thDecember	19.5

```
26   Chimp     Uninduced 27thDecember   20.65
27   Chimp     Uninduced 27thDecember   20.55
28   Chimp     Induced   27thDecember   10.98
29   Chimp     Induced   27thDecember   13.07
30   Chimp     Induced   27thDecember   11.49
31   Chimp     Uninduced 3rdJanuary     89.47
32   Chimp     Uninduced 3rdJanuary     90.3
33   Chimp     Uninduced 3rdJanuary     88.77
34   Chimp     Induced   3rdJanuary     63.65
35   Chimp     Induced   3rdJanuary     92.33
36   Chimp     Induced   3rdJanuary     69.15
37   Bonobo    Uninduced 20thDecember   21.46
38   Bonobo    Uninduced 20thDecember   33.9
39   Bonobo    Uninduced 20thDecember   38.33
40   Bonobo    Induced   20thDecember   30.14
41   Bonobo    Induced   20thDecember   37.91
42   Bonobo    Induced   20thDecember   32.07
43   Bonobo    Uninduced 27thDecember   37.68
44   Bonobo    Uninduced 27thDecember   45.19
45   Bonobo    Uninduced 27thDecember   50.21
46   Bonobo    Induced   27thDecember   30.88
47   Bonobo    Induced   27thDecember   31.94
48   Bonobo    Induced   27thDecember   32.52
49   Bonobo    Uninduced 3rdJanuary     66.19
50   Bonobo    Uninduced 3rdJanuary     83.23
51   Bonobo    Uninduced 3rdJanuary     71.73
52   Bonobo    Induced   3rdJanuary     54.45
53   Bonobo    Induced   3rdJanuary     74.18
54   Bonobo    Induced   3rdJanuary     74.55
```

2.2.10.2 The R codes and results

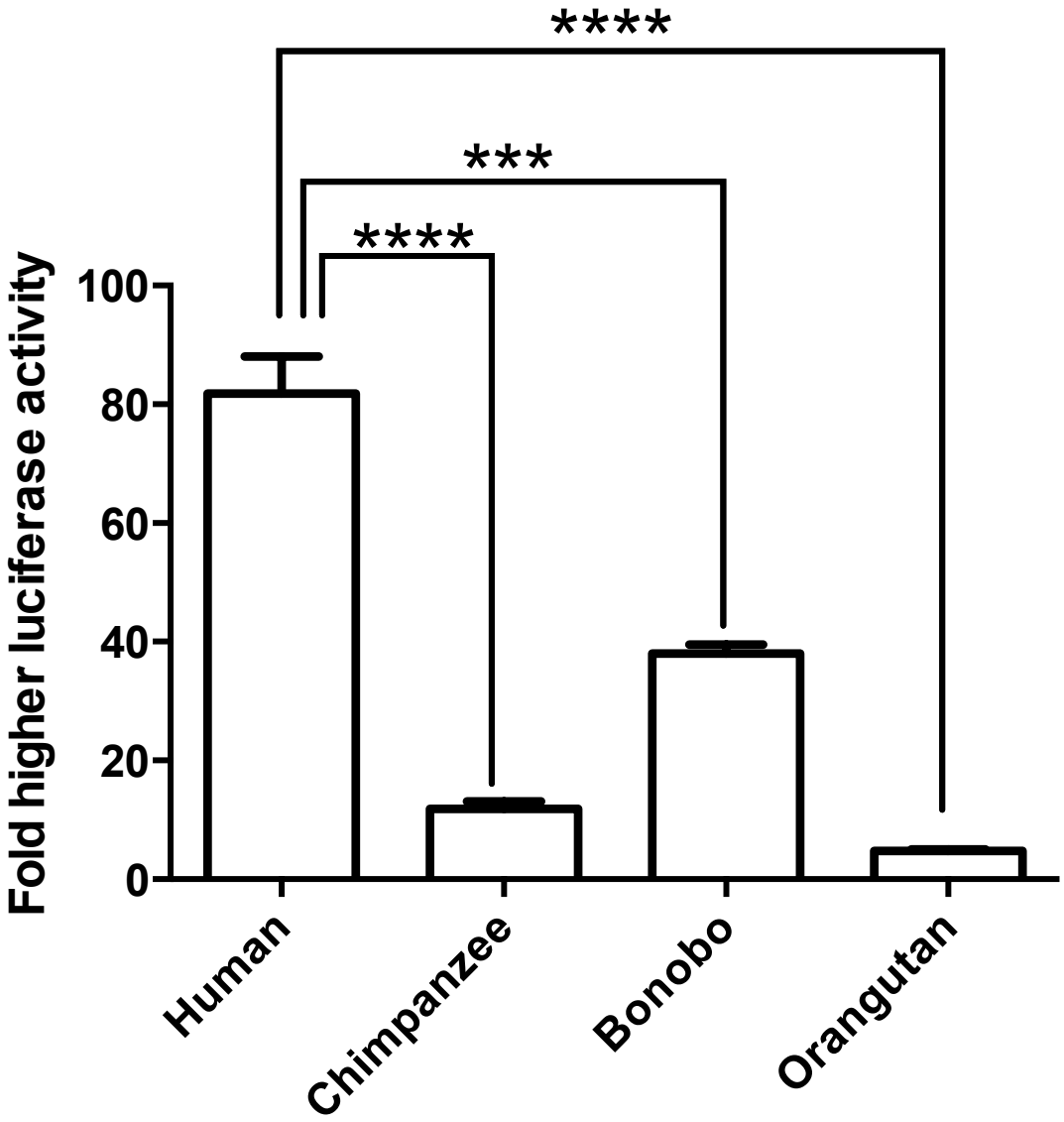
```
a <- read.table("two_way_anova_for_R1881.txt", header =T)

k <- aov(Reading~(Species*Induction)+(Species*Day)+
(Induction*Day),a)
summary (k)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Species	2	414	207	5.050	0.011075	*
Induction	1	642	642	15.649	0.000304	***
Day	2	28976	14488	353.267	< 2e-16	***
Species:Induction	2	19	10	0.235	0.791487	
Species:Day	4	2420	605	14.751	1.74e-07	***
Induction:Day	2	539	270	6.575	0.003396	**
Residuals	40	1640	41			

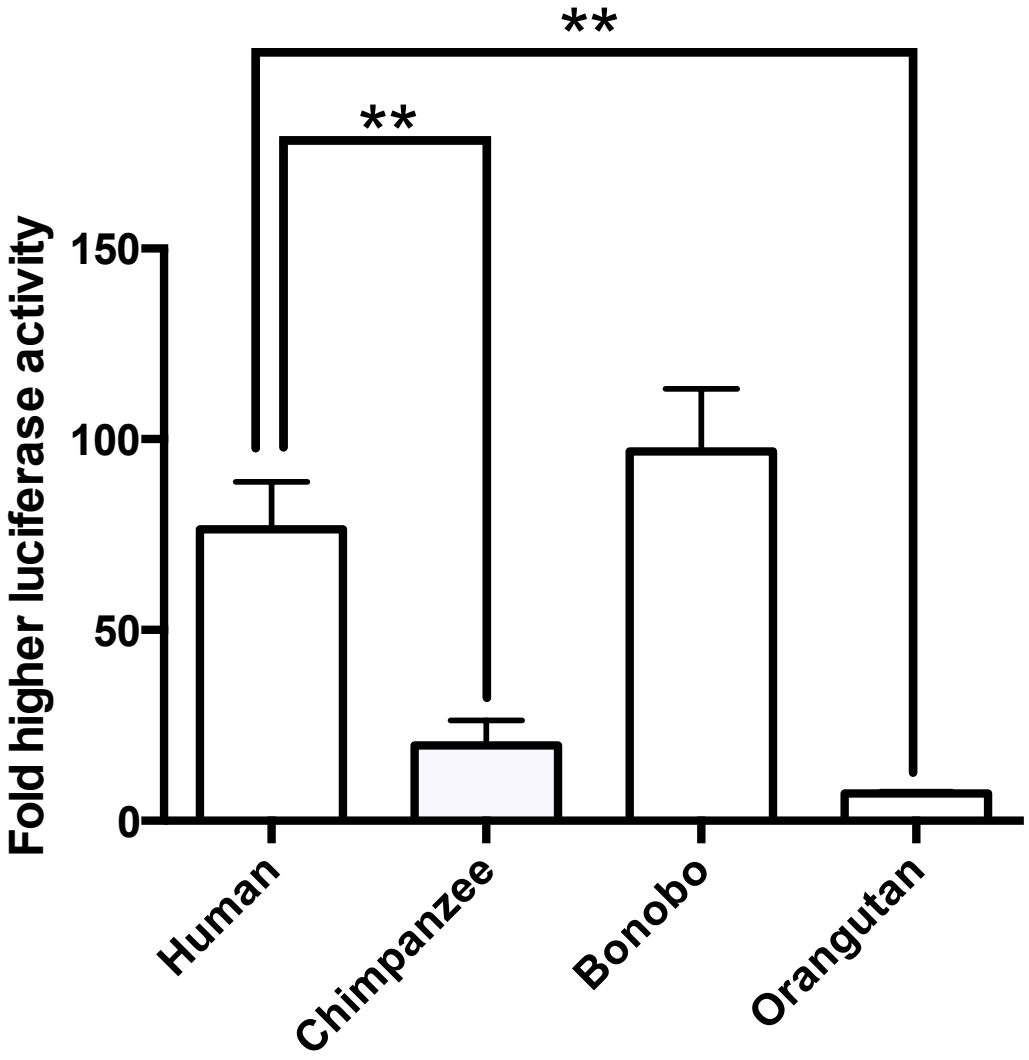
2.2.11 Additional graphs

2.2.11.1 Promoter only transfections

4th October 2012

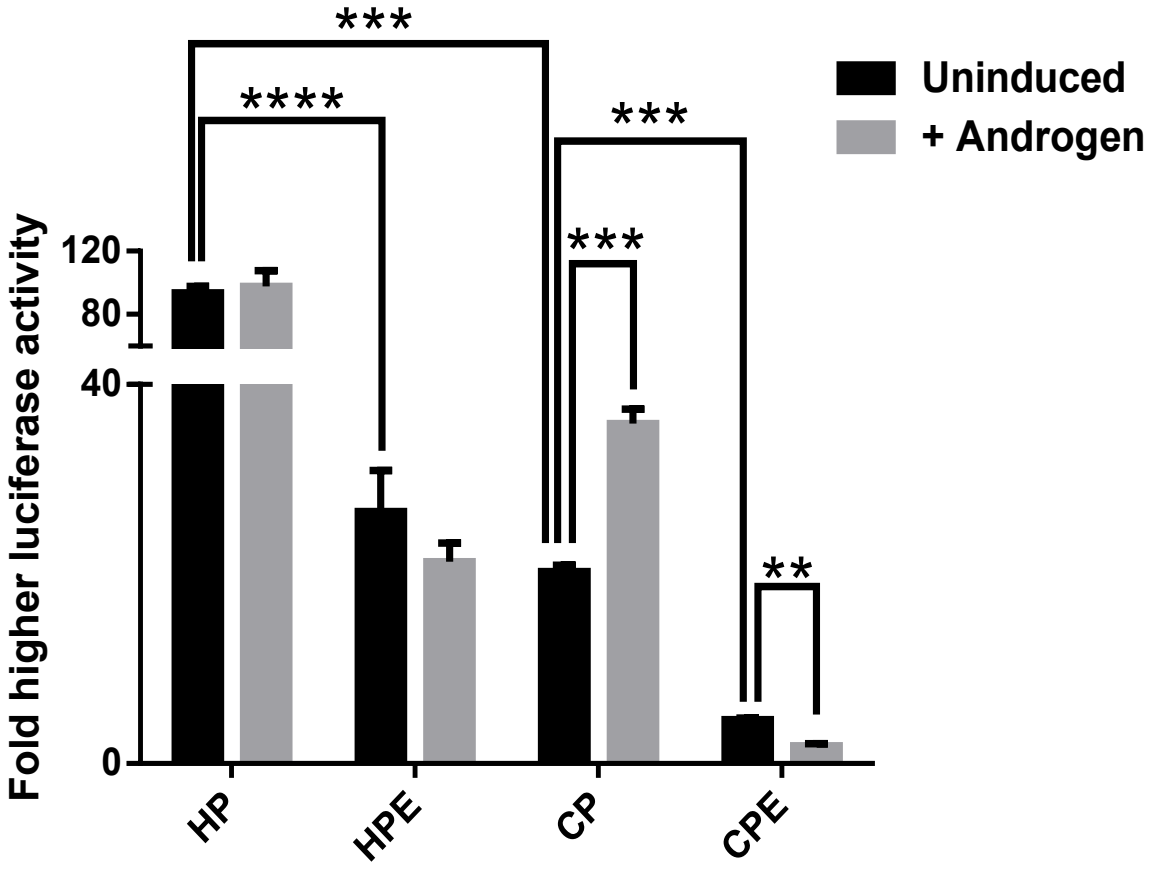


7th September 2012

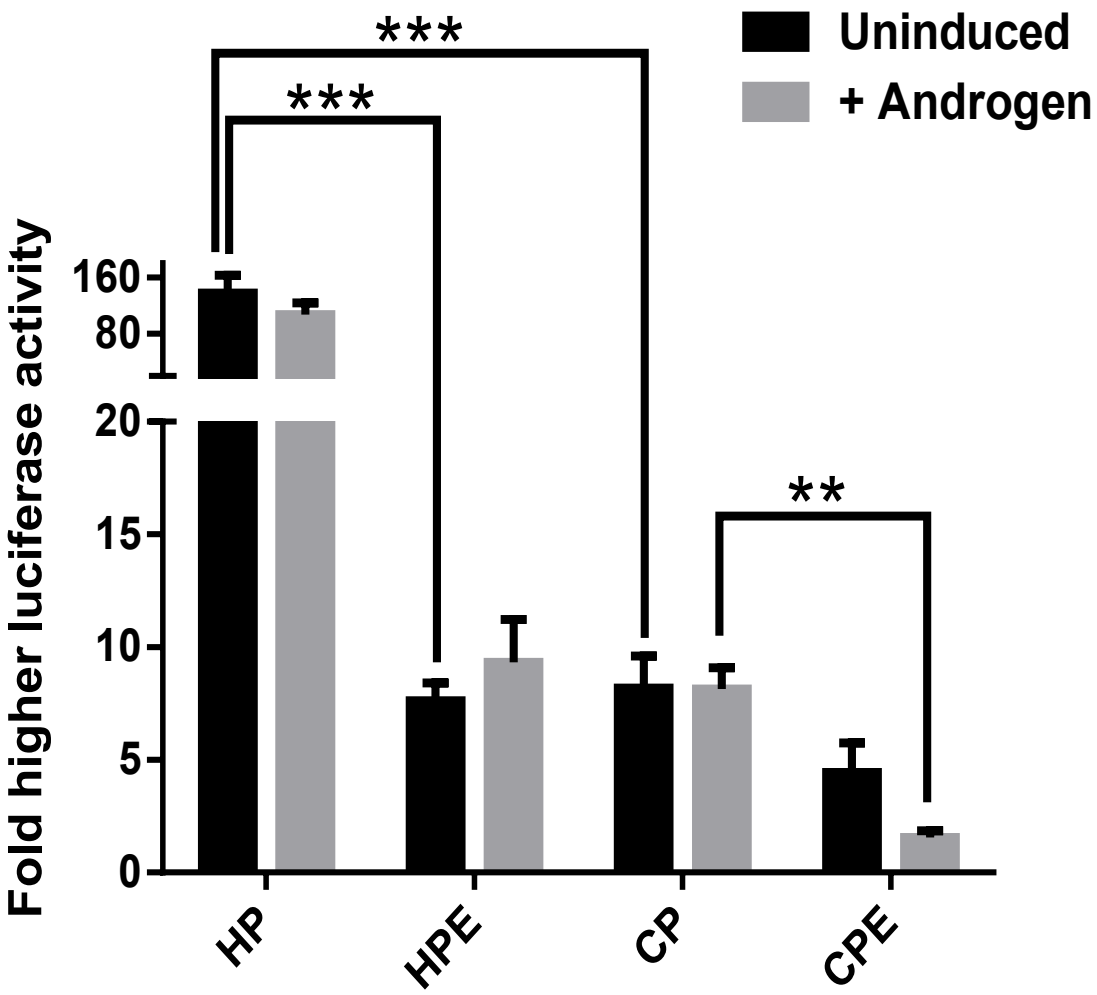


2.2.12.2 Promoter + Silencer transfections

11th April 2013



18th April 2013



Appendix 3: Data from Chapter 4

3.1 Alignment of uniquely gained and loss miRNAs and their homologs in other hominoids

MIR3124

