Comparative genomics of indels in primate lineages and the possible effect of the *MET* promoter indel on gene expression

by

Frances Marie Marín Maldonado

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
BIOLOGY

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2016

Approved by:

_____	_____
Nanette Diffoot Carlo, Ph.D. Member, Graduate Committee	Date
_____	_____
Taras K. Oleksyk, Ph. D. Member, Graduate Committee	Date
_____	_____
Carlos Rodríguez Minguela, Ph. D. Member, Graduate Committee	Date
_____	_____
Juan C. Martínez Cruzado, Ph. D. President, Graduate Committee	Date
_____	
Arnaldo Carrasquillo, Ph. D. Representative of Graduate Studies	_____ Date
_____	
Matías Cafaro, Ph.D. Chairperson of the Department	_____ Date

Abstract

Approximately six million years ago, the human and chimpanzee lineages diverged from a common ancestor resulting in two distinct species with a great number of different morphological, behavioral, cognitive and other phenotypic traits. However, their genomes are more than 98.5% identical at protein-coding loci, and genomic differences between these two species hover around 2.4%, excluding repeats and low complexity DNA and including insertions and deletions (indels). It is believed that most of the genetic foundation for the differences among these two lineages lies at the level of gene regulation. Our work focused on lineage-specific large (> 10 bp) indels located in promoter regions that may affect gene expression and protein product levels. Our goal was to validate indels, in order to differentiate real features from computational artifacts. We started by identifying 64 indels, through the alignment of orthologous regions of human, chimpanzee, gorilla, orangutan and macaque (*in-silico* analysis), located at the 5' side of the gene and at distance no more than 2 kb away from its nearest gene transcription start site. To validate the indels, PCR and electrophoresis (*in-vitro*) analysis was performed with all of them. We found two distinct indels, which were unique to the human lineage, in genes associated with neurodevelopment and the female-sexual development, *MET* and *DMRTA2*. Since previous studies suggest that most of the differences that exist between humans and chimpanzees are in areas related to the cognitive ability and fertility, our results could indicate that the derived variants observed in the human lineage might be important for processes that make the humans distinct to the other hominids. Moreover, given that most of the divergences that exist between humans and chimpanzees are due to differential gene expression, the promoters described herein could serve as models for *in-vitro* gene expression assays for evaluating how these fixed indels may affect gene expression in both species.

Resumen

Hace aproximadamente 6 millones de años, el linaje de los humanos y el chimpancé divergieron de un ancestro en común, resultando en dos distintas especies con una gran cantidad de diferencias a nivel morfológico, cognitivo y de comportamiento, entre otras características fenotípicas. Sin embargo, sus genomas son más de un 98.5% idénticos en loci codificantes para proteínas, y las diferencias genómicas entre ambas especies son el 2.4%, excluyendo regiones repetitivas y de baja complejidad, e incluyendo inserciones y deleciones (indels). Por lo tanto, se cree que la base genética para las diferencias observadas entre ambas especies subyace a nivel de regulación génica. Nuestro trabajo se enfocó en indels (>10 bp) localizados en regiones promotoras que pueden afectar la expresión de los genes y la cantidad de productos proteicos. Nuestra meta fue validar la presencia de indels, para así diferenciar entre artefactos computacionales o indels reales. Se comenzó identificando 64 indels, a través de alineamientos de secuencias de ADN de regiones ortólogas en humanos, chimpancé, gorila, orangután y macaco (análisis *in-silico*), localizados en el lado 5' a una distancia de no más de 2 kilo bases del comienzo del sitio de transcripción más cercano. Se llevó a cabo la validación de los indels a través de su amplificación mediante PCR y electroforesis (análisis *in-vitro*). Se encontraron dos indels que eran únicos para el linaje de los humanos, ambos en genes asociados al desarrollo cognitivo y desarrollo sexual femenino, *MET* y *DMRTA2*. Dado que estudios previos sugieren que la mayoría de las diferencias fenotípicas que existen entre el humano y el chimpancé se observan en áreas relacionadas a la habilidad cognitiva y la fertilidad, nuestros resultados podrían indicar que las variantes derivadas observadas en el linaje de los humanos es importante para aquellos procesos que hacen a los humanos distintos de los otros homínidos. En adición, puesto que la mayoría de las divergencias que existe entre el humanos y el chimpancé se debe a expresión génica diferencial, los promotores previamente descritos podrían servir como modelos para ensayos de expresión *in-vitro,* para evaluar cómo estos indels fijados podrían afectar la expresión de los genes en ambas especies.

Acknowledgments

I would like to express my gratitude to my mentor, Dr. Juan C. Martínez Cruzado, at the University of Puerto Rico at Mayaguez, I don't know where I would be without the opportunity you gave me to work at your laboratory, back in 2013. I have acquired so much knowledge in so many different areas that I could never imagine. Thanks for the opportunity to work at your laboratory! In addition, I would like to thank to Dr. Taras Oleksyk for providing me the research topic for the thesis. I remembered you were excited when I joined the laboratory and you provided me the topic very quickly. Thanks for your mentorship and help.

In addition, I would like to thanks to all my friends that were supporting me during the entire journey of my studies. I am really grateful to Jaime Torres, Pedro Vázquez, Viviana Román and Ingrid Rivera, for all the great moments that we spent together and all the memories that we made, I cherish them very much. Thanks for being there when I needed most!

Special thanks, to Irimar Torres, the Laboratory Technician of Virology, you were incredible. I am ever grateful for your advice; for teaching me all the necessary skills to work with cell cultures and for letting me use your facilities and equipment every time I needed them. In addition, thank you for your friendship and your counseling, it was invaluable to me. In addition, I would like to express my gratitude to Dr. Nanette Diffoot for all the mentorship provided during the progress of my research and for letting me use the facilities of your laboratory.

Moreover, I am really grateful to Dr. Carlos Rodríguez, for sharing with me droplets of knowledge every time I needed help with the research, thank you!

Also, I would like to express my gratitude to the following undergraduate students; Zahudy Figueroa, Coralys Soto and Roberto Cazaldilla, at the University of Puerto Rico at Mayaguez, who helped me perform the indel validation through PCR and electrophoresis analysis. In addition, I would like to thank to Gabriel Fernández, an undergraduate student at the

University of Mayaguez, who has been an incredible help regarding the cell culture and the luciferase expression assay. You have been an excellent student and your contribution was essential for my success.

Likewise, I would like to thank to my Laboratory Team, for the contribution to this project. There were several people that gave me different kinds of advice during the progress of my research. Special thanks to Wilfried Guiblet for your help, you have taught me a lot, specifically for the indel validation part of the thesis. Thanks, to Angelia Caro for helping me during the summer of 2013, with the process of indel validation through PCR and electrophoresis analysis. Also, express my gratitude to Yashira Afanador, your help was crucial for the progress of my research, you were always there for advice when I needed the most, I will be forever grateful. In addition, special thanks to Joseline Serrano, Priscila Rodriguez and Nikole Ayala, for all the moments that we spent together and your friendship.

Additionally, I would like to thank the 2012 HHMI Undergraduate Science Program and the Enhancing Advanced Educational Opportunities in STEM Fields for Minority Students at UPR-M, for making it financially possible for me to carry out the research.

Lastly, but not less important I am really grateful to my family, for all the sacrifices that you have made during all of these years, in order to see me succeed; none of this would have been possible if not were for your instruction and support. I am grateful to my dog Plin Plin, who has been invaluable part of our family and provided me a lot of joy whenever I visited home. To my boyfriend Saul Dastas, for all the time that you were giving me support and bringing me motivation to not give up; to believe in myself and stay positive during the hard times; thanks for everything. I hope one day, I would make my contribution to my community. Finally, thanks to God for giving me the strength to continue during this long journey.

To the memory of my grandparents, hope I have made you proud.

To the science, hope I have made my contribution.

List of Tables

Table 1 Composition and Conditions of the PCR Reactions……...……7
Table 2: Indel Coordinates…….8-10
Table 3 Results of indel validation……...10-19
Table 4 Results of indel validation categorized…….....20
Table 5 Chi-Square Goodness of Fit test for the human genome…….……27
Table 6 Chi-Square Goodness of Fit test for the chimpanzee genome…….……29

List of Figures

Figure 1 *In -silico* analysis of orthologous BRWD1-DNA promoter sequences ……..…21
Figure 2 Electrophoresis Validation of *DMRTA2*-indel in promoter region…….22
Figure 3 Electrophoresis Validation of *MET*-indel in promoter region…….23
Figure 4 Pie Chart of the Biological processes attributed to promoters that presented a deletion in the chimpanzee lineage……. 24
Figure 5 Pie Chart of the biological processes of the human genome……...30
Figure 6 Pie Chart of the biological processes of the chimpanzee genome ……...…...31

Table of Contents

Abstract ii
Resumen iii
Acknowledgments iv
List of Tables viii
List of Figures ix
Table of Contents x
Literature Review 1
Specific Objectives 5
Materials and Methods 5
Results 7
Discussion 21
Conclusion 32
Chapter 2: The effect of human-specific promoter indels on gene expression 34
List of Figures 35
List of tables 36
Literature Review 37
Introduction 39
Objectives 40
Material and Methods 40
Results 58
Discussion 59
Conclusion 67
References 68
Appendix 1

Chapter 1: Comparative genomic of indels in primate lineages

Literature Review

Approximately six million years ago the human and chimpanzee lineages diverged from a common ancestor (Wetternbom, Sevov, Cavelier & Bergstrom, 2006). Subsequently, these two species have developed different morphological, behavioral, cognitive and other phenotypic traits that make them distinct. To gain insights about the genetic basis of the phenotypic differences that distinguish human from chimpanzees, a couple of comparative genomic studies have been made in the last few years (Polavarapu, Arora, Mittal & McDonald, 2011). The most surprising finding of these studies is the shortage of protein-nucleotide differences between these two species. As a consequence, these results support an early hypothesis that the foundation of phenotypic differences lies in the realm of gene regulation.

In fact, it was demonstrated that the significant differences in gene expression patterns that exist between human and chimpanzee, especially in organs (for example brain and testis) and functions such as cognitive ability and fertility, are directly related to some of the major phenotypic traits that are distinct in both species (Polavarapu et al., 2011). Previously, Volfovsky et al. (2009) reported that genome differences between humans and chimpanzee oscillates between 2.4% (excluding repeats and low complexity DNA, and including insertions and deletions (indels)). They included indels in their measurements because it is reasonable to think that indels may play an important role in primate evolution.

Indels are detected as fragments missing in sequence comparisons between closely related species. Although they are less abundant than simple nucleotide mutations, they account for most of the differences observed between species (Sjödin, Bataillon & Schierup, 2010). It has been suggested that the human genome might contain as many as 1.6-2.5 million indel polymorphisms (Volfovsky et al., 2009) and a small portion of them are present within regulatory elements of genes; promoters, splice and terminator sites, untranslated sequences as well as in coding regions. It is suggested that the paucity of indels within coding sequences is due to selective constraints acting on the functional elements of genes (Volfovsky et al., 2009). Therefore, more indels are expected in the intergenic regions. In some cases insertions and deletions could be located within genes, affecting gene expression or affecting the protein product. For example, insertions in coding sequences could result in translational frameshifts, triggering a premature stop codon and nonsense-mediated decay.

The origin of indels seems to be primarily related to their size and sequence context. Large scale indels are suggested to be caused by the proliferation and illegitimate recombination of transposable elements. For example, Alu, L1 and SVA retrotransposons are constantly generating insertions when they retrotranscribe their RNA sequences and insert them into new locations in the human genome. Indeed, Alu elements cause insertions of approximately 300 bp, while L1 and SVA elements cause insertions ranging from 10 bp to 3kb and 10 pb to 6 kb, respectively (Mullaney, Mills, Pittard & Devine, 2010). In contrast, it has been argued that short indels are frequently generated by DNA polymerase slippage and mispair, as in microsatellites (Sjödin et al., 2010).

Even though most of the indels have no adaptive value, some of them are known to change important human traits and many others are known to be involved in certain diseases. Indels located in important gene regions could affect gene function in different ways. For example, DNA insertions in the form of a simple repeat expansion within the 5' untranslated region of the *FMR1* gene is known to cause Fragile X Syndrome in humans (Mullaney et al., 2010). Similarly, an indel located in the promoter region of the *ACE* (Angiotensin-converting enzyme) gene has been shown to be a causative factor for coronary heart disease (Ruiz et al., 1994). Likewise, indels within transcription factor binding sites or enhancers could cause a decrease or abrogation of gene expression. In the same way, indels that are located within promoter regions could alter the phasing and spacing of the DNA sequences in that region, and as a consequence alter interactions in protein binding sites and affect gene expression (Mullaney, et al., 2010). An example of the latter was mentioned in a study performed by Sun et al. (2007). In this study, a deletion of six nucleotides (CTTACT) was found in the promoter region of the *CASP8* (*Caspase 8*) gene. Caspases are necessary for the immune surveillance of malignancies. This deletion was associated to a decrease in CASP8 transcription because it abrogates a stimulatory protein 1 binding site. In addition, they found in a case-control study in the Chinese population, that the indel was linked to a reduced susceptibility of lung, cervical, breast and gastric cancer, among others.

One last example of an indel that could disable a functional gene is the well documented deletion of 32 bp that is located in the open reading frame (ORF) of the *CCR5* locus. This deletion results in a premature termination of translation and the formation of a truncated protein (Al-Mahruqi et al., 2014). *CCR5* is a proinflamatory receptor that binds to chemokines and helps the entry of the R-5 strain of HIV-1.

Previously, King and Wilson (1975) suggested that the extensive differences observed in the phenotype of human and chimpanzees are not a consequence of changes in the amino acids of proteins, but instead, are the result of differential regulation of homologous genes between these two species. Consequently, that differential regulation of expression could account for the great phenotypic divergence that exists between humans and chimpanzee. Moreover, the fact that the chimpanzee and humans proteomes differ only by about 50,000 changes in the amino acids sequences further supports King and Wilson's proposal (Chimpanzee Sequencing and Analysis Consortium, 2005).

Taken together, these observations provide grounds to hypothesize that indels within promoter regions might explain many of the phenotype differences between humans and chimpanzees. Hence, our goal is to evaluate how indels in gene promoters regions can contribute to the regulatory differences between these two species.

Introduction

Approximately six million years ago the human and chimpanzee lineages diverged from a common ancestor resulting in two distinct species with major morphological, behavioral, and cognitive differences. However, their genomes are more than 98.5% identical at protein-coding loci, and differences between these two species hover around 2.4%, excluding repeats and low complexity DNA and including insertions and deletions (indels). It is thus believed that most of the genetic foundation for the profound differences among these two species lies at the level of gene regulation; specifically it may arise from differential regulation of orthologous genes.

As mentioned by Wray et al. (2003) "a gene embedded in random DNA is inert. Every gene with a phenotypic impact is flanked by regulatory sequences that, in conjunction with the expression and activity of proteins encoded elsewhere, regulate when the expression occurs, at what level, under what environmental conditions, and in which cells or tissues." It has been argued that the phenotypic impact of a gene is a function of two different components; 1) the biochemical activity of the protein it encodes and 2) the specific conditions under which that protein is expressed and is therefore capable of performing its activity (Wray et al., 2003). Variation in how transcription is regulated can result in significant phenotypic differences.

Eukaryotes have different methods to regulate gene expression; chromatin modification, DNA methylation, transcriptional initiation, alternative splicing of RNA, mRNA stability, translational controls, different forms of post-translational modifications, intracellular trafficking and protein degradation (Wray et

al., 2003). Nonetheless, the most frequent point of control are the rate of transcriptional initiation and mRNA stability.

The transcription of most genes in an eukaryotic genome is controlled independently. Most genes are differentially transcribed through the life cycle according to environmental conditions, in different cell types and regions and sexes. The majority of the genes have spatially and temporally heterogeneous expression profiles. For example, genes encoding regulatory proteins have some of the most complex expression profiles. In contrast, "housekeeping" genes have simpler transcription profiles, even though most of them are transcribed at different levels among cell types and are shut down in response to extreme environmental conditions (Wray et al., 2003).

Promoters or cis-regulatory sequences are responsible for the alteration of the rate at which transcription initiation is performed. This function is accomplished through transcription factors (proteins) that bind to the DNA in a specific manner; a step which is crucial to determine whether a gene is transcribed or not and how much mRNA is generated from it. The organization of promoters are by far very different to that of coding sequences; they do not present a consistent sequence motifs but they do have two functional features. Promoters have: (1) a basal promoter (core promoter) which is a region of ~ 100 bp that serves as a landing site for the transcription complex and whose function is to position the start of transcription relative to the coding sequences and (2) a group of different binding sites for discriminating transcription factors responsible for the specificity of transcription and triggering a scalar response (the frequency at which new transcripts are initiated) (Wray et al., 2003).

Basal promoter sequences differ among genes. Some promoters have a TATA box usually located 25-30 bp upstream from the transcription start site, but others lack the TATA box and instead have an initiator element. Also, there are promoters that do not have either the TATA box or the initiator element. Genes could also have more than one basal promoter, each of which could initiate transcription at different positions (Wray et al., 2003).

An important step in the transcriptional initiation is the binding of the TATA-binding protein (TBP) to the promoter region. Once the TBP binds, several TBP-associated factors (TAFs) guide the RNA Polymerase II holoenzyme complex onto the DNA. This latter step is one of the most prominent points of transcriptional regulation because it can be positively or negatively regulated by transcription factors located at other sides (Wray et al., 2003).

Transcription factor binding sites also have an important role in the production of appropriate levels of mRNA in cells. These regions are nucleotides sequences that can differ among genes and also dictate which transcription factors bind to the appropriate promoters. Moreover, promoter recognition by transcription factors can be influenced by the course of development of an organism as well as by environmental conditions at the tissue and cellular levels. Therefore, the change in the availability of the transcription factors provide a mechanism to control when, where, and at what level and under what circumstances a given gene is transcribed (Wray et al., 2003).

This work focuses on lineage-specific large (> 10 bp) indels in promoter regions that may affect gene expression and protein product levels. These were identified through the alignment of orthologous regions of human, chimpanzee, gorilla, orangutan and macaque. A total of 64 indels, located 5' and at distance no more than 2 kb away from the nearest transcription start site were detected. In order to differentiate real features from computational artifacts, PCR amplification and electrophoresis analysis were performed resulting in the validation of 47 indels.

Specific Objectives

- Identification of indels in promoter regions from pairwise genome comparisons between five primate species
- Validation of promoter indel and their variation in human and chimpanzee populations

Materials and Methods

Identification of indels in promoter regions from pairwise genome comparisons between five primate species

Preliminary analysis:

Orthologous regions of human, chimpanzee, gorilla, orangutan and macaque genomes were obtained and aligned using the MUMmer program (Delcher, Phillippy, Carlton & Salzberg, 2002) in order to identify small insertions and deletions with an approximate length of 10 bp to 10 kb. These sequences were surrounded by at least 10 bp of perfectly aligned sequences and with no more than 50% of undetermined bases. The selected fragments were evaluated with RepeatMasker and Tandem Repeat Finder to exclude those sequences that overlapped with known repetitive elements or short tandem repeats ($\geq$ 10 bp).

Insertion sequences were extracted from human and chimpanzee chromosomes, and BLAT searches were performed with all selected fragments against the available genome assemblies in order to identify genomic locations containing identical sequences (± flanking 18 bp). The local repeat structure of the indel

region was examined in the 5 Kbp of flanking genomic regions on both sides. With this information we focused on the indels that had no similarity with the flanking region (flanking region of 5 Kbp on both sides), and this type of indel was categorized as a unique indel. A total of 64 indels were selected by these criteria.

Since indels located in promoter regions were the intended target, the UCSC genomic database (https://genome.ucsc.edu/index.html) was used to annotate the human genes that were located at a distance of no more than 2 kb downstream the indel. The reference genome used for the annotation of genes was (Mar. 2006 (NCBI36/hg18)). Afterwards, the retrieved DNA sequence containing the indel was used to perform BLAT searches against the chimpanzee and gorilla genomes. The latter was performed to obtain the orthologous sequences for each indel in the species.

Indel Validation

In order to validate the existence of the indels by PCR and electrophoresis, orthologous sequences from humans (*Homo sapiens*), chimpanzee (*Pan troglodytes*) and gorilla (*Gorilla gorilla*) were aligned using MEGA. The conserved regions that flanked the selected indel were used to design primer pairs manually, which were analyzed for possible self-dimerization and dimerization between members of each pair using the OligoAnalyzer Tool 3.1 from IDTDNA (http://www.idtdna.com/calc/analyzer). Primers were also checked for their uniqueness in relation to the copies on the chromosomes and in the rest of the genome with BLAST from NCBI.

The designed primers were at least 20 base pair long and their annealing temperatures were optimized for the specific amplification of the intended amplicon. The primers pairs were tested on a set of 14 unrelated human DNA samples representing different ethnic groups, one sample from chimpanzee and one sample from gorilla. The primate samples were purchased from either the Integrated Primate Biomaterials and Information Resource or the Coriell Institute for Medical Research, in Camdem, NJ. In total, sixty-four indels were selected for experimental analysis and PCR and electrophoresis validation. Master mix composition and PCR parameters are described in Table 1.

Table 1 Composition and Conditions of the PCR Reactions

Components		
	25 ul Reaction	Final Concentration
5X One Taq Standard Reaction Buffer	5 µl	1x
10 mM dNTPs	0.5 µl	200 µM
10 uM Forward Primer	0.5 µl	0.2 µM
10 uM Reverse Primer	0.5 µl	0.2 µM
Template DNA	5 µl (15ng/µl)	3 ng/µl
One Taq DNA Polymerase **(5 U**/µl)	0.12 µl	
Nuclease-Free Water	13.38 µl	
Thermocycling Conditions		
	Temperature	Time
Initial Denaturation	94^0C	2:30 minutes
Denaturation	94^0C	1 minute
Annealing (35 cycles)	*Depends on the primer	40 seconds
Extension	68 ^{0}C[1]	1 minute
Final Extension	68 ^{0}C[1]	10 minutes
Hold	4^0 C	∞

[1] **As suggested from NEB, the extension temperature for the OneTaq® Polymerase is 68°C.**

PCR products were resolved by agarose (3%) gel electrophoresis. The Geneious (Biomatters Inc.) software was used to confirm the expected amplicon size in the three different species (human, chimpanzee and gorilla). The indel was considered "validated" only if the species expected to have an insertion or deletion relative to the other species showed a fragment of a different size as predicted. All indels found to be false showed no difference in fragment sizes between the three species.