```
Human             GAGTCCGGACGCTGGCGGGCTTCGCGGGCGAAGGCAAAGTCGATTTCCAAAAGTGACTTT 66
hsa-mir-3124      --------------GCGGGCTTCGCGGGCGAAGGCAAAGTCGATTTCCAAAAGTGACTTT 46
Gorilla           GAGTCCGGACGCTGGCGGACTTCGCGGGCGAAGGCGAAGTCGATTTCCAAAAGTGACTTT 66
Chimp             AAGTCCAGGTGTTAATGGACTCTGCTGTTGA-GATGCAGTTGATTTCCAAAAGTGACTTT 119
Orangutan         AAGTCCAGGTGTTAATGGATTCTGCTGTTGA-GATGCAGTTAATTTCCAAAAGTGACTTT 118
                                **  *  ** *  ** *    ***  ******************

Human             CCTCACTCCCGTGAA----GTCGGCGGAACCCTCCACTAACGGCG--------------- 107
hsa-mir-3124      CCTCACTCCCGTGAA----GTCGGC---------------------------------- 67
Gorilla           CCTCACTCCAGTGAA----GTCGGCGGAACCCTT-ACTAACGGCG--------------- 106
Chimp             TCTCAATCCAGTAAACTAAGTCAGCAGGACCCTT-AGTAATGGCGTCTAAGTGACAATAC 178
Orangutan         TCTCAACCCAGTAAACTAAGTCAGCAGGACCCAT-AGTAATGGCATCTAAGTGACAATAC 177
                   ****  ** ** **    *** **
```

MIR941

```
Human             AGAGGACGCACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAG 60
hsa-mir-941       ------------------TGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAG 42
Chimp             AGAGGATGCACCCGGCTGTGTGCACATGTGCCCAGGGCCCAGGACAGCGCCACGGAAGAG 60
Orangutan         AGAGGACGCACCCGGCTGTGTGGACATGTGCCCAGGGCCCAGGACAGCGCCACGGAAGAG 60
                            **** ***************** ******************

Human             GACGCACCCGGCTGTGT-GCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGAC 119
hsa-mir-941       GACGCACCCGGCTGTGT-GCACATGTGCCCA---------------------------- 72
Chimp             GACGCACAGGACAGCGCCACGGAAGAGGACG----------------------------- 91
Orangutan         GACACACCCTGCTGT---------------------------------------------- 75
                  *** ***    * *
```

MIR620

```
Human             TGACAATCCATACTAAAAACTTCTTTGCTTGTTTTATTTCTATATCTATCTCCATATATA 60
Chimp             TGACAATCCATACTAAAAACTTCTTTGCTTGTTTTATATCTATATCTATCTCCATATATA 60
Orangutan         TGACAATCCACACTAAAAACTTCTTTGCTTGTTTTATATCTATATCTATCTCCATATATA 60
hsa-mir           ---------------------------------TATATCTATATCTAGCTCCGTATATA 26
                                                  *** ********** **** ******

Human             --------TGGAGATATCTATATATATATATATATATACGGAGCTAGATATAGATATA 112
Chimp             --------TGGAGATATCTATATATATATATATA------CGGAGCTAGATACAGATATA 106
Orangutan         GATATAGATATAGATATAGATATAGATATAGATATAGATATAGATATAGATATA 120
hsa-mir           --------TATATATATATATATAGATATCTCCATATATATGGAGATAGATATAGAAATA 78
                        *  * **** ***** ****    *        **  ****** *** ***
```

MIR3913

```
Gorilla          AACTTAGAGAAAGTAGAAAACTGTCTATAATAAACA-----A----AGACTAATAATTCG 51
Orangutan        ----------AAATATTTCA--GTTTATTATAAACA-----A----AGACTAATAAGTCG 39
Chimp            AACTTAGAGAAAGTAGAAAACTGTCTATAATAAACTGAAATA----AGACTAATAAGTCG 56
Human            AACTTAGAGAAAGTAGAAAACTGTCTATAATAAACTGAAATATTTGGGACTGATC-TTGA 59
hsa-mir-3913     --------------------TGTCTATAATAAACTGAAATATTTGGGACTGATC-TTGA 38
                                    ** *** ******      *     **** **   *

Gorilla          TCTCTACTA---TTGGAGGAAAAAAAAAAAA--GACCGAAATGCAAAAATTCCTGAAAAT 106
Orangutan        TCTCTACTA---TTGGAGGAAAAAAAAAAAA---GACCGAAATGCAAAAATTCCTGAAAAT 93
Chimp            TCTCTACCC---TTGTAGGAAAAAAAAAAAAAAGACCGAAATGCAAAAA----------- 102
Human            TGTCTGCCAAGGTTTTGGCAGACATCAAGATCAGTCCCAAATATTTCAGTTTATTATAA- 118
hsa-mir-3913     TGTCTGCCAAGGTTTTGGCAGACATCAAGATCAGTCCCAAATATTTCAGTTTATTATAA- 97
                 * *** *     **   * * * *  ** *   * ** ****     *
```

MIR3919

```
Human            TCCAAGTAGTTAACCCTTTACCTGAGCACCATTTACTGAGTCCTTTGTTCTCTACTA--- 57
hsa-mir-3919     --------------------CCTGAGCACCATTTACTGAGTCCTTTGTTCTCTACTA--- 37
Orangutan        TCCAAGTAGTTAACCCTTTACCTGAGCACCATTTACTGAGTCCTTTGTTCTCTACTA--- 57
Gorilla          --------------------CTGAGCACCATTTACTGAGTCCTTTGTTCTCTGCTACGA 39
Chimp            ------------------AATTCGAGCTCCCCTTCGGGGG------GATC-CCACT---- 31
                                    **** **  **   * *      * ** *  **

Human            GTTTGTAGTAGTTCGTAGCAGAGAACAAAGGACTCAGTAAATGGTGCTCAGGAATCTTTA 117
hsa-mir-3919     GTTTGTAGTAGTTCGTAGCAGAGAACAAAGGACTCAGTAAATGGTGCTCAGG-------- 89
Orangutan        GTTTGTAG----------CAGAGAACAAAGGACTCAGTAAATGGTACTCAGGAATCTTTA 107
Gorilla          ACTACTACAAAC---TAGTAGAGAACAAAGGACTCAGTAAATGGTGCTCAGGAATCTTTA 96
Chimp            --TCGTAGAAG--------ACATGCTAGGGGACTCAGTAAATGGTGCTCAGGAATCTTTA 81
                   *  **            * *    *  ***************** ******
```

MIR3941

```
Human            CTGTACAATGTCTGCAGGTAGAGTCAGAATTCTCATCAGGCTGTGATGCTCAGTTGTGTG 60
hsa-mir-3941     --------------------GAGTCAGAATTCTCATCAGGCTGTGATGCTCAGTTGTGTG 40
Chimp            CTGTACAATGTCTGCAGGTAGAGTCAGAATTCTCATCAGGCTGTGATGCTCAGTTGTGTG 60
Gorilla          CTGTAGAATGTCTGCAGGTAGAGTCAGAATTCTCATCAGGCTGTGATGCTCAGTTGTGTG 60
Orangutan        CTGTAGAATGTCTGCAGGTAGAGTCGGAATTCTCATGAGGCTGTCATGCTCGGTTGTGTG 60
                               ***** ********** ******* ****** ********

Human            TAGATTGAAAGCCCTAATTTTACACACAACTGAGGATCATAGCCTGATGGTTCCTTTTTG 120
hsa-mir-3941     TAGATTGAAAGCCCTAATTTTACACACAACTGAGGATCATAGCCTGATGGTTCCTTTTTG 100
Chimp            TAGATTGAAAGCCCTAATTT-ACACACAACTGAGGCTCATAGCCTGATGGTTCCTTTTTG 119
Gorilla          TAGATTAAAAGCCCTAATTT-ACACACAACTGAGGATCATAGCCTGATGGTTCCTTTTTG 119
Orangutan        TAGACTGAAAGCCCCAATTT-ACACACAACTGAGCATCATAGCCTGATGGTTCCTTTTTG 119
                 **** * ******* ***** *************  ************************
```

MIR4329

```
Chimp            AGACTTCAAAGTGTGGAGGAGCTCCAGAATGTGGAACTAGGGTCTCAGGAGGAAAAAAAA 60
Orangutan        AGACTTCAAAGTGTGGAGGAGCTCCAGAATGTGGAACTAGGGTCTCAGGAGGAAAAAAAA 60
Human            AGACTTCAAAGTGTGGAGGAGCTCCAGAATGTGGAACTAGGGTCTCAGGAGGAAAAAAAA 60
hsa-mir-4329     --------------------GCTCCAGAATGTGGAACTAGGGTCTCAGGAGGAAAAAAAA 40
Gorilla          AGACTTCAAAGTGTGGAGGAGCTCCAGAATGTGGAACTAGGGTCTCAGGAGGGAAAAAAA 60
                              ********************************** *******

Chimp            A----AACCCCAAACCCTGGTACACCTTCCTCTCTAGAACCTGGACCATCATCCCT 112
Orangutan        AAAAAAACCCAAATCCCTGGTACACCTTTCTCTCTAGAACCTGGACCATCATCCCT 116
Human            A-----CTCCAAAACCCTGGTACACCTTCCTCTCTAGAACCTGGACCATCATCCCT 111
hsa-mir-4329     A-----CTCCAAAACCCTGGTACACCTTCCTCTCTA------------------- 71
Gorilla          A-----C-CCAAAACCCTGGTACACCTTCCTCTCTAGAACCTGGACCATCATCCCT 110
                 *       ** ** ************** *******
```

MIR1283

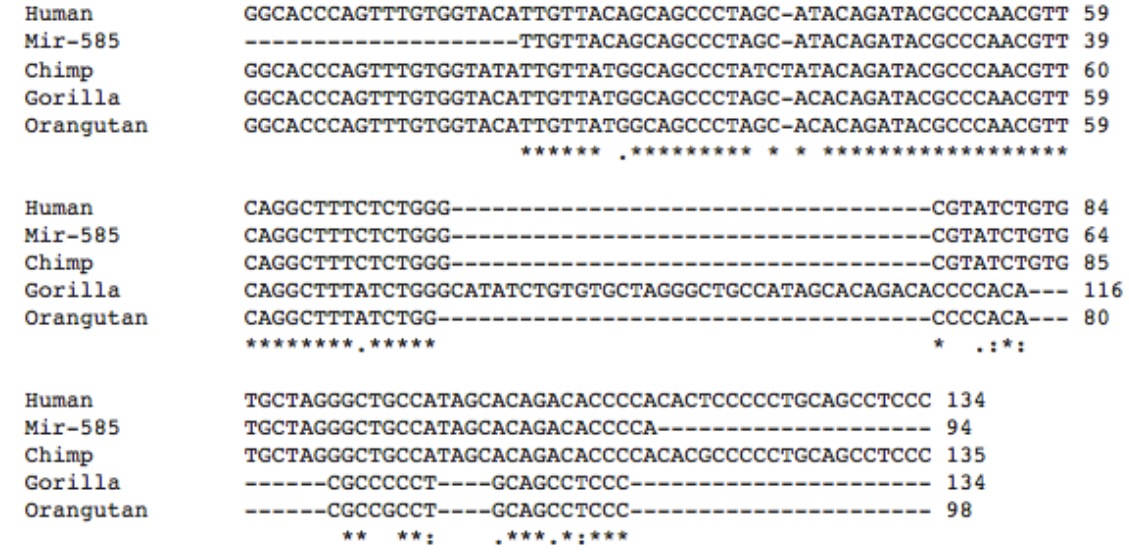
```
Chimp            CTCGTGCTGTGACTCTGCAAAGGGAAGCCCTTTCTGTTGTCTAAAAGAATAGAAAGCGCT 60
ppy-mir-1283     CTCATGCTGTGACTCTGCAAAGGGAAGCCCTTTCTGTTGTCTAAAAGAAAAGAAAGCGCT 60
Orangutan        CTCATGCTGTGACTCTGCAAAGGGAAGCCCTTTCTGTTGTCTAAAAGAAAAGAAAGCGCT 60
Gorilla          CTCATGCTGTGAGTCTGCAAAGGGAAGCCCTTTCTGTTGTCTGAAAGAAAAGAAATCGCT 60
Human            CTCAAGCTGTGA--CTGCAAAGGGAAGCCCTTTCTGTTGTCTAAAAGAAAAGAAAGTGCT 58
                 ***  *******  ***************************** ****** *****  ***