Results

In order to differentiate real features from computational artifacts, 64 indels were selected for PCR and electrophoresis analysis (*in-vitro* analysis) (Table 2). Table 3 shows the genes that were located

downstream of the indel, the primers that were used for the PCR amplification, the expected amplicon size of the PCR products for humans, chimpanzee and gorilla (provided by the software *Geneious*), the outcome or status of the corresponding indel and the indel size. For some genes shown in table 3, the indel size does not match the expected amplicon size of the species that does not present the indel. This occurs in those cases where there are small gaps (1-3 bp) in the amplified fragment (e.g., *CCT4, DDX6, EBPL, FBXO33, ITGA11, MEPE, NSMCE1 and TMEM60*). The software takes this into consideration when predicting the size of the PCR products.

Table 2: Indel Coordinates (cont.)

Indel ID	***Chromosome Coordinates***	***Gene located downstream of the indel***
ind7681	chr1: 36461798	*THRAP3*
ind7403	chr1: 50662899	*DMRTA2*
ind7413	chr1: 154350990	*LMNA transcript variant1*
ind21624	chr10: 77210905	*C10ORF11*
ind9210	chr11: 4346785	*OR52B4*
ind9347	chr11: 71316665	*RNF121 transcript variant 2*
ind9420	chr11: 107969780	*EXPH5*
ind9426	chr11: 118168728	*DDX6 transcript varian 1*
ind14024	chr12: 7489209	*CD163L1*
ind14307	chr12: 115661958	*C12ORF49*
ind20282	chr13: 22906685	*SACS*
ind20660	chr13: 49163670	*EBPL*
ind4933	chr14: 38973065	*FBXO33*
ind5019	chr14: 94678955	*DICER1 transcript variant 2*
ind15212	chr15: 38613924	*MRPL42P5*
ind15360	chr15: 66513445	*ITGA11*
ind14966	chr15: 90736840	*ST8SIA2*
ind12328	chr16:27187711	*NSMCE1*

ind12216	chr16: 29783870	*CDIPT*
ind11845	chr16: 86357187	*KLHDC4 transcript variant 1*
ind12525	chr18: 11740248	*GNAL transcript variant 3*
ind12412	chr18: 53861320	*NEDD4L transcript variant J*
ind6984	chr19: 7890437	*SNAPC2 transcript variant 1*
ind6934	chr19: 12374584	*ZNF799*
ind7052	chr19: 51910877	*PRKD2 transcript variant 4*
ind17978	chr2: 61969824	*CCT4 transcript variant 1*
ind18298	chr2: 198359626	*BOLL transcript variant 1*
ind18403	chr2: 224975001	*FAM124B transcript variant 2*
ind18086	chr2: 232028945	*SNORA75*
ind17950	chr2: 232038389	*NCL*
ind4408	chr20: 18215515	*ZNF133 transcript varinat 2*
ind4456	chr20: 33336176	*EIF6 transcript variant 1*
ind4193	chr20: 35158292	*RBL1 transcript variant 1*
ind11499	chr21: 33024012	*SYNJ1 transcript variant 3*
ind11622	chr21: 39608783	*BRWD1 transcript variant 2*
ind4657	chr22: 23220340	*UPB1*
ind4777	chr22 :28565517	*ASCC2 transcript variant 2*
ind4660	chr22: 34351282	*MB transcript variant 2*
ind4692	chr22: 39581372	*XPNPEP3 transcript variant 2*
ind13539	chr3: 121797709	*NDUFB4 transcript variant 1*
ind12845	chr3: 122108488	*STXBP5L*
ind15972	chr4: 88972216	*MEPE transcript variant 3*
ind16187	chr4: 119494521	*PRSS12*
ind9998	chr6: 39123571	*GLP1R*
ind10256	chr6: 106651651	*PRDM1 transcript variant 2*
ind2852	chr7: 38694633	*FAM183B*
ind2026	chr7: 77266351	*TMEM60*

ind782	chr7: 99219862	*CYP3A4 transcript variant 2*
ind692	chr7: 116098619	*MET transcript variant 2*
ind1133	chr7: 127667278	*LEP*
ind2231	chr7: 128367216	*IRF5 transcript variant 3*
ind2530	chr7: 149102659	*ZNF467*
ind2173	chr7: 150768077	*CRYGN*
ind2767	chr7: 151206474	*PRKAG2 transcript variant A*
ind6323	chr8: 37737977	*PROSC*
ind6201	chr8: 42246807	*IKBKB transcript variant 7*
ind5684	chr8: 59628132	*SDCBP transcript variant 4*
ind6536	chr8: 144787713	*ZNF623 transcript variant 3*
ind19614	chr9: 33438118	*AQP3*
ind20109	chr9: 89577558	*CTSL3P*
ind19494	chr9: 139268407	*NELFB*
ind17801	chrX: 130019056	*ARHGAP36*
ind5578	chrY: 7201869	*PRKY*
ind9795	chr6: 36271910	*BRPF3*

Table 3 Results of indel validation (cont.)

Gene	**Primer Sequence**	**Amplicon Size of the PCR Product**			**Status[1]**	**Indel Size (bp)**
		Human	Chimp	Gorilla		
AQP3	Forward CACTGTCTCTTCTGTCAGGACAGATAAGG Reverse CGACGTGCTCATAGCACAGGGAGAAG	157	139	157	+	18
ARHGAP36	Forward AATCCAACACAGTACCACCCTCC	134	117	134	+	17

	Reverse CAGTTGCTGCCGACTTACAGATTCC					
ASCC2 Transcript variant 2					*	12
BOLL transcript variant 1					/	17
BRPF3	Forward TGTCCTCAGGTCCTGGGCAT Reverse ACGGTTTGGAGCAGCAAGCC	168	157	168	+	11
BRWD1 transcript variant 2	Forward GTAAACCGTAGGTAATTCCTC Reverse GCAAGACCCTGTCTCAAAAT	171	149	195	+	22
C10orf11	Forward GGTAGTCATACCTTTCCTCATTG Reverse CTGGGACAAAGAGGTGAATGAA	144	132	144	+	12
C12orf49	Forward AAATGCAGTTGGAAACAGATGATGC Reverse AGTCCTGATTGGGGAAGCCAGAA	136	118	136	+	18
CCT4[2] transcript variant 1	Forward AGTGGTGAGGACATCCGCATTTCC REVERSE CACCTGTCGTCCTGGCTAGTTGG	514	497	___	+	18
CD163L1	Forward TGCTAGATATGAGCGAGATGTGCC	204	193	204	+	11

	Reverse CCATCATGTGACTTATCCTAGTTAGTG					
CDIPT	Forward CCTCAGGACCACCGTGTCTAGAGAACC Reverse AGAAGAGATGCGGCCAGGGCAGA	228	214	228	+	14
CRYGN	Forward TTTCCTTGGGGTTGAGGGACGCACTCAC CT Reverse AGCTAAACGCGGAGAGCACAGGGAGA CC	277	258	277	0	19
CTSL3P	Forward TGTTTCTTGACTTGGAGAACATCTCCCA Reverse CACTGCTCTCCTCCATCCTTCTTC	175	164	175	-	11
CYP3A4 transcript variant 2					/	15
DDX6 transcript variant 1	Forward GGGAAGCTAGTGAAGCGTCAGT Reverse AAGAAATAGGGAAGCTGCCGGTTAAAG AT	380	359	382	+	22
DICER1 transcript variant 2	Forward CCTGCAAAATCCTGTTTCATAGGCCCC Reverse CACAGCAGAAAGTTCTAGGTGCCTTTA GGC	212	184	212	+	28
DMRTA2	Forward	176	164	164	+	12

	CTTTCGCCTTATTCTCGTCTT Reverse ATCAGAGCTACAGAAACCGAGG					
EBPL	Forward GTCTAGGCGATGGCACTTACC Reverse GTTGGTGAGCACGTTCCTTCTAC	471	447	472	0	24
EIF6 transcript variant 1	Forward CGTTCATTCATTCATTATTCATTCAG Reverse ATCAATACCTCCATCATCAACC	139	128	139	+	11
EXPH5	Forward TCCAACCGAGATGCAAAGTGAACG Reverse CACTTAAGAGGCAGAATTATGTAGGCT TGT	206	189	206	+	17
FAM124B transcript variant 2	Forward GGTCACGTCATCCAGCTTGTGC Reverse CTAACAGAGCCAACTCATGTTCCTT	181	169	181	+	12
FAM183B	Forward CTCTTGCTCCTTATGAGTTATC Reverse CTTTATTAGCGGCATGAGAACA	193	179	193	+	14
FBXO33	Forward ATGTATGAGGGTTTCCAGTTCT Reverse CATTTTACTCAGGAACCATTCC	169	156	166	+	11
GLP1R	Forward CTCTGCCACAACTCATCTTTCC	154	138	154	+	16

	Reverse AGGTTCCCTTCTTACACCCAACAAG					
GNAL transcript variant 3	Forward ATGGAATCTGGCATAACTCCCACC Reverse TTGATCCCTCCTAGAGACTGCATTGG	396	382	396	+	14
IKBKB transcript variant 7	Forward CTTAACATGCTACTTTTAGCCACGG Reverse AAGTGTTGAGATTACAGGCTGTGAGCC	151	133	151	+	18
IRF5 transcript variant 3	Forward ACCGAACTTCCAAAGTCATGG Reverse TCTAACCCGAACAGCATCCATCCT	107	96	107	+	11
ITGA11	Forward GAATGCAAATTCCCCAAGGATCAGG Reverse AGGACGGTGAAGGTGAATGAATG	192	164	192	+	26
KLHDC4 transcript variant 1	Forward CTTAGACACCTTCTTCTCCATCTTGG Reverse AACTCCTGGACTCAAGCGATCC	411	390	411	0	21
LEP	Forward TCACTCTTGTTGCCCAGGCTGTAGTG Reverse AGGTGTGGTGGTGAGTGTCTGTAATCC	133	119	133	-	14
LMNA transcript variant1	Forward TGAGCAGGCAGGAGCCAAGAGA Reverse GCTCTAATAGGTCCTCCTCTGAAGGG	199	184	199	+	15

MB transcript variant 2	Forward GAGCATTGAGAGGTGGTAGGAGG Reverse AGTCCAGGAGTCTCATTCCAAAGC	123	123	134	+	11
MEPE transcript variant 3	Forward TCTGAGTCTGTTCATGCTGCTACTACAG Reverse CCTCTGGAGATGCAGTAGCAAGGT	186	167	185	+	19
MET transcript variant 2	Forward TATGCCATGCCGTATCAGGA Reverse GGTAATAGGATCTAAGGAACGGGCATT GCC	545	568	568	+	23
MRPL42P 5	Forward CTCACACTGCTATAAGAACATACTGGAG AC Reverse TCTCATTCTCCTTCCTGCTGCC	167	156	167	-	11
NCL	Forward CAGTGACTTCCACGGTTAGCTT Reverse GGAACAATGAGGTATGGATGGATC	145	132	145	+	13
NDUFB4 transcript variant 1	Forward TGGTTATGTCAGTAGAGACAAGGCT Reverse CTGACCTTCAGCTTCAGGAAGTG	233	206	233	+	27
NEDD4L transcript variant j	Forward AGGCAAAGGAGAGCTTAAGGTAGCATG Reverse TCATATTCCTCGTTGCCAGCTACATA	103	85	103	+	18

NELFB	Forward CCACCTTGCAGGAGCCTTACCT Reverse CCTTTTTGAAGGGGTTGAGCTCTGGGCCT	105	92	105	+	13
NSMCE1	Forward GTCCGTTGAGTGACGCACTTCCGGTTCTCC Reverse TAAGGATCGAGAGCGGGCGTAATTTGG	100	88	99	+	12
OR52B4	Forward TTCGATCACAACTTCACTCTTAGG Reverse GTCTACTCCCTGCATTAAGAACGTAC	256	252	263	-	11
PRDM1 transcript variant 2	Forward TTCCCAAGCAAAAGAGGGTAGT Reverse GGTTCCACCATTGTAAGTCAGGTGAT	116	100	116	+	16
PRKAG2 transcript variant a	Forward TCATCTCTTAGGAAGCAGGCGTG Reverse ATGAGCCAGGGAACCACCCAAG	137	126	137	+	11
PRKD2 transcript variant 4	Forward TTCTCCTGTTTCCACCAAATGG Reverse ATTTCATGTCCCTGCATGTCGTTG	154	144	154	-	10
PRKY					*	15
PROSC	Forward GCATCCCACTTCATTAAGTTGA Reverse	100	87	100	+	13

	GCAAGGAGAAGGAGATCAATT					
PRSS12	Forward GCAGGCTCCAAAAAACTGTTGGCTTAG Reverse GTTGCCTTCTTTACGGGCTTCCTT	142	130	142	+	12
RBL1 transcript variant 1	Forward TTGAGTTAGGAGATGGAGGCTGG Reverse TCCTTCCCCTCTGTCATTTAGCGAACC	123	101	123	+	22
RNF121 transcript variant 2	Forward CATCCTAGTTCCATTCACCAGACAG Reverse ATACACCATTCACTGTCCCACC	147	120	147	+	27
SACS transcript variant 1					/	11
SDCBP transcript variant 4	Forward TGAATTGGAGGCGACGAGAACCAAGC Reverse AAAGTGGGGCGGTTCATGCC	167	153	167	+	14
SNAPC2 transcript variant 1					/	11
SNORA75	Forward ACTGCCACTGATAGACAGAAAGGAGC Reverse GGAAACTGGGTCCTCCAAAGGGTAAG	133	119	133	+	14
ST8SIA2	Forward AAGGGCTCTGGGCACAAGCA	111	100	111	+	11

	Reverse AGAGGAGTTTGGCGAGGGGCT					
STXBP5L	Forward CCAAAAGGAGTGAAATAGTCATTCAGG Reverse GGATGATGACTTGTCCAAGGACATAGC	122	111	122	+	11
SYNJ1 transcript variant 3	Forward TTGGACTGAGGAGATTTAGGTC Reverse CTAAAATGAGATGTGGAAAAGCTGTGG	139	123	139	+	16
THRAP3	Forward GTGGGAACACAGGCATAGGGAGGAG Reverse CTCTCGGGTGATTCTTACCCACCTACA	178	154	178	+	24
TMEM60	Forward TGTGGAGGTCGTCGAGTTCTGACAG Reverse AAGGAAGCAAAGGACAGGGGCCTGGAA	313	289	313	-	21
UPB1	Forward CGCTTCTTTGCTCCTGAAGGGATGG Reverse AAGGTCAGTTGCTCGCCTGC	256	244	256	+	12
XPNPEP3 transcript variant 2	Forward TGCAGTGATCCCGCCACTGTATGC Reverse GCAACTATGCAGACCATTGCGACTCCT	129	115	129	+	14
ZNF133 Transcript variant 2	Forward TTGAAGGACCAGATGGATTCAGAGG Reverse	101	90	101	+	11

	TTGCTGCCATGTCCTCTTTTTCTCCT					
ZNF467	Forward TCTGCCTCCAGGATGAAGGGAGC Reverse CACACATGCACACTCTCACTGTAAGTCT C	115	101	115	-	14
ZNF623 transcript variant 3	Forward ACAGGGAGGCTGAAAGGTCTGTAAGC Reverse GGCTGCTGCAACCCTATTGCCAG	106	83	106	+	23
ZNF799	Forward GCATATATTGAACCATCCTTGC Reverse ACAGAGGAGAAGGCAATACTTCC	237	225	237	-	12

[1]Validated indels are represented by a (+) sign, indels that demonstrated not to exist are represented by (-) sign, indels that were not observed during the *in-silico* analysis are presented by (/), indels whose PCR amplification was not successful are represented by (0) and indels whose primers were not designed are presented by (*) [2]The predicted amplicon size for the gorilla species, corresponding to the *CCT4* gene, could not be determined because the DNA sequence has undetermined nucleotides.

Sixty four indels were selected for PCR and electrophoresis analysis, however four indels were not observed when performing the DNA alignment (*in-silico* analysis) of the three different species (*SNAPC2*, *SACS*, *CYP3A4* & *BOLL*). Likewise, a total of five indels were not analyzed because in three of them (*CRYGN*, *EBPL* & KLHDC4) the PCR amplification (*in- vitro* analysis) was not successful and primers were not designed for the remaining two (*ASCC2* & PRKY).

As shown in table 3, only 47 indels were validated. We found that 44 indels presented a deletion in the chimpanzee lineage relative to the humans and gorilla. We also found two different genes whose indels were unique for the human lineage; *DMRTA2* and *MET*. *DMRTA2* presented an insertion in the human lineage relative to the chimpanzee and gorilla whereas *MET* presented a deletion in humans relative to the chimpanzee and gorilla. In addition, we observed that the *MB* gene presented a deletion in humans and chimpanzee lineages relative to the gorilla.

Therefore, indels were classified into three different categories (see table 4): (1) null/not validated- when the indel was not observed when performing reference genome alignment (*in-silico* analysis), (2) indel/not validated- when the *in-vitro* and *in-silico* evidence did not match, and (3) indel/validated- when the presence of the indel *in-silico* was validated by *in-vitro* analysis. In order to see all the agarose-gel pictures for every gene, refer to the appendix.

Table 4 Results of indel validation categorized.

Status	Number of genes	Genes
Indel/validated	47	*AQP3, ARHGAP36, BRPF3, BRWD1, C10orf11, C12orf49, CCT4, CD163L1, CDIPT, DDX6, DICER1, DMRTA2, EIF6, EXPH5, FAM124B, FAM183B, FBXO33, GLP1R, GNAL, IKBKB, IRF5, ITGA11, LMNA, MB, MEPE, MET, NCL, NDUFB4, NEDD4L, NELFB, NSMCE1, PRDM1, PRKAG2, PROSC, PRSS12, RBL1, RNF121, SDCBP, SNORA75, ST8SIA2, STXBP5L, SYNJ1, THRAP3, UPB1, XPNPEP3, ZNF133, ZNF623*
Null/not validated	4	*BOLL, CYP3A4, SACS, SNAPC2*
Indel/not validated	8	*CTSL3P, ZNF799, ZNF467, TMEM60 PRKD2, OR52B4, MRPL42P5, LEP*

Discussion

As previously shown in table 3 and 4, we validated 47 indels. From this total, 44 genes presented a deletion in the promoter region of the chimpanzee lineage relative to the human and gorilla. In addition, we found two different genes, *DMRTA2* and *MET* whose indels were unique for the human lineage relative to the other two primates analyzed and one gene (*MB*) whose indel was unique for the gorilla (See Appendix). Furthermore, we also observed one gene (*BRWD1*) whose indel was different in every species analyzed.

The observation that all indels in the chimpanzee lineages were deletions is intriguing. It could be hypothesized that a genome-wide selective pressure to shorten the genome is operating in the chimpanzee. Alternatively, some chimpanzee-specific factor, common to certain kinds of promoters could be driving their deletions.

Regarding *BRWD1 (bromodomain and WD repeat domain containing 1)*, this gene had unique indels for each lineage. The indel of *BRDW1* was located 1,201 bp (Mar.2006 (NCB1 36/hg 18 assembly)) from the nearest transcription start site and we confirmed the expected deletion of 22 bp in the chimpanzee lineage (relative to humans).In addition, we also observed an insertion of 24 bp (relative to humans) in the gorilla lineage, in the *in-silico* analysis, that was confirmed by the electrophoresis analysis (Figure 1). It is important to note that the insertion in the gorilla lineage, relative to the human lineage, is a different sequence from that observed in the human lineage. The sequence of 24 bp in the gorilla lineage has three 12 bp sequence imperfectly repeated. The 12 bp unit is composed of 3 copies of a 4 bp element. This structure suggests an expansion of the repeat, in the gorilla lineage, by a DNA polymerase slippage and mispair mechanism.

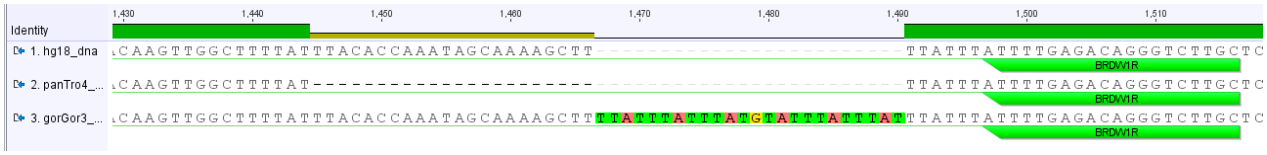


Figure 1 *In -silico* analysis of orthologous *BRWD1*-DNA promoter sequences. DNA alignment of orthologous sequences from human (hg 18_dna), chimpanzee (panTro4) and gorilla (GorGor3) were performed by the *Geneious* software. A deletion of 22 bp is observed in the chimpanzee lineage relative to the human lineage. In addition, an insertion of 24 bp is observed in the gorilla lineage relative to the human lineages.

BRWD1 is located at chromosome 21, specifically, within the Down syndrome region-2. The protein encoded by *BRWD1* has seven WD40 repeat domains and dual bromodomains. WD40 are repeats observed in different eukaryotic proteins such as G proteins, phosphatases and transcription regulators. These ~40 amino acids repeats form a structure that enables interactions with other proteins or ligands. Bromodomain-

containing proteins typically bind acetylated histones and are related to relaxed chromatin. They are involved in different cellular functions. For example; transcriptional activation, transcriptional silencing, chromatin remodeling, mRNA splicing and DNA replication (Philipps et al., 2008). Interestingly, it has been shown that *BRWD1* has a key role in the epigenetic control of chromosome structure during female meiosis. In addition, it was found that in males, this gene is involved in the control of haploid gene transcription during postmeiotic differentiation events of spermiogenesis (Pattabiraman et al., 2015).