Chimp            TCCCTTTGGAGTGTTACGGTTTGAGAA---------------- 87
ppy-mir-1283     TCCCTTTGGAGTGTTACGGTTTGAGAA---------------- 87
Orangutan        TCCCTTTGGAGTGTTACGGTTTGAGAACTGGACCATCATCCCT 103
Gorilla          TCCCTTTGGAGTGTTACGGTTTGAGAA---------------- 87
Human            TCCCTTTGGTGAATTACGGTTTGAGAA---------------- 85
                 ********* *  ***************
```

MIR466

```
Human            GTGCCATTACGTGGGTGTGTGCATGTATATATGTGTGTTGCGTGTATGTGTATGTGTACA 60
Mir-466          --------------------GCATGTATATATGTGTGTTGCGTGTATGTGTATGTGTACA 40
Orangutan        GTGTCATTACGTGGGTGTGTGCATGTATATATGTGTGTTGTGTGTGTGTGTATGTGTACA 60
Chimp            GTGCCATTACGTGGGTGTGTGCATGTATATATGTGTGTTGCGTGTGTGTGTATGTGTACA 60
Gorilla          GTGCCATTACGTGGGTGTGTGCATGTATATATGTGTGTTGCGTGTGTGTGTATGTGTACA 60
                              ******************** **** **************

Human            TATATACACACATATACACACATGCAACACACATATACACACACATACAAATATACACAT 120
Mir-466          TATATACACACATATACACACATGCAACACACATATACACACAC---------------- 84
Orangutan        TATATACACACATATACACACATGCAACACACATATACACACACATACAAATATACATAT 120
Chimp            TATATATACACATATACACATA--CAACACGCATATACACACACATACAAATATACATAT 118
Gorilla          TATATATACACATATACACACA--CACAACACATATACACACACACACAAATATACATAT 118
                 ****** ************** *  **  ** **************
```

MIR578

MIR585



MIR611

MIR1270

```
Human          AGGGTCCCAGGTGAGCACAATGCAAAGAGCCACATAGAAGATAAAGAAAAGCCTGTTACT 60
Mir-1270       --------------------TGCAAAGAGCCACATAGAAGATAAAGAAAAGCCTGTTACT 40
Gorilla        AGGGTCCCAGGTGAGCACAATGCAAAGAGCCACATAGAAGATAAAGAAAAGCCTGTTACT 60
Chimp          ------CCAGGTGAGCACAATGCAAAGAGCCACATAGAA-AAAAAGAAAAGTCTGTTACT 53
Orangutan      ------CCAGGTGAGCACAATGCAAAGAGCCACATAGAA-AAAAAGAAAAGTCTGTTACT 53
                           ******************* *:********* ********

Human          TATACCCAACACAGCTCTTCCATATCTCCAGTATAACTCTGTGCCTTTAAAAGTAAATTA 120
Mir-1270       TATACCCAACACAGCTCTTCCATATCTCCAGTATAACTCTGTG----------------- 83
Gorilla        TATACCCAACACAGCTCTTCCATATCTCCAGTATAACACTGTGCCTTTAAAAGTAAATTA 120
Chimp          TATACCCAACACAGCTCTTACTGCTCTCCAGTATAAAATTGTGCCTTTAAAAGTAAATT- 112
Orangutan      TATACCCAACACAGCTCTTCCTGCTCTCCAGTGTAACATTGTGCCTTTAAAAGTAAATT- 112
               ******************.*: .********.***.: ****
```

MIR1287

```
Gorilla        ACTCCCTGGAGTGTGGGGACTTGTCATGGCTCCAATCACTGCATCTGTGGCTAGAGTGGC 60
Mir-1287       --------------------TTGTCATGGCTCCAATCACTGCATCTGTGGCTAGAGTGGC 40
Human          ACTCCCCGGAGTGTGGGGACTTGTCATGGCTCCAATCACTGCATCTGTGGCTAGAGTGGC 60
Chimp          ACTCCCCGGAGTGTGGGGACTTGTCATGGCTCCAATCACTGCATCTGTGGCTAGAGTGGC 60
Orangutan      ACTCCCCGGAGTGTGGGGACTTGTCATGGCTCCAATCACAGCATCTGTGGCTAGAGTGGC 60
                           ********************:**********************

Gorilla        TTTTAAAGGCTCAGACTCGAACCACTGATCCAGCACCTGGACAGCACAACAGCAGAGAGA 120
Mir-1287       TTTTAAAGGCTCAGACTCGAACCACTGATCCAGCACCTGGACAGCACAAC---------- 90
Human          TTTTAAAGGCTCAGACTCGAACCACTGATCCAGCACCTGGACAGCACAACAGCAGAGAGA 120
Chimp          TTTTAAAGGCTCAGACTCGAACCACTGAT-CAGCACCTGGACAGCACAACAGCAGAGAGA 119
Orangutan      TTTTAGAGGCTCAGACTCAAACCACTGATCCAGCACCTGGACAGCACAACAGCAGAGAGA 120
               *****.************.********** ********************
```

MIR2116

```
Gorilla        --------AGGA---CTTAGCATGGGAGGACTCCTAGTCCGGGAGTTCTTGGCATGGGAG 49
Orangutan      --------AGGA---CTTAGCATGGGAGGACTCCTAGTCTGGGAGTTCTTGGCAGGGGAG 49
Human          --------AGGA---CTTAGCATGGGAGGACTCCTAGTCTGGGAGTTCTTGGCATGGGAG 49
Mir-2116       ------------------------------TCCTAGTCTGGGAGTTCTTGGCATGGGAG 29
Chimp          CCTAGTCTAGGAGTTCTTACCATGGGAGGACTTC------GGGAGGACTGAGCATGGGAG 54
                                             * *      *****  **  *** *****

Gorilla        GACT--------------TCTTAGCATAGGAAGACCTCCTATGCTAAGAACTCCTAGCCT 95
Orangutan      GACT--------------TCTTAGCATGGGAAGACCTCCTGTGCTAAGAACTCCTAGCCT 95
Human          GACT--------------TCTTAGCATGGGAAGACCTCCTATGCTAAGAACCCCTAGCCT 95
Mir-2116       GACT--------------TCTTAGCATGGGAAGACCTCCTATGCTAAGAACCCCTAGCCT 75
Chimp          GACTCCTAGTCTGGGAGTTCTTGGCACGGGAAGATCTCCTATGCTAAGAACCCCTAGCCT 114
               ****              **** ***  ****** ***** ********** ********

Gorilla        AGGTCCCACTGACGGAGGAGATGCT 120
Orangutan      AGGTCCCACTGATGGAGGAGATGCT 120
Human          AGGTCCCACTGACGGAGGAGATGCT 120
Mir-2116       AGGTC-------------------- 80
Chimp          AGGTCCCACTGACGGAGGGGATGCT 139
               *****
```

MIR3125

MIR3179

MIR3679

MIR3916

```
Human          GCAAGTTCTAAACACATAACATCCCAGAAGAAAAGAGGAGACACATCCCCTGAACCCAGA 60
Mir-3916       --------------------ATCCCAGAAGAAAAGAGGAGACACATCCCCTGAACCCAGA 40
Chimp          GCAAGTTCTAAGCACATAACATCCCAGAAGAAAAGAGGA--CACATCCCCTGAACCCAGA 58
Orangutan      GCAAGTTCCAAACACATAACATCCCAGAAGAAAAGAGGA--CACATCCCCTGAACCCAGA 58
Gorilla        GCAAGTTCTAAACACATAACATCCCAGAAGAAAAGAGGA--CACATCCCCTGAACCCAGA 58
                                ******************  ******************

Human          CACATTCACCTGAGAACCAGCCATTTCTTCCTCTTCTTCCTTCTTCTCTGGGATTCCTTC 120
Mir-3916       CACATTCACCTGAGAACCAGCCATTTCTTCCTCTTCTTCCTTCTTCTCTGGGAT------ 94
Chimp          CACATTCACCTGAGAACCAGCCATTTCTTCCTCTTCTTCCTTCTTCTCTGGGATTCCTTC 118
Orangutan      CACATTCACCTGAGAACCAGCCATTTCTTCCTCTTCTTCCTTCTTCTCTGGGATTCCTTC 118
Gorilla        CACATTCACCTGAGAACCAGCCATTTCTTCCTCTTCTTCCTTCTTCTCTGGGATTCCTTC 118
               ************************************************************
```

MIR3925

```
Human          AGGAAGTGATGAAAGATGTGGTGGAAATAGCAAGAGAACTAAAACTGGAGTCTGGAGATG 60
Chimp          AGGAAGTGATGAAAGATGTGGTGGAAATAGCAAGAGAACTAAAACTG-AGTCTGGAGATG 59
Gorilla        AGGAAGTGATGAAAGATGTGGTGGAAATAGCAAGAGAACTAAAACTGGAGTCTGGAGATG 60
Mir-3925       --------------------GTGGAAATAGCAAGAGAACTAAAACTGGAGTCTGGAGATG 40
Orangutan      AGGAAGTGATGAAAGATGTGGTGGAAATAGCAAGAGAACTAAAACTGGAGTCTGGAGATG 60
                                ***************************** ************

Human          TGACAGGCTCCACTTTCAGTTCTCTTGCTATTCCCACATGACCATGGTCCAGGATCC 117
Chimp          TGACAGGCTCCACTTTCAGTTCTCTTGCTATTCCCACATGACCATGGTCCAGGATCC 116
Gorilla        TGACAGGCTCCACTTTCAGTTCTCTTGCTATTCCCACATGACCATGGTCCAGGATCC 117
Mir-3925       TGACAGGCTCCACTTTCAGTTCTCTTGCTATTCCCAC------------------- 77
Orangutan      TGACGGGCTCCACTTTCAGTTCTCTTGCTATTCCCACATGACCGTGGTCCAGGATCC 117
               **** ********************************
```

MIR3938

```
Gorilla        GCTCTACACCTCATGACCTCCATGGAGCCATCCAACCTGATCACCAGGTTATCTACAAGG 60
Orangutan      GCTCTACACCTCATGACCTCCATGGAGCCATCCAACCTGATCACCAGGTTATCTACAAGG 60
Chimp          GCTCTACACCTCATGACCTCCATGGAGCCATCCAACCTGACCACCGGGTTATCTACAAGG 60
Human          GCTCTACACCTCATGACCTCCATGGAGCCATCCAACCTGACCACCGGGTTATCTACAAGG 60
Mir-3938       --------------------CATGGAGCCATCCAACCTGACCACCGGGTTATCTACAAGG 40
                                  ******************** **** **************

Gorilla        GAATTTTAAAAATTAAAAAAA--TTCCCTTGTAGATAATCTAGTGATCAGGTTAAAAATT 118
Orangutan      GAATTTTAAAAATTAAAAAAA--TTCCCTTGTAGATAATCTGGTGATCAGGTTAAAAATT 118
Chimp          GAATTTTAAAAATTAAAAAAA--TTCCCTAGTAGATAATCTAGTGATCAGGTTAAAAATT 118
Human          GAATTTTTTAAATTAAAAAAAAATTCCCTTGTAGATAATCTAGTGATCGGGTTAAAAATT 120
Mir-3938       GAATTTTTTAAATTAAAAAAAAATTCCCTTGTAGATAATCTAGTGATCGGGTTAAAAATT 100
               *******  ************  ****** *********** ****** ***********

Gorilla        CCTTTAAAAAAATCTTGCAGCAG 141
Orangutan      CCTTTAAAAAAACCTTGCAGCAG 141
Chimp          CCTTTAAAAAAATCTTGCAGCAG 141
Human          CCTTTAAAAAAATCTTGCAGCAG 143
Mir-3938       CCT-------------------- 103
               ***
```

MIR4327

```
Human          TGCTAACTATGAAGAACCCACTACTGAGTAAAGG-CTTGATGAGAACTCCCTGGACTAAC 59
Mir-4327       --------------------CTACTGAGTAAAGG-CTTGATGAGAACTCCCTGGACTAAC 39
Chimp          TGCTAACTATGAAGGACCCACTACTGAGTAAAGGGCTTGATGAGAACTCCCTGGACTAAC 60
Orangutan      TGCTAACTATGAAGGACCCACTACTGAGTAAAGGGCTTGATGAGAACTCGCTGGACTAAC 60
Gorilla        TGCTAACTATGAAGGACCCACTACTCAGTAAAGGGCTTGATGAGAACTCCCTGGACTAAC 60
                           ***** ******** ************** **********

Human          CTGTTCATGGTCTCTTCCC-AGTCCCCCATGCAAGCCTACCCAGGCCTGTGGCTTTCTCG 118
Mir-4327       CTGTTCATGGTCTCTTCCC-AGTCCCCCATGCAAGCCTACCCAGGCC------------- 85
Chimp          CTGTTCATGGTCTCTTCCC-AGTCCCCCATGCAAGCCTACCCAGGCCTGTGGCTTTCTCG 119
Orangutan      CTGTTCATGGTCTCTTCCCCAGTCCCCCATGCAAGCCTACCCAGGCCTGTGGCTTTCTCG 120
Gorilla        CTGTTCATGGTCTCTTCCC-AGTCTCCCATGCAAGCCTACCCAGGCCTGTGGCTTTCTCG 119
               ******************* **** **********************
```

MIR1-1

```
Chimp          CTGCATACAGACTGCCTGCTTGGGAAACATACTTCTTTATATGCCCATATGGACCTGCTA 60
Mir-1-1        --------------------TGGGAAACATACTTCTTTATATGCCCATATGGACCTGCTA 40
Orangutan      -------CAGACTGCCTGCTTGGGAAACATACTTCTTTATATGCCCATATGGACCTGCTA 53
Human          CTGCATGCAGACTGCCTGCTTGGGAAACATACTTCTTTATATGCCCATATGGACCTGCTA 60
Gorilla        -------------------------ACATACTTCTTTATGTACCCATATGAACATACAA 34
                                ************** * ******** ** * * * *