Furthermore, from all the analyzed genes, only one indel was unique for gorillas. The indel for the *MB* gene was located 1,935 bp (Mar.2006 (NCB1 36/hg 18 assembly)) from its nearest transcription start site and we confirmed an insertion of 11 bp in the gorilla lineage (relative to humans and chimpanzees). This gene, is located on chromosome 22, it has 6 exons, encodes a member of the globin superfamily and is expressed in the skeletal and cardiac muscles. In addition, according to NCBI (2016), this protein (haemoprotein) contributes to the intracellular storage of oxygen and the transcellular facilitated diffusion of oxygen.

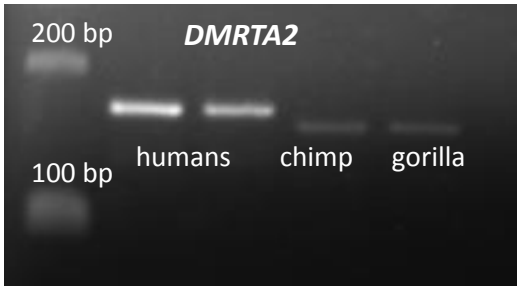


Figure 2 Electrophoresis Validation of *DMRTA2*-indel in promoter region.

On the other hand, we also found two different promoters, in which their indels were unique to the human lineage. These indels were located upstream *DMRTA2* and *MET* genes. The presence of the12 bp indel, 1,192 bp (Mar.2006 (NCB1 36/hg 18 assembly)) from the nearest transcription start site of *DMRTA2* was confirmed *in-silico* for the human lineage (see figure 2, for more details). *DMRTA2*, also called *DMRT5*, is located on chromosome 1, has 3 exons and it encodes for a transcription factor that required during sexual development. More specifically, *DMRTA2* is necessary for the proper differentiation of oogonia during the female embryonic germ cell development (Poulain et al., 2014).

In contrast, *MET* (*proto-oncogene, receptor tyrosine kinase*) which is located on chromosome 7, had a deletion of 23 bp unique to the human lineage as predicted by the *in-silico* analysis. The *in-silico* analysis predicted an indel located 1,076 bp (Mar.2006 (NCB1 36/hg 18 assembly)) from the nearest transcription start site of *MET*. Therefore, a MetaPhor-agarose gel at 3 % was prepared, as shown in figure 3, to validate our predictions. *MET* which is also known as *AUTS9* is a proto-oncogene and encodes the hepatocyte growth factor receptor. This gene has 24 exons and two different isoforms (NCBI, 2016). In addition, it has been well documented that variants within the promoter region of this gene are involved with the severity of

autistic symptomatology (Rudie et al., 2012). Moreover, it is known that this gene has a key role in the early brain development and different studies have reported that reduced expression of MET is associated to neurodevelopmental disorders (Hedrick et al., 2012).

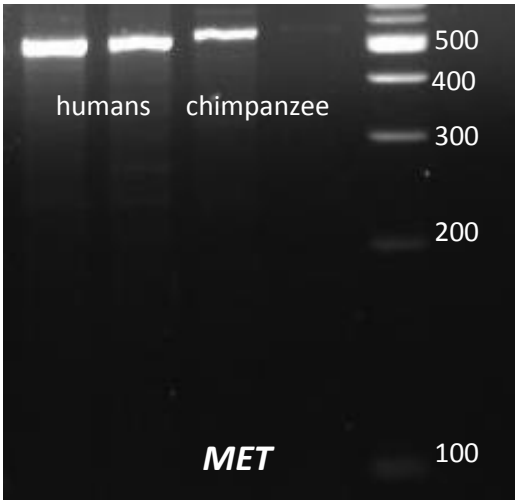


Figure 3 Electrophoresis Validation of *MET*-indel in promoter region.

In order to explore the biological implications of our results, we used the PANTHER biological process categories (Mi, Muruganujan, Casagrande, & Thomas, 2013) to gain insights about the biological processes that could be related to the genes that presented a deletion in the chimpanzee lineage relative to the human lineage. To accomplish this, we excluded from the analysis *DMRTA2*, *MB* and *MET*. We also had to exclude *SNORA75* and *C10orf11* because they were not recognized by PANTHER Classification system. Perhaps *SNORA75* was excluded from the analysis because this gene had a provisional RefSeq status in *Homo sapiens*, as reported by the Gene database from NCBI (NCBI, 2016). *C10orf11* was probably excluded because its gene symbol in *Pan troglodytes* is *C10H10orf11*, therefore it was not recognized by the software. However, when the gene symbol, *C10H10orf11* was used instead, this gene appears to be uncharacterized. For that reason we analyzed a total of 42 genes using as a reference the *Pan troglodytes* genome.

When performing the analysis we found that 54.8% of the total genes were classified, by PANTHER Classification system (http://pantherdb.org/), as being involved in metabolic processes (represented by red color in the pie chart, figure 4). According to the software's literature, this category is defined as "the chemical reactions and pathways, including anabolism and catabolism, by which living organisms transform chemical substances. Metabolic processes typically transform small molecules, but also include macromolecular processes such as DNA repair and replication, and protein synthesis and degradation". The genes with validated indels involved in metabolic processes were *DDX6, CD163L1, NDUFB4, PRSS12, IKBKB, SYNJ1, RBL1, BRPF3, GNAL, PROSC, DICER1, RNF121, XPNPEP3, BRWD1, EIF6, UPB1, CCT4, ZNF623, NEDD4L, PRKAG2, IRF5, ZNF133* and *CDIPT*.

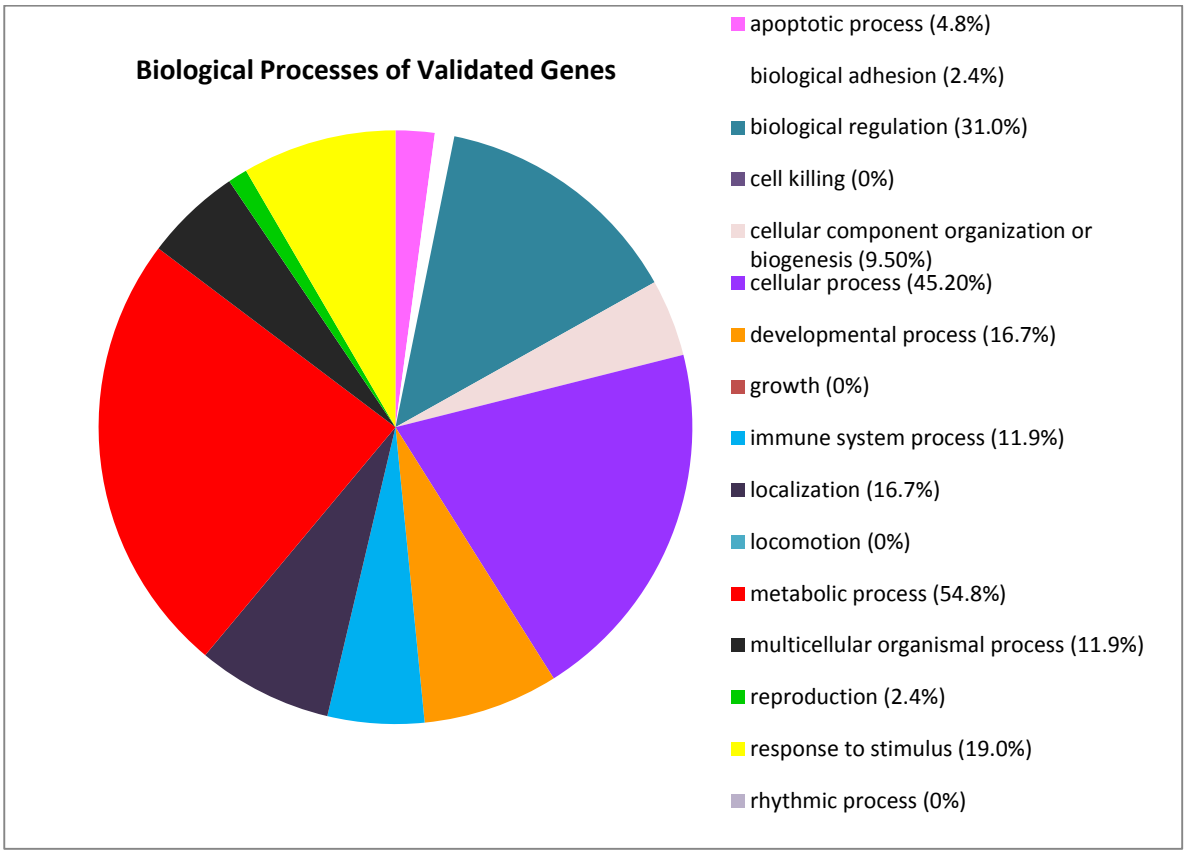


Figure 4 Pie Chart of the Biological processes attributed to promoters that presented a deletion in the chimpanzee lineage. A total of 42 genes were evaluated and categorized as the biological function that they perform in *Pan Troglodytes*.

When analyzing the pie chart provided by the PANTHER Classification System, we found, in many cases, that one gene could be classified by different biological functions. For example, besides being involved in apoptotic processes (representing a 4.80%), *PRSS12* was also linked to cellular adhesion (the only gene in this function, representing a 2.4%). *PRSS12* is located on chromosome 4 and previous studies in murine models have suggested that the encoded enzyme could be associated with structural reorganizations related with learning and memory. In addition, defects in this gene are linked to mental retardation (NCBI, 2015).

The second largest category was cellular processes. Genes grouping in this functional category were: *CD163L1, PRSS12, IKBKB, GLP1R, SYNJ1, RBL1, LMNA, GNAL, PROSC, DICER1, STXBP5L, ARHGAP36, AQP3, RNF121, BRWD1, ZNF623, NEDD4L, PRKAG2* and *CDIPT*, representing a 45.2%. In this classification we found activities related to cell communication, cell cycle, cell growth, cell proliferation,

cell recognition, cellular component movement, chromosome segregation, cytokinesis and viral processes. The third largest category was biological regulation with a 31.0 % of the total promoters (represented by dark green in the pie chart), regulating the following genes; *DDX6, PRSS12, RBL1, GNAL, DICER1, AQP3, RNF121, BRWD1, EIF6, ZNF623, PRKAG2, IRF5 and ZNF133.*

On the other side, 9.5% of the total genes belonged to the category of cellular component organization or biogenesis we found *LMNA, DICER1, BRWD1* and *CTT4*. This category was defined by the software as "a process that results in the biosynthesis of constituent macromolecules, assembly, arrangement of constituent parts, or disassembly of a cellular component."

PRSS12, GLP1R, LMNA, DICER1, STXBP5L, NCL and BRWD1 (16%) were associated with developmental processes. Developmental processes were referred as "any biological process whose specific outcome is the progression of an integrated living unit: an anatomical structure". According to PANTHER within this category we could find activities such as cell differentiation, anatomical morphogenesis, death, ectoderm development, embryo development, endoderm development, mesoderm development, patter specification process, sex determination and system development.

Five genes, *PRSS12, IKBKB, GLP1R, PRKAG2* and *IRF 5* (11.9%) were found to be associated with antigen processing and presentation, immune response and macrophage activation functions while *SDCBP, CD163L1, PRSS12, GLP1R, SYNJ1, STXBP5L,* and *AQP3* (16.7%) were found to be related to RNA localization, protein localization and protein transport processes. In the category of multicellular organismal processes, we found *GLP1R, SYNJ1, GNAL, STXBP5L* and *NCL*, for 11.9%. This area is defined by PANTHER as "Any biological process, occurring at the level of a multicellular organism, pertinent to its function". In the area of reproduction we found only *GLP1R*, representing 2.4%.