Chimp          AGCTATGGAATGTAAAGAAGTATGTATCTCAGGCCGGGACCTCTCTCGCCG 111
Mir-1-1        AGCTATGGAATGTAAAGAAGTATGTATCTCA-------------------- 71
Orangutan      AGCTATGGAATGTAAAGAAGTATGTATCTCAGGCCGGGACCTCTCTCGCCG 104
Human          AGCTATGGAATGTAAAGAAGTATGTATCTCAGGCCGGGACCTCTCTCGCCG 111
Gorilla        TGCTATGGAATGTAAAGAAGTATGTAT------------------------ 61
                ***************************
```

MIR132

```
Orangutan      -----TCGGGGCGCGGCGTGGCGCGGCGCGTGGGCGTGCTGCGGGGCGACCATGGCTGTA 55
Mir-132        ------------------------GCGCGTGGGCGTGCTGCGGGGCGACCATGGCTGTA 35
Human          -----TCGGGGCGCGGCGTGGCGCGGCGCGTGGGCGTGCTGCGGGGCGACCATGGCTGTA 55
Gorilla        TCGGGGCGCGGCGTGGCGTGGCGCGGCGCGTGGGCGTGCTGCGGGGCGACCATGGCTGTA 60
Chimp          TCGGGGCGCGGCGTG-----GCGCGGCCCGTGGGCGTGCTGCGGGGCGACCATGGATGTA 55
                                  ** ************************** ****

Orangutan      GACTGTTACCTCCAGTTCCCACAGTAACAATCGAAAGCCACGGTTGCCCTGGAGACGCGG 115
Mir-132        GACTGTTACCTCCAGTTCCCACAGTAACAATCGAAAGCCACGGTTGCCCTGGAGACGCGG 95
Human          GACTGTTACCTCCAGTTCCCACAGTAACAATCGAAAGCCACGGTTGCCCTGGAGACGCGG 115
Gorilla        GACTGTTACCTCCAGTTCCC---------------------------------------- 80
Chimp          CACTGTTACCTCC------------------------------------------------ 68
                ************

Orangutan      GGGCGGGGCGCGCGGACGGCGCCGGC 141
Mir-132        GGGCGG--------------------- 101
Human          GGGCGGGGCGCGCGGACGGCGCCGGC 141
Gorilla        -------------------------
Chimp          -------------------------
```

MIR184

```
Chimp          CATCAAAACTTCTTTGCCGGCCAGTCACGTCCCCTTATCACTTTTCCAGCCCAGCTTTGT 60
Orangutan      CATCAAAACTTCTTTGCCGGCCAGTCACGTCCCCTTATCACTTTTCCAGCCCAGCTTTGT 60
Human          CATCAAAACTTCTTTGCCGGCCAGTCACGTCCCCTTATCACTTTTCCAGCCCAGCTTTGT 60
Mir-184        --------------------CCAGTCACGTCCCCTTATCACTTTTCCAGCCCAGCTTTGT 40
Gorilla        ---------TTCCAAGC---CCAG--------CTTTGTGACTGTTAC----CAGTTTTAG 36
                               ****        * ** * *** ** *    *** ***

Chimp          GACT---GTAAGTGTTGGACGGAGAACTG-ATAAGGGTAGGTGATTGACACTCACAGCCT 116
Orangutan      GACT---GTAAGTGTTGGACGGAGAACTG-ATAAGGGTAGGTGATTGACACTCACAGCCT 116
Human          GACT---GTAAGTGTTGGACGGAGAACTG-ATAAGGGTAGGTGATTGACACTCACAGCCT 116
Mir-184        GACT---GTAAGTGTTGGACGGAGAACTG-ATAAGGGTAGGTGATTGA------------ 84
Gorilla        GACCTTGGTAACTGT--GATGGTTAACTTTATATGTCAACTTGACTGGC-CTAAGAGATG 93
               ***    **** ***  ** **  ****  *** *   *  *** **
```

MIR3153

```
Chimp          CGTACTAGAACATGGTGGCTGTAATCGATCAGTAAAATT---------------------- 99
Orangutan      TGTACTAGAACATGGTGGCTGTAATCAATCAGTAAAATTAAACAAATTTTAAATGTCTCT 120
Gorilla        CGTACTAGAACATGGTGGCTGTAATCGATCAGTAAAATTAGACAAATTTTAAATGTCCCT 120
Human          CGTACTAGAACATGGTGGCTGTAATCGATCAGTAAAATTAGACAAATTTTAAATGTCCCT 120
Mir-3153       ---------------------------------------GACAAATTTTAAATGTCCCT 20

Chimp          ------------------------------------------------------------TG 101
Orangutan      GTCCCCTTCCCGCCAATTAAACTAGATTGGGGGAAAGCGGGTAGGGACATTTAAAATTTG 180
Gorilla        GTCCCCTTCCCCCCAATTAAAGTAGATTGGGGGAAAGCGAGTAGAGACATTTAAAATTTG 180
Human          GTCCCCTTCCCCCCAATTAAAGTAGATTGGGGGAAAGCGAGTAGGGACATTTAAAATTTG 180
Mir-3153       GTCCCCTTCCCCCCAATTAAAGTAGATTGGGGGAAAGCGAGTAGGGACATTTAAAATTTG 80

Chimp          TTGTCCTTACTGTATGCGGGACACTTCATTTAGGGCGTGGCTTCAGAAAACAGTCAACCC 161
Orangutan      TTGTCCTTATTGTATGCGGGACACTTCATTTAGGGCGTGGCTTCAGAGAACAGTCAACCC 240
Gorilla        TTGTCCTTACTGTATGCGGGACACTTCATTTAGGGCGTGGCTTCAGAAAACAGTCAACCC 240
Human          TTGTCCTTACTGTATGCGGGACACTTCATTTAGGGCGTGGCTTCAGAAAACAGTCAACCC 240
Mir-3153       TT------------------------------------------------------------ 82
```

MIR320C1

```
Human          AATTTTTAAGGAAAATGTGTTTTGCATTAAAAATGAGGCCTTCTCTTCCCAGTTCTTCCC 60
Chimp          AATTTTTAAGGAAAATGTGTTTTGCATTAAAAATGAGGCCTTCTCTTCCCAGTTCTTCCC 60
Mir-320C1      --------------------TTTGCATTAAAAATGAGGCCTTCTCTTCCCAGTTCTTCCC 40
Gorilla        AATTTTTAAGGAAAATGTGTTTTGCATTAAAAATGAGGCCTTCTCTTCCCAGTTCTTCCC 60
Orangutan      AATTTTTAAGGAAAATGTGTTTTGCATTAAAAATGAGGCCTTCTCTTCCCACTTCTTCCC 60
                             ********************************* ********

Human          AGAGTCAGGAAAAGCTGGGTTGAGAGGGTAGAAAAAAAA-TGATGTAGGTGAATAACTAT 119
Chimp          AGAGTCAGGAAAAGCTGGGTTGAGAGGGTAGAAAAAAAA-TGATGTAGGTGAATAACTAT 119
Mir-320C1      AGAGTCAGGAAAAGCTGGGTTGAGAGGGTAGAAAAAAAA-TGATGTAGG----------- 88
Gorilla        AGAGTCAGGAAAAGCTGGGTTGAGAGGGTAGAAAAAAAAAATGATGTAGGTGAATAACTAT 120
Orangutan      AGAGTCAGGAAAAGCTGGGTTGAGAGGGTAGAAAAAAAA-TGATGTACGTGAATAACTA- 118
               ****************************************** ******* *
```

MIR567

```
Human          TGTGATGAATCTATGGAAGAGGATTCTTATAGGACAGTATGTTCTTCCAGGACAGAACAT 60
Chimp          TGTGATGAATCTATGGAAGAGGATTCTTATAGGACAGTATGTTCTTCCAGGACAGAACAT 60
Gorilla        TGTGATCAATCTATGGAAGAGGATTCTTATAGGACAGTATGTTCTTCCAGGACAGAACAT 60
Mir-567        -------------------GGATTCTTATAGGACAGTATGTTCTTCCAGGACAGAACAT 40
Orangutan      TGTGATGAATCTATGGAAGAGGATTCTTATAGGACAGTATGTTCTTCCAGGGCAGAACAT 60
                           ******************************** ********

Human          TCTTTGCTATTTTGTACTGGAAGAACATGCAAAACTAAA---------------AAAAAA 105
Chimp          TCTTTGCTATTTTGTACTGGAAGAACATGCAAAACTAAA---------------AAAAAA 105
Gorilla        TCTTTGCTATTTTGTACTGGAAGAACATGCAAAACTAAAAAAAAAAAAAAAAAAAAAAAA 120
Mir-567        TCTTTGCTATTTTGTACTGGAAGAACATGCAAAACTAAA---------------AAAAAA 85
Orangutan      TCTTTGCTATTTTGTACTGGAAGAACATGCAAAACTAAA---------------AAAAAA 105
               ************************************                 ******

Human          AAAA------GTTATTGCTTAAATTCATTTCTTGAGTTG 138
Chimp          AAAAAAAAAAGTTATTGCTTAAATTCATTTCTTGAGTTG 144
Gorilla        AAAA------GTTATTGCTTAAATTCATTTCTTGAGTTG 153
Mir-567        AAAA------GTTATTGCT-------------------- 98
Orangutan      AAAA------GTTATTGCCTACATTCATTTCTTGA---- 134
               ****      ********
```

MIR708

```
Orangutan        GAGAGATGACTAGG--GTAAGGTCCCTGCCCTCTAGAAGCTCACAGTCTAGTTGTGTTCA 58
Mir-708          ----------------------TCCCTGCCCTCTAGAAGCTCACAGTCTAGTTGTGTTCA 38
Chimp            GAGAGATGACTAGG--GTAAGGTCCCTGCCCTCTAGAAGCTCACAGTCTAGTTGTGTTCA 58
Human            GAGAGATGACTAGG--GTAAGGTCCCTGCCCTCTAGAAGCTCACAGTCTAGTTGTGTTCA 58
Gorilla          ---------CTAGATTGTAAGCTCCTTGAAGGCAGGGA--TCCCAGCCTGG-------CA 42
                               *** **    *  * *  ** *** ** *       **

Orangutan        TGTGCAAGTCATTTACCCCCAGCTAGATTGTAAGCTCCTTGAGGGCAGTTACCATTTCAC 118
Mir-708          TGTGCAAGTCATTTACCCCCAGCTAGATTGTAAGCTCCTTGAGGGCAGTT---------- 88
Chimp            TGTGCAAGTCATTTACCCCCAGCTAGATTGTAAGCTCCTTGAGGGCAGTTACCATTTCAC 118
Human            TGTGCAAGTCATTTACCCCCAGCTAGATTGTAAGCTCCTTGAGGGCAGTTACCATTTCAC 118
Gorilla          CATAGTAGGGGCTT-CTCTTAGTCAAGATCCAG---CCAGGAAACCAGAACTCAAGTCAG 98
                   *   **    ** * *  **  *   *  *    **  **   ***
```

MIR1237

```
Chimp            GTCAAGATCCTCAGTCGCAGGTGGGAGGGCCCAGGCGCGGGCAGGGGTGGGGGTGGCAGA 60
Mir-1237         --------------------GTGGGAGGGCCCAGGCGCGGGCAGGGGTGGGGGTGGCAGA 40
Human            GTCAAGATCCTCAGTCGCAGGTGGGAGGGCCCAGGCGCGGGCAGGGGTGGGGGTGGCAGA 60
Orangutan        GTCAAGATCCTCAGTCGCAGGTGGGGGGGCCCAGGCGCGGGAAGGGGTGGGGGTGGCAGA 60
Gorilla          ------------------------------------------------------------

Chimp            GCGCTGTCCCGGGGGCGGGGCCGAAGCGCGGCGACCGTAACTCCTTCTGCTCCGTCCCCC 120
Mir-1237         GCGCTGTCCCGGGGGCGGGGCCGAAGCGCGGCGACCGTAACTCCTTCTGCTCCGTCCCCC 100
Human            GCGCTGTCCCGGGGGCGGGGCCGAAGCGCGGCGACCGTAACTCCTTCTGCTCCGTCCCCC 120
Orangutan        GCGCTGTCCCGGGGGCAGGGCTGAAGCGCGGCGACAGTAACTCTTTCTGCTCCGTCCCCC 120
Gorilla          ----TGTCCCGGGGGCGGGGCCGACG---GGCGCTCAGGTCTGGGCCACTTTCATCC--- 50
                     ************ **** ** *   ****       **    *   * * ***
```

MIR1263

```
Human          TACTGATGGGACTTACAGCATGCTACCCCAAAATATGGCAACATGGCATACTGAGTATGC 60
Mir-1263       --------------------TGCTACCCCAAAATATGGCAACATGGCATACTGAGTATGC 40
Gorilla        TACTGATGGGACTTAGAGCATGCTACCCCAAAATATGGCACCATGGCATACTGAGTATGC 60
Chimp          TACTGATGGGACTTAGAGCATGCTACCCCAAAATATGGCACCATGGCATACTGAGTATGC 60
Orangutan      -ACTGATGGGACTTAGAGCATACTACCCCAAAATATGGCACTATGGTATACTGAGTATGC 59
                           * ******************  **** *************

Human          CAGTATTAAAATACTCAGTATGCCAGG--------------------GTACCATATTTTG 100
Mir-1263       CAGTATTAAAATACTCAGTATGCCAGG--------------------GTACCATATTTTG 80
Gorilla        AAGTATTAAAATACTCAGTATGCCAGGGTAATACTCAGTATGCCAGGGTATCATATTTTG 120
Chimp          CAGTATTAAAATACTCAGTATGCCAGG--------------------GTACCATATTTTG 100
Orangutan      CCGTATTAAAATACTCAGTATGCCAGG--------------------GTACCATATTTTG 99
                 *************************                    *** *********

Human          GGGTAGCTGAAGGAATTAGAGAGTCA 126
Mir-1263       GGGTAG-------------------- 86
Gorilla        GGGTAGCTGAAGGAATTAGAGAG--- 143
Chimp          GGGTAGCTGAAGGAATTAGAGAG--- 123
Orangutan      GGGTAGCTAAAGGAATTTGAGAG--- 122
               ******
```

MIR1289