Finally, the last category is response to stimulus with a 19.0% of the total genes. This category spans "any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a stimulus. The process begins with detection of the stimulus and ends with a change in state or activity or the cell or organism", as defined by PANTHER. This category comprises responses to abiotic and biotic stimulus, endogenous and external stimulus, responses to stress, pheromones and toxic substances. It also includes cellular defense response, defense response to bacterium and immune response. Genes associated to these responses were *IRF5*,

PRKAG2, RNF121, AQP3, DICER1, GNAL, GLP1R and *IKBKB.* There were also four more categories; cell killing, growth, locomotion and rhythmic processes, however none of them showed any particular gene.

We hypothesize that a general selective pressure to shorten the genome in the chimpanzee would have an equivalent effect in all gene promoters, independently of their function. By contrast, if deletions found were the consequence of some chimpanzee-specific factor for certain kinds of promoters, we hypothesize that the frequency of such categories would differ from its frequency for the whole genome.

In order to see if our sample of 42 genes that were analyzed with PANTHER were a representative sample, we used PANTHER's whole genome function tool to visualize how the human (Figure 5) and chimpanzee genomes (Figure 6) were arranged according to their biological function. Additionally the frequency of each function was compared between them in order to detect significant differences.

Part I

Question: Test if the frequencies of the chimpanzee sample (42 genes) differ significantly or not from the whole human genome frequencies (Figure 5).

We found that 9.5% of the 42 genes were classified as belonging to cellular component organization or biogenesis, 45.2% to cellular process, 16.7% to localization, 4.8% to apoptotic process, 2.4% to reproduction, 31% to biological regulation, 19.0% to response to stimulus, 16.7% to developmental process, 11.9% to multicellular organismal process, 2.4% to biological adhesion, 54.8% to metabolic process, 11.9% to immune system process, 0% to cell killing, 0% to growth, 0% to locomotion and 0% to rhythmic process.

In order to answer our question, we did a Chi-Square Goodness of Fit Test for each biological function separately. We did not apply the Chi-Square Goodness of Fit Test for all the functions together because this approach can only be done with unique categories. Since in our analysis, some genes can be found in more than one category, we did the test for each category separately.

Stated hypotheses:

- **Null hypothesis:** The frequencies of our sample do not differ from the frequencies of the whole human genome. According to the whole human genome frequencies, 6.9% of the 19,184 genes were classified as belonging to cellular component organization or biogenesis, 35% to cellular process, 13.6% to localization, 2.9% to apoptotic process, 2.1% to reproduction, 20.4% to biological regulation, 11.3% to response to stimulus, 12.8% to developmental process, 8.5% to multicellular

organismal process, 3.2% to biological adhesion, 43.0% to metabolic process and 7.3% to immune system process, 0% cell killing, 0% to growth, 0.3% to locomotion and 0% to rhythmic process.

- **Alternative hypothesis:** At least one of the proportions in the null hypothesis is false.

Formulated analysis plan: For this analysis, we used separate Chi-Square Goodness of Fit Tests and the significance level was 0.05.

Analyzed sample data: Using the Chi-Square Goodness of Fit Test, we calculated the degrees of freedom, the expected frequency counts and the chi-square test statistic.

Table 5 Chi-Square Goodness of Fit test for the human genome

Chimpanzee Sample against Human Genome	
Biological Processes	**X^2**
Apoptotic process	0.536402938
Biological adhesion	0.665585428
Biological regulation	2.279863633
Cell killing	0.049515107
Cellular component organization or biogenesis	1.115847983
Cellular process	1.339168897
Developmental process	0.489904409
Growth	0.109026197
Immune system process	1.300612069
Localization	0.293144034
Locomotion	1.16338959
Metabolic process	1.405864218
Multicellular organismal process	0.553580989
Reproduction	0.295094322
Response to stimulus	2.222154705
Rhythmic process	0.008757298

Interpretation of results: For each biological function the x^2 was less than the critical value 3. 84 (α= 0.05), therefore we do not reject the null hypothesis. We can conclude that for each category (biological function) the frequencies of our sample fit the frequencies of the whole human genome.

Part II

Question: Test if the frequencies of the chimpanzee sample (42 genes) differ significantly or not from the chimpanzee genome frequencies.

Stated hypotheses:

- **Null hypothesis:** The frequencies of our sample do not differ from the frequencies of the chimpanzee genome. According to the whole chimpanzee genome distribution, 7.1% of the 17387 genes were classified as belonging to cellular component organization or biogenesis, 35.1% to cellular process, 13.9% to localization, 2.9% to apoptotic process, 2.1% to reproduction, 20.4% to biological regulation, 11.3% to response to stimulus, 13.1% to developmental process, 8.4% to multicellular organismal process, 3.2% to biological adhesion, 43.6% to metabolic process, 7.5% to immune system process, 0% cell killing, 0% to growth, 0.4% to locomotion and 0% to rhythmic process.
- **Alternative hypothesis:** At least one of the proportions in the null hypothesis is false.

Formulated analysis plan: For this analysis, we used the Chi-Square Goodness of Fit Test and the significance level was 0.05.

Analyzed sample data: Using the Chi-Square Goodness of Fit Test, we calculated the degrees of freedom, the expected frequency counts and the chi-square test statistic.

Table 6 Chi-Square Goodness of Fit test for the chimpanzee genome

Chimpanzee Sample against Chimpanzee Genome	
Biological Processes	**X^2**
Apoptotic process	0.501429127
Biological adhesion	0.088304285
Biological regulation	2.872421778
Cell killing	0.019333679
Cellular component organization or biogenesis	0.381276984
Cellular process	1.88547395
Developmental process	0.480728023
Growth	0.009664615
Immune system process	1.198246569
Localization	0.264759804
Locomotion	0.147870253
Metabolic process	2.125084258
Multicellular organismal process	0.679713994
Reproduction	0.017661021
Response to stimulus	2.518813371
Rhythmic process	0.501429127

Interpretation of results: For each biological function the x^2 was less than the critical value 3. 84 (α= 0.05), therefore we do not reject the null hypothesis. We can conclude that for each category (biological function) the frequencies of our sample fit the frequencies of the whole chimpanzee genome.

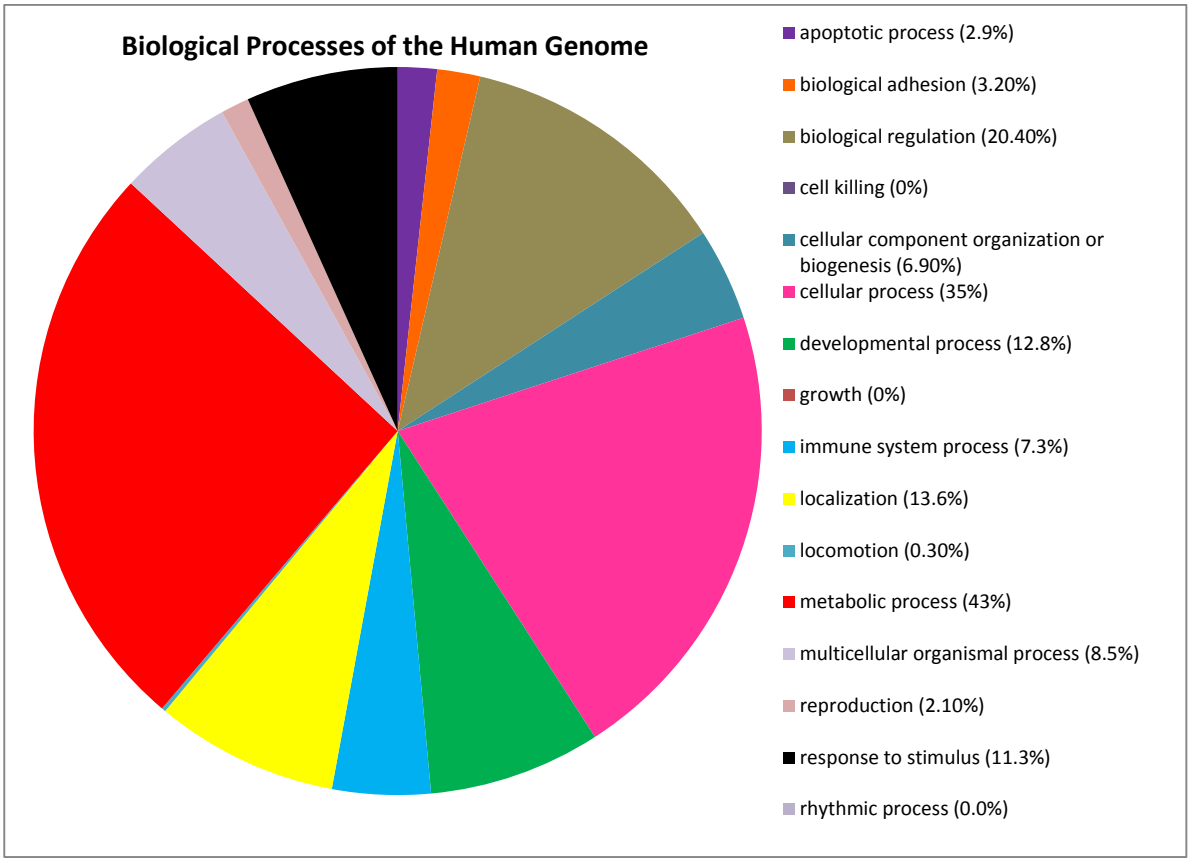


Figure 5 Pie Chart of the biological processes of the human genome. A total of 19184 human genes were analyzed with the Panther Classification System.