```
Human          ATGCAGCGCTTCCTAAAACGTGTTCCCAAT-TATCAGAATCACCCTGGGTACTTTCTAAA 59
Mir-1289-1     --------------------TGTTCCCAAT-TATCAGAATCACCCTGGGTACTTTCTAAA 39
Gorilla        ATGCAGCGCTTCCTAAAACGTGTTCCCAATGTATCAGAATCACCCTGGGTACTTTCTAAA 60
Chimp          ATGCAGCGCTTCCTAAAACGTGTTCCCAATGTATCAGAATCACCCTGGGTACTTTCTAAA 60
Orangutan      ATGCAGCGCCTCCTAAAACGTGTTCCCAATGTATCAGAATCACCCTGGGTACTTTCTAAA 60
                           ********** ******************************

Human          ATGCAGATTCCTGGACTCCACCCCCGGTTTAGGGATTCTCTGGGGGTGGAAACCAAGAGT 119
Mir-1289-1     ATGCAGATTCCTGGACTCCACCCCCGGTTTAGGGATTCTCTGGGGGTGGAAACCAAGAGT 99
Gorilla        ATGCAGATTCCTGGATTCCACCCCCGGTTTAGGGATTCTCTGGGGGTGGAATCCAAGAGT 120
Chimp          ATGCAGATTCCTGGACTCCACCCCCGGTTTAGGGATTCTCTGGGGGTGGAATCCAAGAGT 120
Orangutan      ATGCAGATTCCTGGACTCCACCCCCGGTTTACGGATTCTCTGGGGGTGGAATCCAAGAGT 120
               *************** *************** ******************* ********

Human          CTGCATTTACCAGTTTTGTTTTTAATTCCTACTAAAAATTGAGAACCACTGAGTGCAAGA 179
Mir-1289-1     CTGCATTTACCAGTTTTGTTTTTAATTCCTACTAAAAATTGAGAA--------------- 144
Gorilla        CTGCATTTACCAGTTTTGTTTTTAATTCCTACTAAAAATTGAGAACCACTGAGTGCAAGA 180
Chimp          CTGCATTTACCAGTTTTGTTTTTAATTCCTACTAAAAATTGAGAACCACTGAGTGCAAGA 180
Orangutan      CTGCATTTACCAGTTTTGTTTTTAATTCCTACTAAAAATTAAGAACCACTGAGTGCAAGA 180
               ****************************************** ****
```

MIR1299

```
Human           TCACACAGAATTCCAGAACACTGCTACGTGGGTCTGAATGATTGTCCCTCACGTAGGATT 84
Mir-1299        TCACACAGAATTCCAGAACACTGCTACGTGGGTCTGAATGATTGTCCCTCACGTAGGATT 64
Chimp           TCACACAGAATTCCAGAACACTGCTACGTGGGTCTGAATGATTGTCCCTCATGTAGGATT 180
Orangutan       TCATATAGGATTCCCGAACACAAGGGCTGGGATCTGAGTCTTTGTCCCACACATTGAATT 84
Gorilla         TCACAAAGGGTTCCAGAGCAGAGCTGTTGGAGTCATAATGTTTGTCTGTCACACAGGATT 84
                *** * **  **** ** **         *  **  * *  *****   **    * ***

Human           CCAGAACACTGCCATGAGGGTCTTAATGTTTGCTCCTCA 123
Mir-1299        CCAGAACACTGCCATGAGG------------------- 83
Chimp           CGAGAACACTGCCACGAGGGTCTTAATGTTTG------- 212
Orangutan       CCAGAACACTGCTGCT----------------------- 100
Gorilla         CTAGAACACTGCTACG----------------------- 100
                * **********
```

MIR1303

```
Human           ----------------------AGGCCGGGAGTTTGAGATCAGGCTGGGCAACATAGC 36
Mir-1303        ----------------------------------------GGCTGGGCAACATAGC 16
Gorilla         ----------------------AGGCCGGGAGTTTGAGATCAGGCTGGGCAACATAGC 36
Chimp           ----------------------AGGCCGGGAGTTTGAGATCAGGCTGGGCAACATAGC 36
Orangutan       AGGCCGAGGTGGGAGAAGCACTTGAGGCCAGGAGTTTGAGATCAGGCTGGGCAACATAGC 60
                                                      ****************

Human           GAGACCTCAACTCTACAATTTTTTTTTTTTTAAATTTTAGAGACGGGGTCTTGCTCTGTT 96
Mir-1303        GAGACCTCAACTCTACAATTTTTTTTTTTTTAAATTTTAGAGACGGGGTCTTGCTCTGTT 76
Gorilla         GAGACCTCAACTCTACAATTTTTTTTTTTTTAA-TTTTAGAGACGGGGTCTTGCTCTGTT 95
Chimp           GAGACCTCAACTCTACAATTTTTTTTTTTTTAA-TTTTAGAGACGGGGTCTTGCTCTGTT 95
Orangutan       GAGACCTCAACTCTACAATTTTTTTTTTTTT-AAATTTTAGAGACAGGGTTTTGCTATGTT 119
                ****************************** ** ********** **** ***** ****

Human           GCCAGGCTTTGTGCAGCATGTAAACTGTAC- 126
Mir-1303        GCCAGGCTT---------------------- 85
Gorilla         GCCAGGCTTTGTGCAGCATGTAAACTGTAC- 125
Chimp           GCCAGGCTTTGTGCAGCATGTAAACTGTAC- 125
Orangutan       GCCAGGCTTTGTGCAGCATGTAAACTGTACA 150
                *********
```

MIR-3118-5

```
Gorilla         ACAGGGCACTCTCAGGTGCCCACACATACTACAATAATTTTCATAATGCAATCACACACA 60
Orangutan       ACAGGGCACTCTCAGGTGCCCACACATACTACAACAATTTTCATAATGCAATCACGCACA 60
Chimp           ACAGGGCACTCTCAGGTGCCCACACATACTACAATAATATTCATAATGCAATCACACACA 60
Human           ACAGGGCACTCTCAGGTGCCCACACATAC---AATAATATTCATAATGCAATCACACACA 57
Mir-3118-5      --------------------CACACATAC---AATAATATTCATAATGCAATCACACACA 37
                                    *********   ** *** ***************** ****

Gorilla         ATCACCATGTGACCGCATTATGAAAATTCTTCTAGTGTGA-------------------- 100
Orangutan       ATCACCATGTGACTGCGTTATGAAAATTCTTCTAGTGTGA-------------------- 100
Chimp           ATCACCATGTGACTGCATTATGAAAATTCTTCTAGTGTGA-------------------- 100
Human           ATCACCATGTGACTGCATTATGAAAATTCTTCTAGTGTGA-------------------- 97
Mir-3118-5      ATCACCATGTGACTGCATTATGAAAATTCTTCTAGTGTGGGCTGGGCAACATAGCGAGAC 97
                ************* ** ***********************
```

MIR320B

```
Mir-320A        --------------------CCATCACCAAAACATGGAAGCACTTACTTCTTTAGTTTCA 40
Human           AAATGTAAAAGGAAGACTTACCATCACCAAAACATGGAAGCACTTACTTCTTTAGTTTCA 60
Chimp           AAATGTAAAAGGAAGACTTACCATCACCAAAACATGGAAGCACTTACTTCTTTAGTTTCA 60
Gorilla         AAATGTAAAAGGAAGACTTACCATCACCAAAACATGGAAGCACTTACTTCTTTAGTTTCA 60
Orangutan       AAATGTAAAAGGAAGATGGACCATCACCAAAACATGGAAGAACGTGATTCGGTAGTTTCA 60
                          ******************** ** *  ***  ********

Mir-320A        AAGCAAGTACATCCACGTTTAAGTG------------------GTGG------------- 69
Human           AAGCAAGTACATCCACGTTTAAGTG------------------GTGGGGAGCCCAGTCTT 102
Chimp           AAGCAAGTACATCCACGTTTAAGTG------------------GTGGGGAGCCCAGTCTT 102
Gorilla         AAGCAAGTACATCCACGTTTAAGTG------------------GTGGGGAGCCCAGTCTT 102
Orangutan       AAGCAAGTACATCCNNNNNNNNNNGTACATCCACGTTTAAGTGGTGGGGAGCCCAGTCTT 120
                **************          *                  ****
```

MIR635

```
Mir-635         --------------------ATTGGAAGCTCAATGGACACAAAACAATGATCAGGAGAAG 40
Human           CTGCACTCCATCCTTTTACAATTGGAAGCTCAATGGACACAAAACAATGATCAGGAGAAG 60
Chimp           CTGCACTCCATCCTTTTACAATTGGAAGCTCAATGGACACAAAACAATGATCAGGAGAAG 60
Gorilla         --------------TTTAGAATTGGAAGCTCAATGTAAGCAA--CAATGAAGTAAAGAAA 44
Orangutan       --------------------TTGGAAG-TCAAT----ACAAAACAATGATCAGGAGAGA 34
                          ******* *****     ***  ******     ***

Mir-635         TTTCCATAACAAAGCCTAA-TGGACATTGTTTCAGTGCCCAAGTGGCAGCTCCTCT-CTG 98
Human           TTTCCATAACAAAGCCTAA-TGGACATTGTTTCAGTGCCCAAGTGGCAGCTCCTCT-CTG 118
Chimp           TTTCCATAACAAAGCCTAA-TGGACATTGTTTCAGTGCCCAAGTGGCAGCTCCTCT-CTG 118
Gorilla         -----ACAACTGAATTTCTCTTGAAATTATTT---TCCTTTACTGACTTGTCCTTTATTG 96
Orangutan       T----AAAAGGTAAGATGA--GGAAAGCAT----GGACTGCAAAGGAGGATAGTGTAGTG 84
                     * **   *   *     ** *   *       *   *  *     *  * *  **
```

MIR718

```
Mir-718         --------------------GCCTCGTGCGACGCCCGGCGGGGCGGAAGGGGCGGTGCCC 40
Human           CTCACTTCCCCTTCGAGCCGGCCTCGTGCGACGCCCGGCGGGGCGGAAGGGGCGGTGCCC 60
Chimp           CTCACTTCCCCTTCGAGCCGGCCTCGTGCGACGCCCGGCGGGGCGGAAGGGGCGGTGCCC 60
Gorilla         CTCACTTCCCCTTCGAGCCGGCCTCGTGCGACGCCCGGCGGGGCGGAAGGGGCGGTGCCC 60
Baboon          CTCACTTCCCCTTCGAGCCGGCCTCGTGCGACACCCGGCGGGGCGGAAGGGGCGGTGCCC 60
Orangutan       CTCACTTCCCCTTCGAGCCGGCCTCGTGCGA-GCCTGGCGGGGCGGAAGGGGCGGTGCCC 59
                          ***********  ** ************************

Mir-718         GGGCCCGCCGCCATCTTGCG-CGCCGCGGCC-------------------- 70
Human           GGGCCCGCCGCCATCTTGCG-CGCCGCGGCCGCGCGCTGGGTCACACGGGG 110
Chimp           GGGCCCGCCGCCATCTTGCG-CGCCGCGGCCGCGCGCTGGGCCACACGGGG 110
Gorilla         GGGCCCGCCGCCATCTTGCG-CGCCGCGGCCGCGCGCTGGGCCACACGGGG 110
Baboon          GGGCTCGCCGCCATCTTGCG-CGCCGCGGCCGCGCGCTCGGCCACACGGGG 110
Orangutan       GGGCCCGCCGCCATCTTGCAACGCCGCGGCCGCGCACTGGGCCACACGGGG 110
                **** ***************  ***********
```

MIR1278

```
Mir-1278        --------------------ATTTGCTCATAGATGATATGCATAGTACTCCCAGAACTCA 40
Rhesus          GTATTAACTCATGTAATCTGATTTGCTCATAGATGATATGCACAGTACTCCCAGAACTCA 60
Chimp           GCATTAACTCATGTAATCGTATTTGCTCATAGATGATATGCATAGTACTCCCAGAACTCA 60
Human           GTATTAACTCATGTAATCGTATTTGCTCATAGATGATATGCATAGTACTCCCAGAACTCA 60
Gorilla         GTATTAACTCATGTAATCGTATTTGCTCATAGATGATATGCATAGTACTCCCAGAACTCG 60
Orangutan       GTATTAACTCATGTAATCCTATTTGCTCATAGATGATATGCGTAGTACTCCCAGAACTCA 60
                                     **********************  ****************

Mir-1278        TTAAGTTGGTAGTACTGTGCATATCATCTATGAGCGAATAG-------------------- 81
Rhesus          TTAAGTTGGTAGAACCATGCATGTCATCTATGAGCGAATAGGCTCAGACAGAGT------ 114
Chimp           TTAAGTTGGTAGTACTGTGCATATCATCTGTGAGCGAATAGGCTCAGACAGAGTGAGTTC 120
Human           TTAAGTTGGTAGTACTGTGCATATCATCTATGAGCGAATAGGCTCAGACAGAGTGAGTTC 120
Gorilla         TTAAGTTGGTAGTACTGTGCATATCATCTATGAGCGAATAGGCTCAGACAGAGTGAGTTC 120
Orangutan       TTAAGTTGGTAGTACTG--------------TGAGCGAATAGGCTCAGACAGAGTGAGTTC 107
                ************ **               ***********
```

MIR2117

```
Mir-2117        --------------------GCTCTGATTTACTTCTGTCCGGCATGGTGAACAGCAGGAT 40
Human           CTTAAGGAACTGGAGAGAATGCTCTGATTTACTTCTGTCCGGCATGGTGAACAGCAGGAT 60
Chimp           CTTAAGGAACTGGAGAGAATGCTCTGATTTACTTCTGTCTGGCCTGGTGAACAGCAGGAA 60
Gorilla         CTTAAGGAACTGGAGAGAATGCTCTGATTTACTTCTGTCCGGCATGGTGAACAGCAGGAA 60
Orangutan       CTTAAGGAACTGGAGAGAATGCTCTGATTTACTTCTATGCGGCATGGTGAACAGCAGGAA 60
Rhesus          CTTAAGGAACTGGAGAGAATGTTCTGATCTGCATCTGTCCGACATGGTAAACAGCAGGAA 60
                          * ****** * * *** *  * * **** **********

Mir-2117        TGGCTGTAGCTGTTCTCTTTGCCAAGGACAGATCTGATCT-------------------- 80
Human           TGGCTGTAGCTGTTCTCTTTGCCAAGGACAGATCTGATCTGATTGCCTTGAGGTGAAAGT 120
Chimp           TGGCTGTAGCTGTTCTCTTTGCCAAGGACAGATCTGATCTGATTGCCTTGAGGTGAAAGT 120
Gorilla         TGACTGTAGCTGTTCTCTTTGCCAAGGACAGATCTGATCTGATTGCCTTGAGGTGAAAGT 120
Orangutan       TGACTGTAGCTGTTTTTTATGC-AAGGACAGATCTGATCTGATTGCCTTGAGGTGAAAGT 119
Rhesus          TGACTGTAGCTGTTCTCTTTGCCAAGGACAGATCTGATCTGATTGCCTTGAGGTGAAAGT 120
                ** ************ * * *** ******************
```

MIR3653

```
Chimp          GACCTACTGGCCCCCAGGCTTCCTTGGGGAAGCAGCCCCTTCAGTCAACTTCTTAGAGGC 60
Gorilla        GACCTACTGGCCCCCAGGCTTCCTTGGGGAAGCAGCCCCTTCAGTCAACTTCTTAGAGGC 60
Human          GACCTACTGGCCCCCAGGCTTCCTTGGGGAAGCAGCCCCTTCAGTCAACTTCTTAGAGGC 60
Orangutan      GACCTACTGGCCCCCAGGCTTCCTTGGGGAAGCAGCCCCTTCAGTCAACTTCTTAGAGGC 60
Mir-3653       --------------------TCCTTGGGGAAGCAGCCCCTTCAGTCAACTTCTTAGAGGC 40
Rhesus         GACCTACTGGCCTCCAGGCTTCCCTGGGGAAGCAGCCCCTTCAGTCAACTTCTTAGAGGC 60
                           *** ************************************

Chimp          TCAGTTTGCTCATCATGAGCGTGCTCAGGAAGAAGAATCATCAGGAGGGGCTGCCAG--- 117
Gorilla        TCAGTTTGCTCATCATGAGCGTGCTCAGGAAGAAGAATCATCAGGAGGGGCTGCCAG--- 117
Human          TCAGTTTGCTCATCATGAGCGTGCTCAGGAAGAAGAATCATCAGGAGGGGCTGCCAG--- 117
Orangutan      TCAGTTTGCTCATCATGAGCGTGCTCAGGAAGAAGAATCATCAGGAGGGGCTGCCAGCAG 120
Mir-3653       TCAGTTTGCTCATCATGAGCGTGCTCAGGAAGAAGAATCATCAGGAGGGGCTGCCAG--- 97
Rhesus         TCAGTTTGCTCATCATGAGCGTGCTCAGGAAGAAGAATCATCAGGAGGGGCTGCCAG--- 117
               ***********************************************************

Chimp          --------------------------GGGTCCCCAG-GGAGTCCTAGCCTCACTGG-CT 148
Gorilla        --------------------------GGGTCCCCAG-GGAGTCCTAGCCTCACTGG-CT 148
Human          --------------------------GGGTCCCCAG-GGAGTCCTAGCCTCACTGG-CT 148
Orangutan      GAAGAAGAATCATCAGGAGGGCTGCAGGGGTCCCCAAAGGAGCCCTAGCCTCACTGGGCT 180
Mir-3653       --------------------------GGGTCCCCAG-GGA------------------- 110
Rhesus         --------------------------GGGTCCCCAG-GGAGCCCTAGCTTCAC------ 143
                                         *********  ***
```

MIR1825

```
Mir-1825       --------------------AGAGACTGGGGTGCTGGGCTCCCCTAGACTAGGACTCCAG 40
Rhesus         TGTTA-CAGTGTTGCAGGTGAGAGACTGGGGTGCTGGGCACCCCTAGACTAGGACTCCAG 59
Human          TGTTAACAGGGTTGCAGGCGAGAGACTGGGGTGCTGGGCTCCCCTAGACTAGGACTCCAG 60
Chimp          TGTTAACAGGGTTGCAGGCGAGAGACTGGGGTGCTGGGCTCCCCTAGACTAGGACTCCAG 60
Gorilla        TGTTAACAGGGCTGCAGGCAAGAGACTGGGGTGCTGGGCTCCCCTAGACTAGGACTCCAG 60
Orangutan      TGTTA-CAGGGTTGCAGGCGAGAGACTAGAGTGCTGGGCTCCCCTAGACTAGGACTCCAG 59
                             ******* * ********* ********************

Mir-1825       TGCCCTCCTCTCC-------------------- 53
Rhesus         TGCCCTCCTCTCCCAAGAGACAAAGGCCATTGC 92
Human          TGCCCTCCTCTCCCAAGAGACAAAGGCCATTGC 93
Chimp          TGCCCTCCTCTCCCAAGAGACAAAGGCCATTGC 93
Gorilla        TGCCCTTCTCTCCCAAGAGACAAAGGCCATTGC 93
Orangutan      TGCCCT---CTCCCAAGAGACAAAGGCCATTGC 89
               ******   ****
```

MIR1272

```
Mir-1272        --------------------TACCAGCCCAGACCACAGAGCACACGAGCTGCAGCCTAC 39
Human           -GGAGCAGGCCCAGCCATCCCTACCAGCCCAGACCACAGAGCACACGAGCTGCAGCCTAC 59
Chimp           GGAGCAGGGCCCAGCCATCCCTACCAGCCCAGACCACAGAGCACACGAGCTGCAGCCTAC 60
Gorilla         GGAGCAGGGCCCAGCCATCCCTACCAGCCCAGACCACAGAGCGCACGAGCTGCAGCCTAC 60
Orangutan       ---------------------------------ACAGAGCACATGAGCTGCAGC-TAC 24
Rhesus          -------GGCCCAGCCATCCCTACCAGCCCAGACCACAGAGCACATGAGCTGCAGCCTAC 53
                                                ******* ** ********** ***

Mir-1272        ATTTCTAAGTTTACGGCTTTCCTGTTATAAAGACACTGAGCACGTTTTCAGAATTTGCTG 99
Human           ATTTCTAAGTTTACGGCTTTCCTGTTATAAAGACACTGAGCACGTTTTCAGAATTTGCTG 119
Chimp           ATTTCTAAGTTTACGACTTTCCTGTTATAAAGACACTGAGCACGTTTTCAGAATTTGCTG 120
Gorilla         ATTTCTAAGTTTACGGCTTTCCTTTTATAAAGACACCAAGCATGTTTTCAGAATTTGCTG 120
Orangutan       ATTCT--AGTTTACAGCTTTC-TGTTATAA-GACACTGAACATGTTTT--AGATTTGCTG 78
Rhesus          ATTTCTAAGTTTACAGCTTTCCTGTTATAAAGACACTGAACAGGTTTTTAGAATTTGCTG 113
                ***    *******  ***** * ****** *****  * ** *****    ********

Mir-1272        CCATCATCATCGCACCCAGATCTGATCTGG-------------------- 129
Human           CCATCATCATCGCACCCAGATCTGATCTGGCTTCCCCAAGGGAAAGAGAC 169
Chimp           CCATCATCATCGCACCCAGATCTGATCTGGCTTCCCCAAGGGAAAGAGAC 170
Gorilla         CCATCATCATCGCACCCAGATCTGATCCGGCTCCCCCAAGGGAAAGAGAC 170
Orangutan       CCATCATCATCGCACCCAGATCTGATCTG--CTCCCCAAGGGAAAGAGAC 126
Rhesus          CCATCATCGTCACACCCAGATCCGATCTGGCTCCCCCAAGGGAAAGAGA- 162
                ******** ** ********** **** *
```

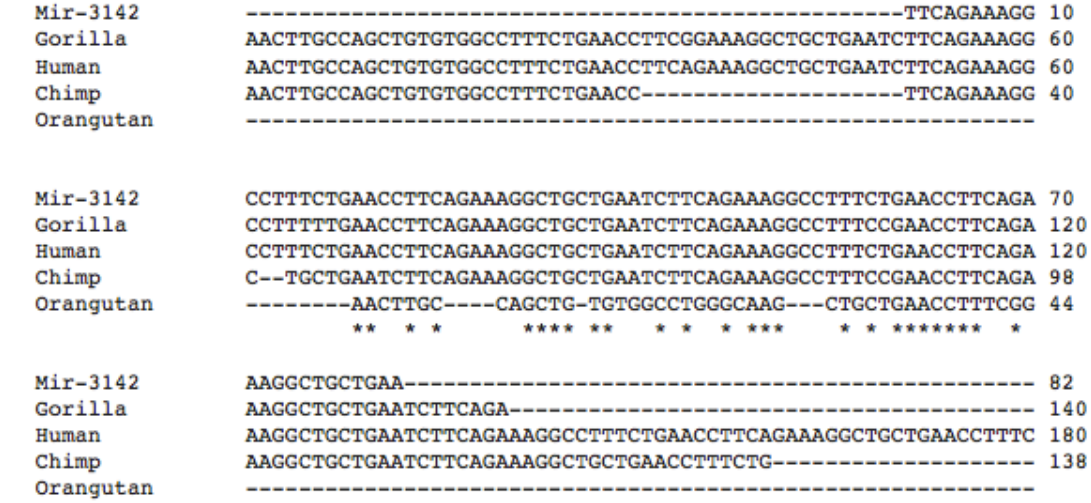
MIR935

```
Human           GCTCGGACCTGCTCAAGGCCGGCGGGG-----GCGCGGGCGGCAGTGGCG--GGAGCGGC 53
Chimp           GCTCGGACCTGCTCAAGGCCGGCGGGG-----GCGCGGGCGGCAGTGGCG--GGAGCGGC 53
Mir-935         --------------------GGCGGGG-----GCGCGGGCGGCAGTGGCG--GGAGCGGC 33
Gorilla         --------------------GGCGGGG-----GTGCGGGCGGCAGTGGCG--GGA--GAA 31
Orangutan       ------------------CCGCCGCCGCTCCCGCTCAGCCAGCTCCAGCTCCGGCCAAGA 42
                                  * **  *     *  * * * **    **   **

Human           CCCTCG-GCCATCCTCCGTCTGCCCAGTTAC-CGCTTCCGCTACCGCCGCCGCTCCCGC- 110
Chimp           CCCTCG-GCCATCCTCCGTCTGCCCAGTTAC-CGCTTCCGCTACCGCCGCCGCTCCCGC- 110
Mir-935         CCCTCG-GCCATCCTCCGTCTGCCCAGTTAC-CGCTTCCGCTACCGCCGCCGCTCCCGC- 90
Gorilla         CCCTGG-GCGG-----CGGCTGCGGAGGCGGGCGGCCCAGCCAGAGCCGGTGTCTCTGCG 85
Orangutan       CTCTGACACCG-----CTGCCGCCGCCATGG-CGCCTCAGCTGCTGGCAG-GCACCCACC 95
                * **    *       *  * **         **   * **    * *   *   *  *

Human           TCTAGCTCCCGCTCCAGCGAG 131
Chimp           TCTAGCTCCCGCTCCAGCGAG 131
Mir-935         T-------------------- 91
Gorilla         CCTTATAAGGGCTTC------ 100
Orangutan       CACCA---------------- 100
```

MIR1470

MIR3130

MIR3142

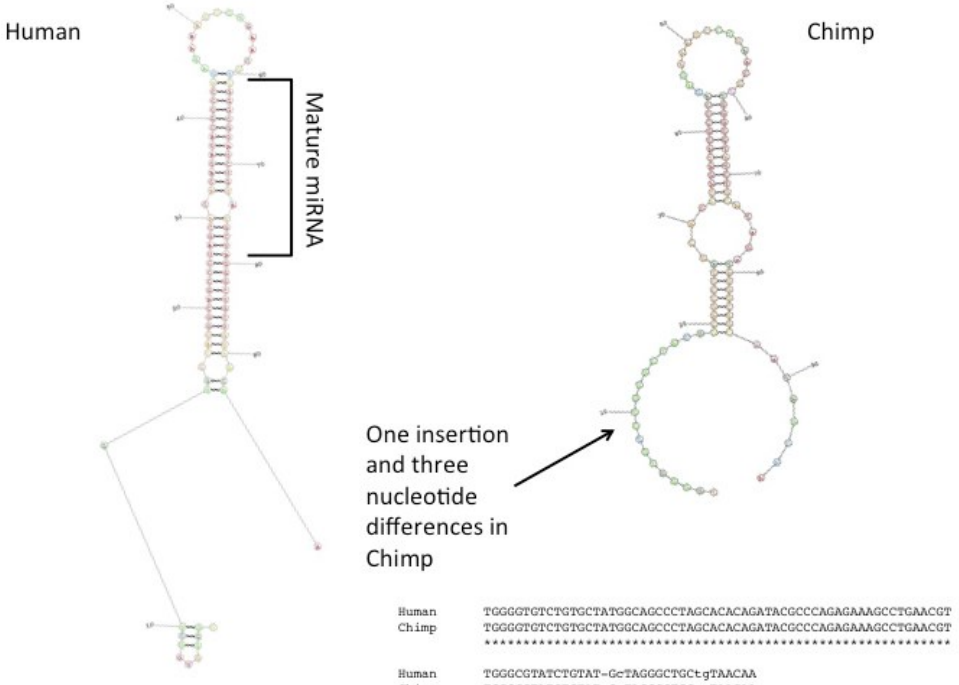


MIR1283B