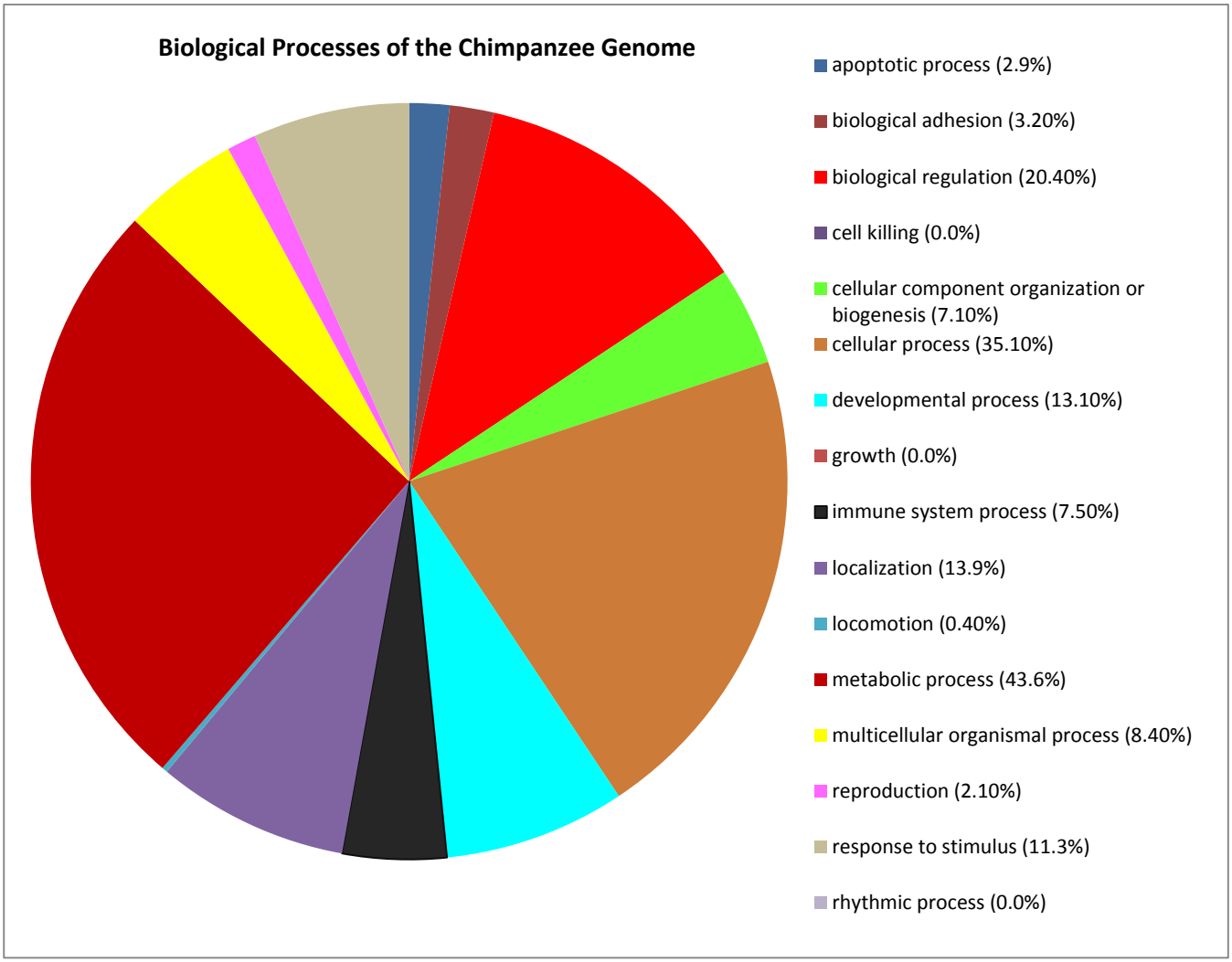


Figure 6 Pie Chart of the biological processes of the chimpanzee genome. A total of 29381 chimpanzee genes were analyzed with the Panther Classification System.

Conclusion

Previously, Polavarapu et al. (2011) suggested that most of the differences that exist between humans and chimpanzees are in areas associated with the cognitive ability and fertility. Consistent with this notion, two distinct indels unique to the human lineage, in promoters of genes associated with neurodevelopment and the female-sexual development; *MET* and *DMRTA2*, respectively were identified. Therefore, our work could indicate that the derived allele observed in the human lineage might be important for processes that make the humans different to the other hominids.

Moreover, since it has been suggested that most of the divergences that exist between humans and chimpanzees are due to differential regulation of gene expression (King & Wilson, 1975), we could hypothesize that the indels that were unique to the human lineage could be associated with different gene expression patterns among these two species.

In summary, our work provides new insights about indels located within cis-regulatory sequences. However in order to make a more comprehensive study, more analysis needs to be done. *In-vitro* gene expression assays could help evaluate how these fixed indels could cause differential expression in both species.

Chapter 2: The effect of human-specific promoter indels on gene expression

List of Figures

Figure 1 Neuroanatomy of humans……..…….38
Figure 2 pGL3-Basic Vector and pRL-SV40 circle map, Promega®……...…….41
Figure 3 A human (hg18_dna) and chimpanzee (panTro4_dna) alignment created by Geneious……..…….51
Figure 4 Primer Design for the chimpanzee promoter using Q5 Site- Directed Mutagenesis Kit from NEB.....52
Figure 5 Primer Design for the human promoter using Q5 Site- Directed Mutagenesis Kit from NEB……..52
Figure 6 Construction of Recombinant Plasmids……...……...58
Figure 7 Single Nucleotide Polymorphism (rs38839) located upstream the 5' of *MET*…….….61
Figure 8 Differences located near the transcription of *MET*……..62
Figure 9 Single Nucleotide Polymorphism (rs34939991) located upstream the 5' of *MET*…….….62
Figure 10 Single Nucleotide Polymorphism (rs62469050) located upstream the 5' of *MET*…….....64
Figure 11 Differences located near the transcription start site of *MET*…….66

List of tables

Table 1 Primers used for cloning promoters into expression vectors……42
Table 2 Composition and conditions of restriction-site adding PCRs……...…43
Table 3 Restriction reaction conditions…….....44
Table 4 Restriction reaction conditions…….45
Table 5 Conditions for the Dephosphorylation Reaction……..46
Table 6 Primers used for sequencing WT recombinant plasmids……49-50
Table 7 PCR Reaction for Mutagenesis……53
Table 8 Primers used for the mutagenesis assays……..……54
Table 9 Primers used for sequencing mutagenized recombinant plasmids……..55-57
Table 10. Comparison between the human reference genome and chimpanzee reference genome……59-60
Table 11 Differences found between the human reference genome (NCBI 36/ hg 18) and the inserted promoter fragment of MET……...……63
Table 12 Differences found between the inserted human fragment and the inserted chimpanzee fragment of *MET*……...…….……...65

Literature Review

From performing PCR and electrophoresis analysis to validate 64 indels located at no more than 2 kb upstream from its nearest transcription start site (TSS), we found two distinct genes with promoter-associated indels only in humans relative to all other primates (see previous Chapter). These genes were *MET* and *DMRTA2*, whose indels were located at 1076 bp and 1192 bp, respectively from the transcription start site (Mar.2006 (NCB1 36/hg18 assembly)).

The human gene *DMRTA2 (*DMRT-like family A2) is located on chromosome locus 1p32.3. It has 3 exons and it belongs to the *DMRT* (*doublesex (dsx) and male abnormal-3 (mab-3) related transcription factor*) gene family. This family of genes has eight members that are conserved among vertebrates and are involved in processes such as, somitogenesis, nervous system development, gonadal differentiation and gametogenesis. *DMRTA2* a member of this family, codes for a transcription factor that is involved in sexual development. More specifically, *DMRTA2* is necessary for the proper differentiation of oogonia during the female embryonic germ cell development (Poulain et al., 2014).

On the other hand, the human gene *MET* is located on chromosome locus 7q31.2. It spans approximately 126 kb, and encodes a tyrosine kinase transmembrane receptor of the hepatocyte growth factor/scatter factor (HGF/SF) (Sousa et al., 2009). Its signaling pathway starts upon the binding of its ligand, the hepatocyte growth factor. Binding of its ligand causes the oligomerization of MET and the intracellular phosphorylation of the tyrosine residues, who then serve as a landing site for the different adaptor proteins that activate different pathways. The MET signaling pathways have been linked to tissue remodeling, wound repair, organ homeostasis and cancer metastasis (Peng et al., 2014).

MET is commonly known as a proto-oncogene that is deregulated in different human tumors (Sousa et al., 2009). Several studies relate this gene with tumor development and progression when its expression is augmented. MET causes cancer development by different ways; 1)its overexpression in human tumors, 2) enhanced activation by its ligand, the hepatocyte growth factor, which may be abnormally secreted by cancer cells, 3) by some mechanism independent of hepatocyte growth factor and 4) structural alterations caused, for example, by missense mutations.

However, even though that this gene has been widely recognized in the context of cancer biology, distinct studies have found that reduced expression of MET is linked to neurodevelopmental disorders. For example, it has been found that MET has a critical role in the early brain development, being important in the

normal growth of the cerebral cortex in humans (Hedrick et al., 2012). In this same line of thinking, it is worth noting that the human gene *MET* is located in a region recognized by the International Molecular Genetic Study of Autism Consortium (IMGSAC), as the Autism Susceptibility Locus I (AUTS1) (Sousa et al., 2009). In this area there is a peak of association with autism that covers more than 40 Mb and contains more than 200 genes. Interestingly, several independent studies have suggested that variation in *MET* affects different areas within the cerebral cortex that are related to social behaviors present in individuals suffering from Autism Spectrum Disorder (ASD) (Hedrick, et al., 2012).

The ASD is a broader definition for autism. It is a diverse group of disorders characterized by a disability in social interactions, communication, and is also associated to repetitive conducts and limited interests. Its prevalence has increased over the last two decades and this could be attributed to improvement in diagnostic techniques and awareness. Autism is mainly present in males, with an onset before 3 years of age, and symptoms progressing throughout the entire life (Sousa et al., 2009).

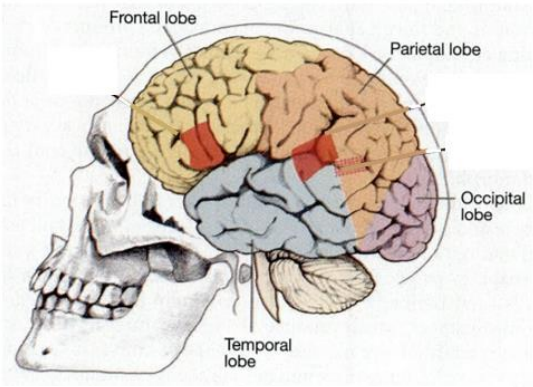


Figure 1 Neuroanatomy of humans. Lateral view of the left hemisphere of a modern human brain. Adapted from Carroll, S.B. (2003).

In humans, *MET* is expressed in the temporal, occipital and parietal cortexes (Figure 1); areas that are essential for the processing of social information (Rudie et al., 2012). However, it is more differentially expressed in the temporal lobe than in other cortical areas as this lobe is involved in language processing, emotional control and affective perception. This is particularly important because these regions are commonly affected in individuals showing ASD.

Three common variants in *MET* have been implicated with autism susceptibility. One of them is the rs1858830 C allele (G→C single nucleotide polymorphism (SNP)). It is found in the promoter region of *MET* and it has been linked to the decreased binding of nuclear proteins to the promoter region (Rudie et al., 2012). In addition, its existence has been shown to reduce *in vitro* gene transcription, by 50%. It is important to point out that common variants are not disease-causing, however they may increase the risk of any disorder. This is particularly interesting in people that have ASD, because if they possess this allele, they will present more severe social interactions and communication patterns, than individuals without the allele. In other words, the C allele rs 1858830 modulates the severity of the communication and social features in

individuals with ASD (Hedrick, et al., 2012). Moreover, the rs1858830 C allele has been related to reduce gray matter in developing children and adolescents (Plummer et al., 2013).
In addition to the rs1858830 C allele, two other alleles have been reported to increase the risk for ASD. There are the rs 38841 G allele(A→G) and the rs38845 A allele (G→A), both of them found in the first intron (Judson, Eagleson, & Levitt, 2011).