```
ppy-mir-1283B      CTCATGCTGTGACTCTGCAAAGGGAAGCCCTTTCTGTTGTCTAAAAGAAAAGAA------ 54
Chimp              CTCGTGCTGTGACTCTGCAAAGGGAAGCCCTTTCTGTTGTCTAAAAGAATAGAA------ 54
Gorilla            CTCATGCTGTGAGTCTGCAAAGGGAAGCCCTTTCTGTTGTCTGAAAGAAAAGAA------ 54
Human              CTCAGGCTGTGACCCTCCAAAGGGAAGAACTTTCTGTTGTCTAAAAGAAAAGAACGCACT 60
                   ***  *******  ** **********  ************* ****** ****

ppy-mir-1283B      ----------------------------------------------AGCGCTTCCCT 65
Chimp              ----------------------------------------------AGCGCTTCCCT 65
Gorilla            ----------------------------------------------ATCGCTTCCCT 65
Human              TCCCTTTAGAGTGTTACCGTGTGAGAAAAGCAACGTTGAAAGAAAAGAAATCGCTTCCCT 120
                                                                 * *********

ppy-mir-1283B      TTGGAGTGTTACGGTTTGAGAA 87
Chimp              TTGGAGTGTTACGGTTTGAGAA 87
Gorilla            TTGGAGTGTTACGGTTTGAGAA 87
Human              TTGGAGTGTTACGGTTTGAGAA 142
                   **********************
```

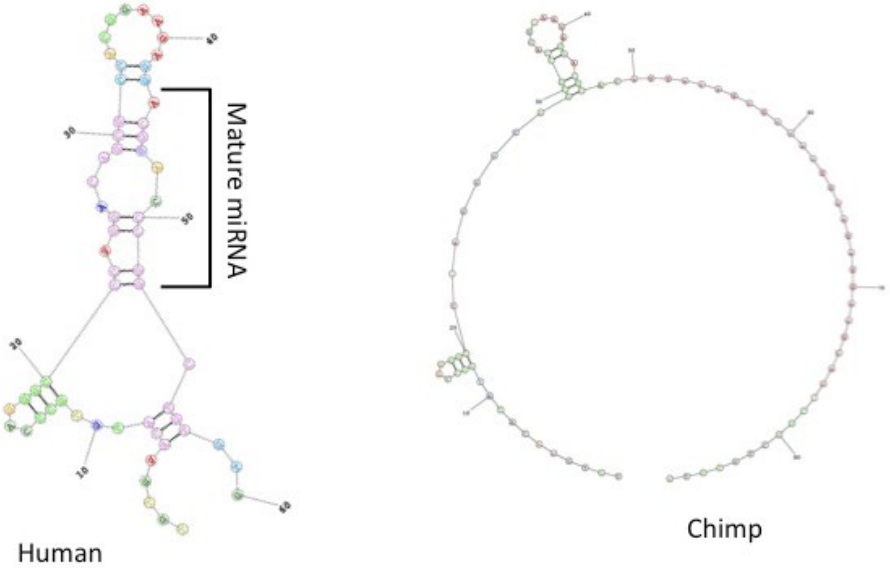
3.2 Human-Chimpanzee miRNA structure comparison

hsa-mir-585 and its chimpanzee homologous sequence



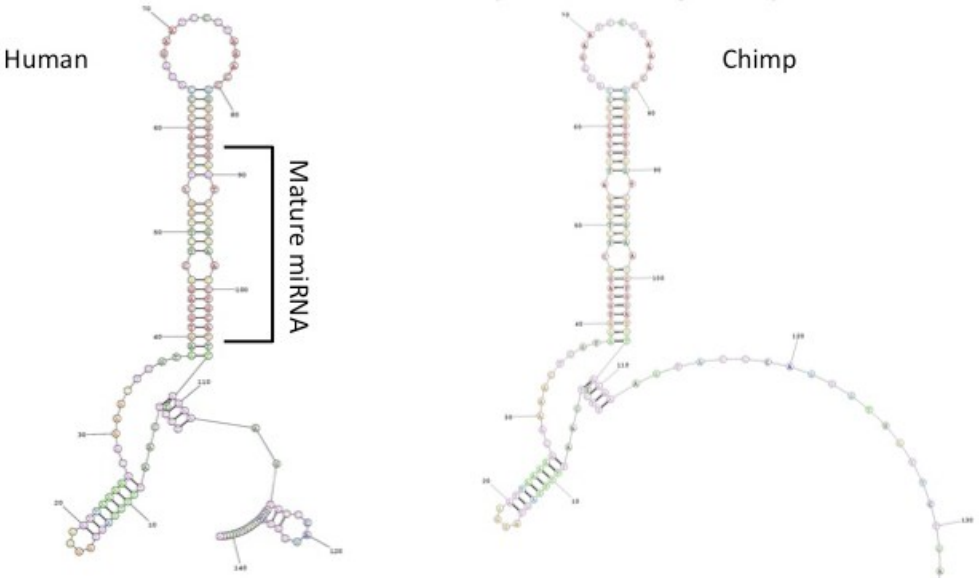

```
Human     TGGGGTGTCTGTGCTATGGCAGCCCTAGCACACAGATACGCCCAGAGAAAGCCTGAACGT
Chimp     TGGGGTGTCTGTGCTATGGCAGCCCTAGCACACAGATACGCCCAGAGAAAGCCTGAACGT
          ************************************************************

Human     TGGGCGTATCTGTAT-GcTAGGGCTGCtgTAACAA
Chimp     TGGGCGTATCTGTATaGaTAGGGCTGCcaTAACAA
          *************** * *********  ******
```

hsa-mir-941 and its chimpanzee homologous sequence



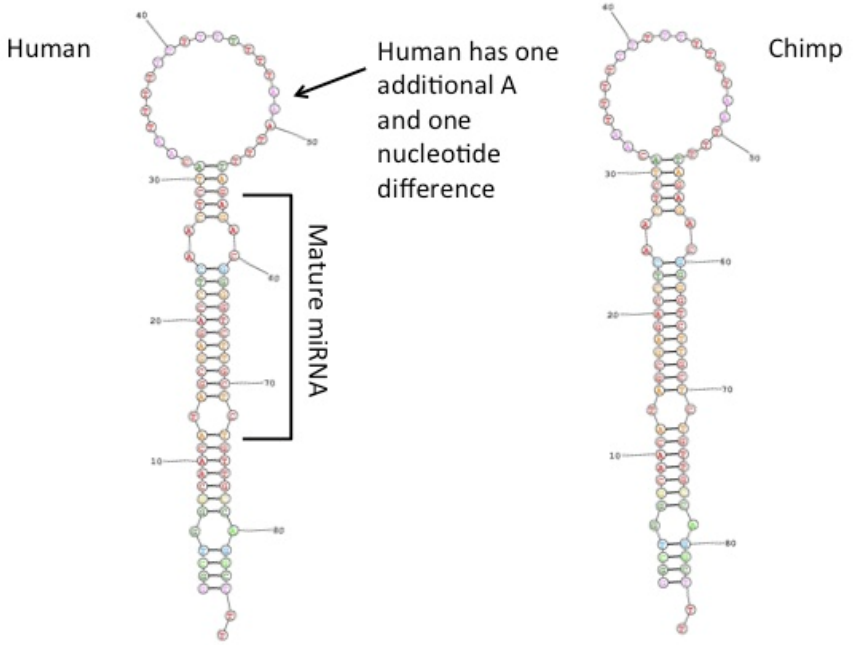

```
Human     TGTGgACATGTGCCCAGGGCCCgGGACAGCGCCACGGAAGAGGACGCAC----------
Chimp     TGTGcACATGTGCCCAGGGCCCaGGACAGCGCCACGGAAGAGGACGCACaggacagcgcc
          **** ****************** ***************************

Human     ----------------CCGGCTGTGTG
Chimp     acggaagaggacgtacCCGGCTGTGTG
                          ***********
```

hsa-mir-1289 and its chimpanzee homologous sequence



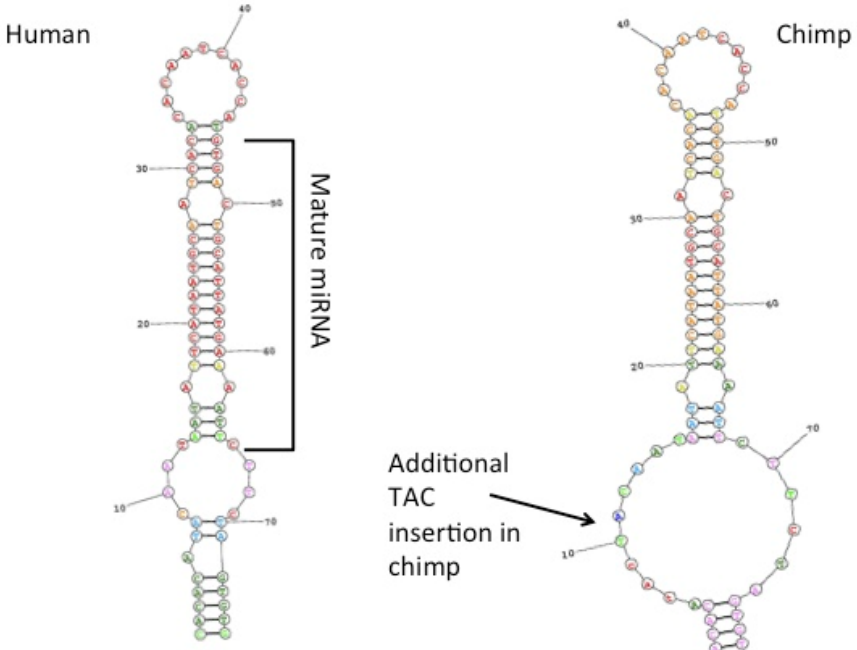

```
Human     TTCTCAATTTTTAGTAGGAATTAAAAACAAAACTGGTAAATGCAGACTCTTGGtTTCCAC
Chimp     TTCTCAATTTTTAGTAGGAATTAAAAACAAAACTGGTAAATGCAGACTCTTGGaTTCCAC
          ********************************************************* ******

Human     CCCCAGAGAATCCCTAAACCGGGGGGTGGAGTCCAGGAATCTGCATTTTAGAAAGTACCCA
Chimp     CCCCAGAGAATCCCTAAACCGGGGGGTGGAGTCCAGGAATCTGCATTTTAGAAAGTACCCA
          ************************************************************

Human     GGGTGATTCTGATA-ATTGGGAACA
Chimp     GGGTGATTCTGATAcATTGGGAACA
          ************** **********
```

hsa-mir-1303 and its chimpanzee homologous sequence



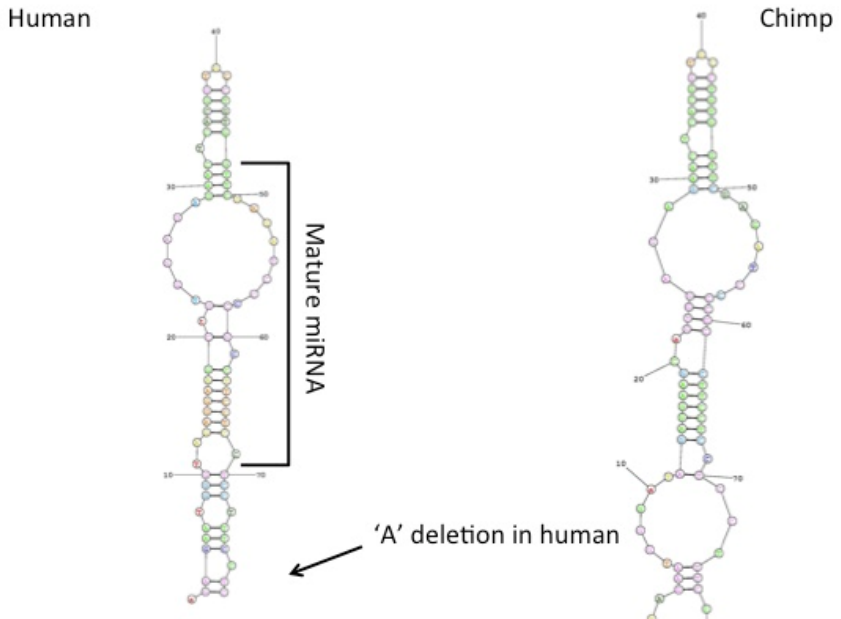

```
Human     GGCTGGGCAACATAGCGAGACCTCAACTCTACAATTTTTTTTTTTTTAAaTTTTAGAGAC
Chimp     GGCTGGGCAACATAGCGAGACCTCAACTCTACAATTTTTTTTTT-TTTAATTTTAGAGAC
          ********************************************* ** ************

Human     GGGGTCTTGCTCTGTTGCCAGGCTTT
Chimp     GGGGTCTTGCTCTGTTGCCAGGCTTT
          **************************
```

hsa-mir-3118-5 and its chimpanzee homologous sequence



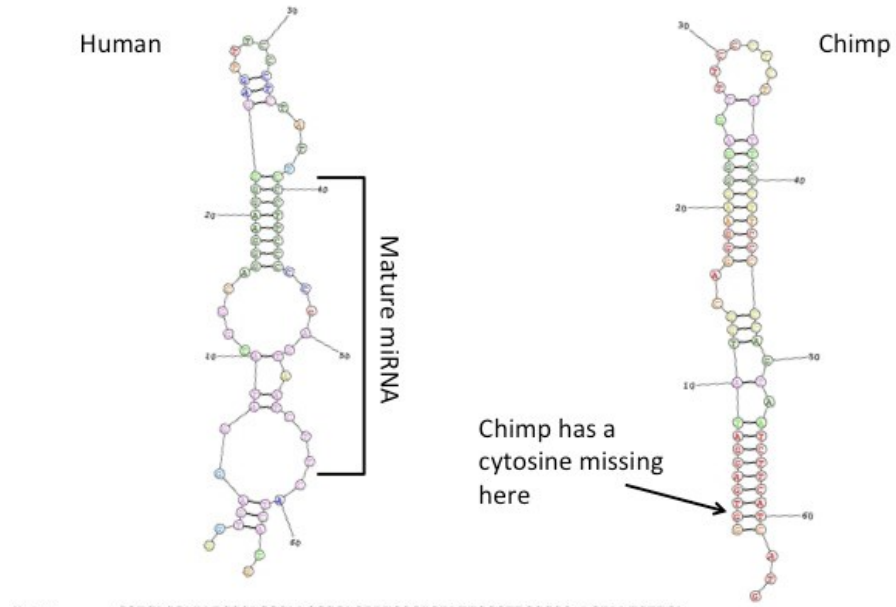

```
Human       CACACA---TACAATAATATTCATAATGCAATCACACACAATCACCATGTGACTGCATTA
Chimp       CACACAtacTACAATAATATTCATAATGCAATCACACACAATCACCATGTGACTGCATTA
            ******   **************************************************

Human       TGAAAATTCTTCTAGTGTG
Chimp       TGAAAATTCTTCTAGTGTG
            ********************
```

hsa-mir-3125 and its chimpanzee homologous sequence



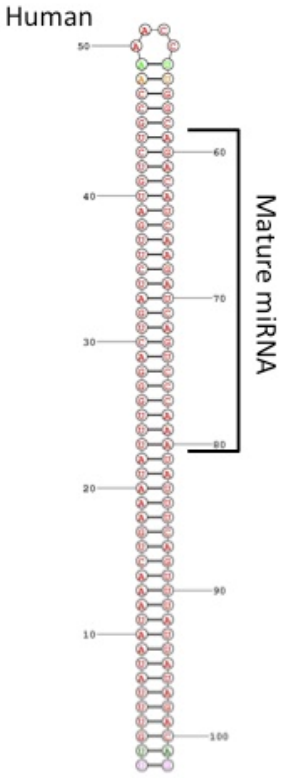

```
Human       aGAGAATGGG-TAGAGGAAGCtGTGGAGAGAACTCACGGTGCCTGTGGTTCGAGATCCCCG
Chimp       -GAGAATGGGaTAGAGGAAGCaGTGGAGAGAACTCACGGTGCCTGTGGTTCGAGATCCCCG
             ********* ********** ****************************************

Human       CCTTCCTCCTCCTTTCCTC
Chimp       CCTTCCTCCTCCTTTCCTC
            ********************
```

hsa-mir-3679 and its chimpanzee homologous sequence




```
Human     GGTGAGGATATGGCAGGGAAGGGGAGTTTCCCTCTATTCCCTTCCCCCcAGTAATCTTCA
Chimp     GGTGAGGATATGGCAGGGAAGGGGAGTTTCCCTCTATTCCCTTCCCCC-AGTAATCTTCA
          ********************************************** ***********

Human     TCATG
Chimp     TCATG
          *****
```

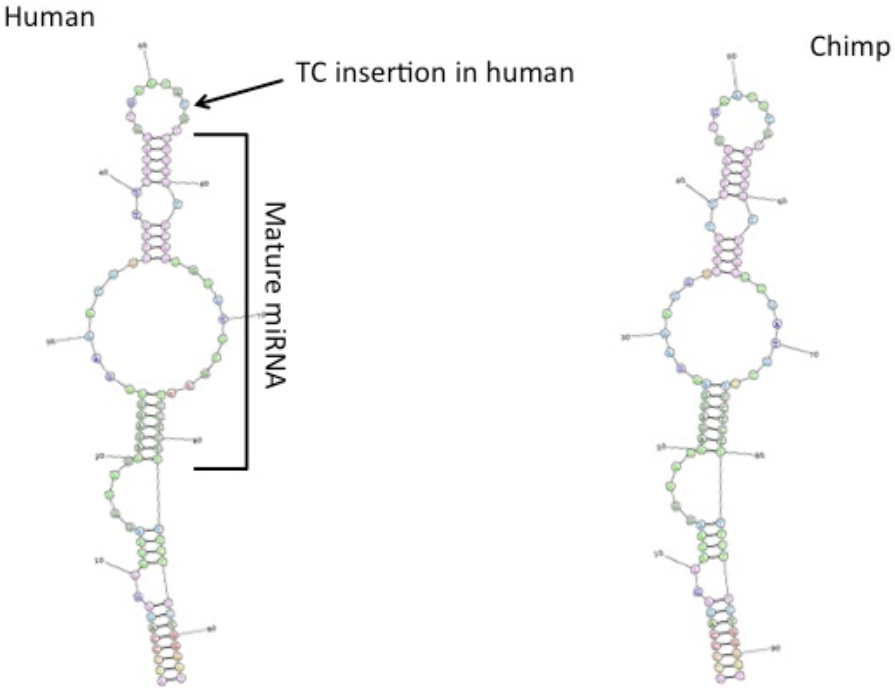
hsa-mir-3913 and its chimpanzee homologous sequence



Chimp

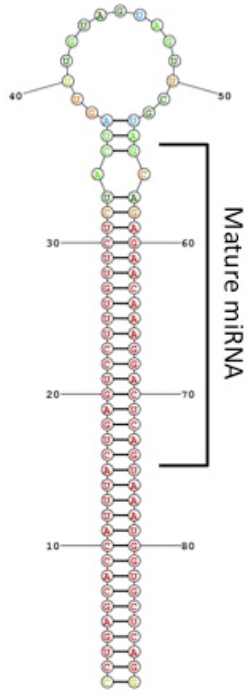
NA: Sequence missing from genome

hsa-mir-3916 and its chimpanzee homologous sequence




```
Human       ATCCCAGAGAAGAAGGAAGAAGAGGAAGAAATGGCTGGTTCTCAGGTGAATGTGTCTGGG
Chimp       ATCCCAGAGAAGAAGGAAGAAGAGGAAGAAATGGCTGGTTCTCAGGTGAATGTGTCTGGG
            ************************************************************

Human       TTCAGGGGATGTGtcTCCTCTTTTCTTCTGGGAT
Chimp       TTCAGGGGATGTGTC--CTCTTTTCTTCTGGGAT
            ***************  ******************
```

hsa-mir-3919 and its chimpanzee homologous sequence

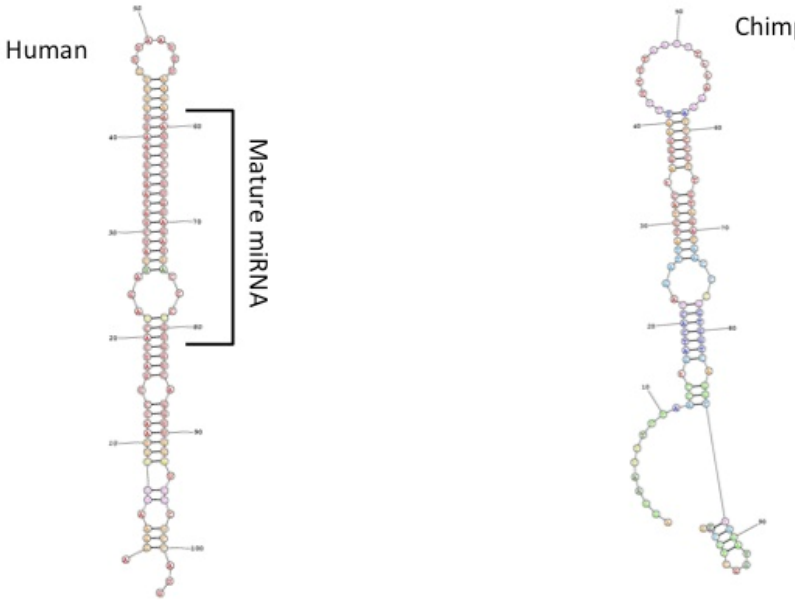


Human

Chimp

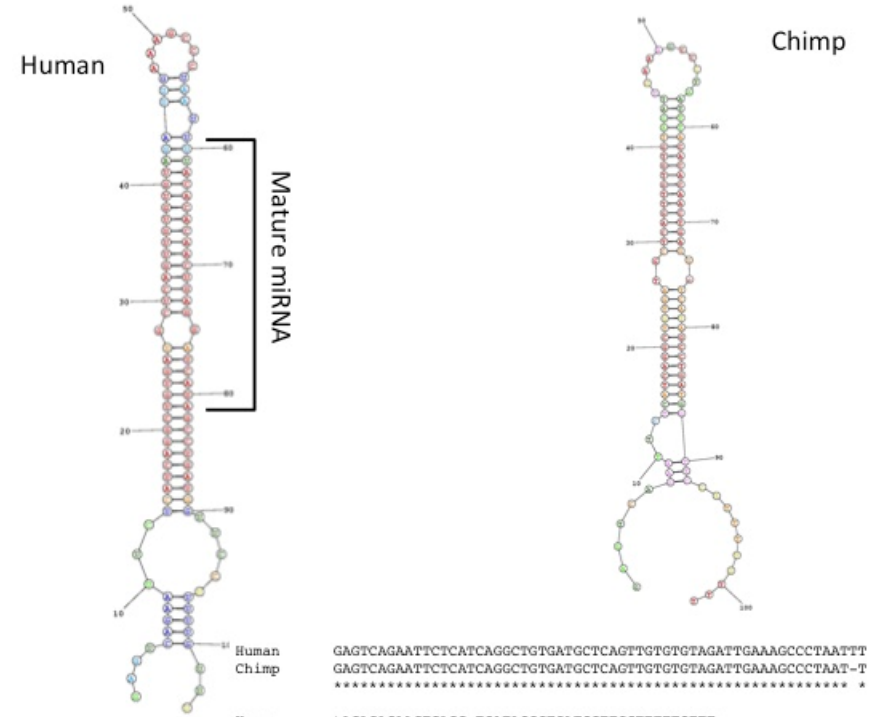
NA: Sequence missing from genome

hsa-mir-3938 and its chimpanzee homologous sequence



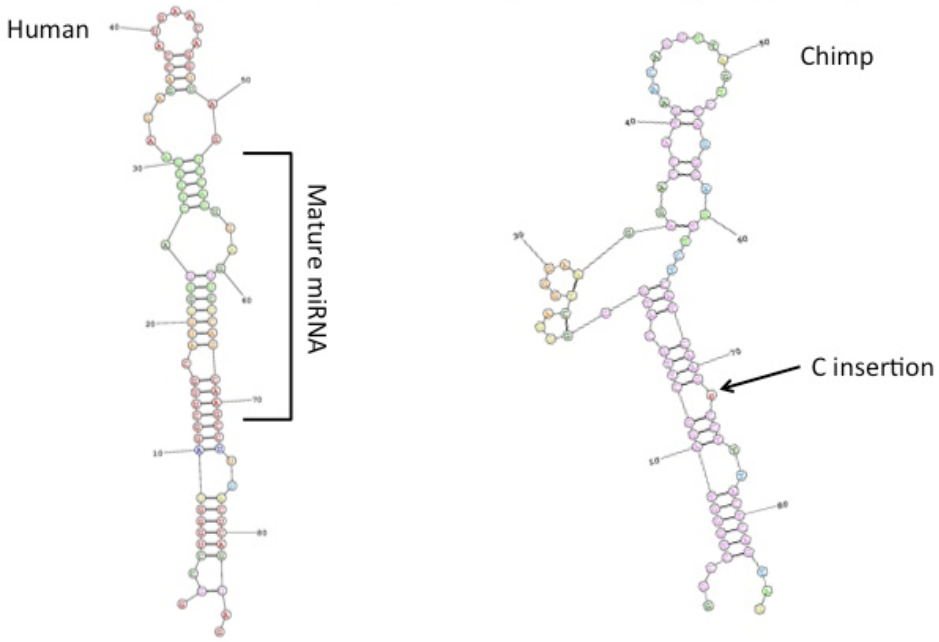

```
Human       AGGAATTTTTAACCcGATCACTAGATTATCTACaAGGGAATTTTTTTttAATTTaaAAAA
Chimp       AGGAATTTTTAACCtGATCACTAGATTATCTACtAGGGAATTTT--TTTAATTTttAAAA
            ************** ****************** **********  ********  ****

Human       TTCCCTTGTAGATAACCCGGTGGTCAGGTTGGATGGCTCCATG
Chimp       TTCCCTTGTAGATAACCCGGTGGTCAGGTTGGATGGCTCCATG
            *******************************************
```

hsa-mir-3941 and its chimpanzee homologous sequence



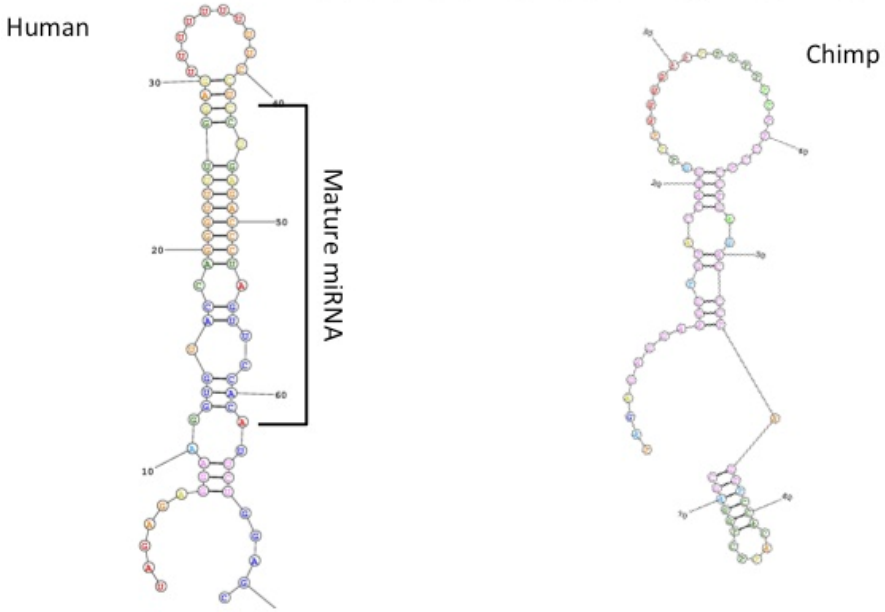

```
Human       GAGTCAGAATTCTCATCAGGCTGTGATGCTCAGTTGTGTGTAGATTGAAAGCCCTAATTT
Chimp       GAGTCAGAATTCTCATCAGGCTGTGATGCTCAGTTGTGTGTAGATTGAAAGCCCTAAT-T
            ************************************************************ *

Human       tACACACAACTGAGGaTCATAGCCTGATGGTTCCTTTTTGTTT
Chimp       TACACACAACTGAGGcTCATAGCCTGATGGTTCCTTTTTGTTT
            *************** ****************************
```

hsa-mir-4327 and its chimpanzee homologous sequence




```
Human       GGCCTGGGTAGGCTTGCATGGGGGACTGGGAAGAGACCATGAACAGGTTAGTCCAGGGAG
Chimp       GGCCTGGGTAGGCTTGCATGGGGGACTGGGAAGAGACCATGAACAGGTTAGTCCAGGGAG
            ************************************************************

Human       TTCTCATCAAG-CCTTTACTCAGTAG
Chimp       TTCTCATCAAGcCCTTTACTCAGTAG
            *********** **************
```

hsa-mir-4329 and its chimpanzee homologous sequence




```
Human       TAGAGAGGAAGGTGTACCAGGGTTTtgg-agTTTTTTTTTCCTCCTGAGACCCTAGTTCC
Chimp       TAGAGAGGAAGGTGTACCAGGGTTTggggttTTTTTTTTTCCTCCTGAGACCCTAGTTCC
            ************************* **   ******************************

Human       ACATTCTGGAGC
Chimp       ACATTCTGGAGC
            ************
```