Another important feature of the AUTS1 is that it is located in one of the few regions of the human genome which is free of Neanderthal DNA. Known as Neanderthal deserts, these genomics regions are located in chromosomes 2, 21, 10 and 7, among others. Interestingly, some of these regions have structural or regulatory features and also often have genes that have been implicated in different diseases affecting the cognitive capabilities in present-day humans. For instance, some of these genes are DYRK1A, NRG3, CADPS2 and AUTS2. Mutations in NRG3 have been related to schizophrenia, and in the same way mutations in CADPS2 and AUTS2 have been associated with autism. The lack of Neanderthal sequences in certain regions of the human genome suggests that the genes in those regions are implicated in the cognitive development, and were positively selected after the divergence of the humans from the Neanderthal lineage 300,000-700,000 years ago (Green et al., 2010) to provide humans with species-specific advantages in cognitive development and social interactions.

Given that *MET* is located in a region desert of Neanderthal DNA and it is mainly expressed in the temporal lobe, a cerebral region involved in language processing, emotional control and affective perception, we might be able to explain why we observed a fixed deletion in the human promoter region relative to chimpanzee and gorilla. Perhaps this derived indel in the human promoter of *MET* could be a result of positive selection, augmenting gene expression and social interactions, which might explain some of the differences that are unique to the present-day humans in comparison to the other hominids, including cognitive capacity and language ability.

Introduction

MET and *DMRTA2* were found to have unique indels in the human lineage relative to the other primates (see previous chapter). *MET* which has been previously described to have a function in language processing, emotional control and affective perception (Rudie et al., 2012), presented a fixed deletion of 23 bp in the human lineage. Likewise, *DMRTA2*, which has been previously shown to have a role in the proper differentiation during the female embryonic germ cell development (Poulain et al., 2014), presented a fixed

insertion of 12 bp in the human lineage. Previously, Polavarapu et al. (2011) suggested that most of the divergences in gene expression patterns that exists between humans and chimpanzees are in organs and functions associated to the cognitive and fertility area. In addition, King and Wilson (1975) proposed that most the phenotypic differences that are observed between human and chimpanzee are a consequence of changes in gene expression of orthologous genes.

The expression of *MET* is given by various regulatory elements and its promoter comprises over 700 bp located upstream the transcription start site (approximately 410 bp) and within the first exon (approximately 360 bp). *MET*- promoter is also distinguished by an extremely high GC content and by the absence of a TATA box sequence (Trzyna, Majka, Faryna, Jurczyszyn, 2009). Therefore, our goal was to incorporate a fragment of the *MET* promoter, from the human and chimpanzee, into a luciferase plasmid to evaluate how the *MET*-indel could be associated to different expression patterns among these two species, using *in-vitro* gene expression assays.

Objectives

Overall Goal

• Assess how the *MET* indel could be associated to different patterns of gene expression between chimpanzees and humans using *in vitro* gene expression assays.

Specific Objectives

- Incorporate the human and chimpanzee promoter fragments of the *MET* gene into a luciferase vector and compare their expression in Human Embryonic Kidney Cells (HEK 293) and Pan troglodytes Fibroblast (chimpanzee) backgrounds.
- Perform an in vitro mutagenesis assay to create the corresponding insertion in the human-promoter fragment and the corresponding deletion in the chimpanzee-promoter fragment and compare their expression in human and chimpanzee backgrounds. In that way, "hybrid" promoter fragments with the indel of one species but the sequence of the other are constructed and assayed.

Material and Methods

Plasmid Selection

In order to study the effect of the indel on the expression of the human and chimpanzee-*MET* promoter, two different vectors were purchased from Promega®. One of them was the pGL3- Basic Vector which lacked eukaryotic and enhancer sequences; therefore we used this one to incorporate the human and chimpanzee-promoter fragments. The other was the pRL-SV40 Vector which contained the SV40 promoter and enhancer sequences, this vector was used as an internal control for the expression assays in the eukaryotic cells (See figure 2 for further details). Both vectors contained the luciferase gene which enabled us to measure promoter activity. The pGL3- Basic Vector contains a firefly (*Photinus pyralis*) luciferase reporter gene. Likewise, the pRL-SV40 Vector contains a Renilla (*Renilla reniformis*, also known as sea pansy) luciferase reporter gene.

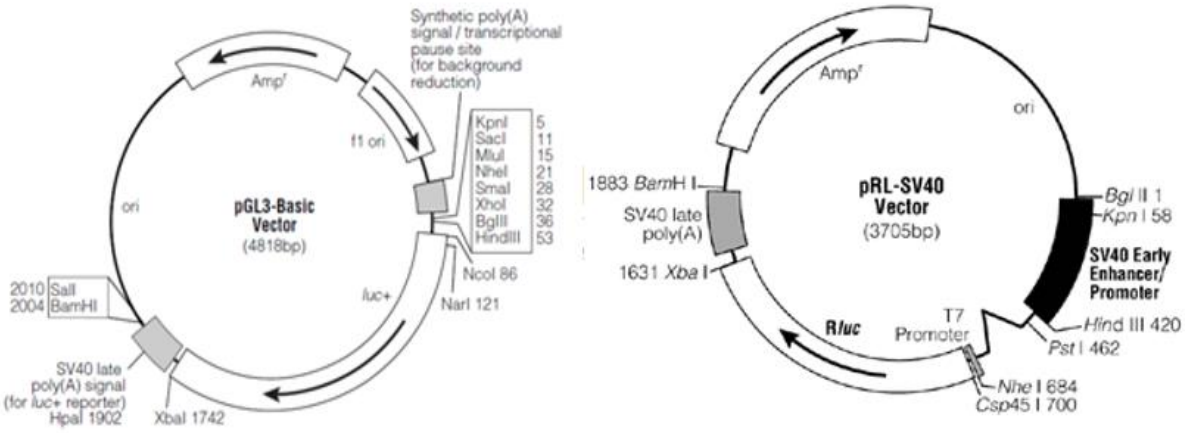


Figure 2 pGL3-Basic Vector and pRL-SV40 circle map, Promega®.

Analysis of the DNA fragments

The NEB Cutter tool V2.0 (http://www.labtools.us/nebcutter-v2-0/) was used to analyze the different restriction sites that were present within the human and chimpanzee promoter fragments. The human promoter fragment comprises from 116,098,490 to 116,099,792 at chromosome 7 ((Mar.2006 (NCB1 36/hg 18 assembly)) and its orthologous sequence from chimpanzee spans from 118,163,863 to 118,165,185 (Feb. 2011 (CSAC 2.1.4/panTro4)) at the same chromosome. Given that none of the restriction sites that were present in the Multiple Cloning Region of the pGL3-Basic Vector were found inside the promoter fragments, we had to add the restriction sites to the terminus of the fragments using 5' tailed primers.

Preparation of Oligonucleotides

The 5' terminus of each primer contained a clamp of 6 nucleotides in length to provide a landing site for the restriction enzymes. The middle portion of each primer contained the restriction sites for the enzymes that were used during the digestion reaction. The restriction sites that were selected are present within the Multiple cloning region of the pGL3-Basic Vector (*Mlu*I and *Bgl*II), but were not present within the promoter fragments that we wanted to incorporate inside the vector. For this reason we added the restriction site for *Mlu*I (ACGCGT) to the forward primer and the restriction site for *Bgl*II (AGATCT) to the reverse primer. The 3'end of the forward primer was an exact complement of 21 consecutive bases at the selected site in the target DNA. The 3' end of the reverse primer was an exact complement of 20 consecutive bases at the selected site in the target DNA (Table 1). In this way, the promoter would be located in the vector in the proper orientation to activate transcription of the luciferase gene. The potential usefulness of the designed primers were studied using the OligoAnalyzer Tool 3.1 from IDTDNA (http://www.idtdna.com/calc/analyzer), to ensure non-major annealing effects between primers or auto-annealing.

Table 1. Primers used for cloning promoters into expression vectors

Primers for MET	Length	GC Content	Melting Temperature
5'- CGG CAG ACG CGT TGA ACC TGC ATA GTA ACT GTT -3' (Forward Primer)	33 bp	51.5%	65.7
5'- GAC GTG AGA TCT TCT CAG CAA GTC AGC TGT CG -3' (Reverse Primer)	32 bp	53.1 %	64.1

PCR Amplification

The PCR was performed in 50 µl volume using the Q5 High-Fidelity DNA Polymerase (NEB #M0491), the 5x Q5 High GC Enhancer and an annealing temperature of 66^0C (Table 2).

Table 2. Composition and conditions of restriction-site adding PCRs

Components	50 ul Reaction	Final Concentration
5X Q5 Reaction Buffer	10 µl	1x
10 mM dNTPs	1 µl	200 µM
10 µM Forward Primer	2.5µl	0.5 µM
10 µM Reverse Primer	2.5 µl	0.5 µM
Template DNA	5 µl (30ng/µl)	3 ng/µl
Q5 High-Fidelity DNA Polymerase	0.5 µl	0.02 U/µl
5x Q5 High GC Enhancer (optional)	10 µl	1x
Nuclease-Free Water	18.5 µl	

Thermocycling Conditions	Temperature	Time
Initial Denaturation	98^0C	30 seconds
Denaturation	98^0C	8 seconds
Annealing (35 cycles)	66^0C	20 seconds
Extension	72^0C	45 seconds
Final Extension	72^0C	2 minutes
Hold	4^0 C	∞

Analysis of PCR Products

In order to analyze the PCR products, an agarose gel at 1.2% was run. The expected amplicon size for the chimpanzee promoter of *MET* was of 1347 bp, including the additional bases of the incorporated restriction sites and the landing sites. On the other hand, the expected amplicon size for the human- promoter fragment was 1327 bp, including the additional bases of the incorporated restriction sites and the landing sites.

Purification of the PCR Product

In order to remove the residual primers, dNTPs and the DNA Polymerase from the PCR products, a purification step was performed using Wizard PCR Preps DNA Purification System from Promega. After purification the DNA was quantified using a spectrophotometer.

Restriction Enzyme Reaction

The digestion of the human and chimpanzee promoter for the *MET* gene was performed using two different enzymes; *Bgl*II (NEB #R0144S) and *Mlu*I (NEB #R0198S). The reaction was performed using the two restriction enzymes simultaneously. In order to select the appropriate buffer for the reaction and to prevent the star activity from the different restriction enzymes, the Double Digest Finder tool from NEB (http://www.labtools.us/nebcutter-v2-0/) was used. The NEB Buffer 3.1 was identified as the most appropriate buffer for this reaction. The reaction was achieved using 1ul of each one of the restriction enzymes with 1 μg of the purified DNA in a final volume of 50 ul. Given that the reaction was performed using Time-Saver qualified restriction enzymes from NEB, the incubation time was 2 hours. However, this could be performed in less time (5-15 minutes) or even overnight (Table 3).

Table 3. Restriction reaction conditions

Reaction for the DNA promoter fragments	
Restriction Enzymes	
Bgl*II *(10,000 U/ml)	1 μl
Mlu*I *(10,000 U/ml)	1 μl
DNA (promoter fragment)	1 μg (33 μl)
Buffer 3.1	5 μl
dH_2O	10 μl
Total Reaction Volume	50 μl
Incubation Temperature	37^0 C [1]
Incubation Time	2 hours

[1]The incubation temperature was the same for both restriction enzymes.

The digestion of the pGL3-Basic Vector was performed with the same enzymes; *Bgl*II and *Mlu*I. The reaction was performed using the NEBuffer 3.1 and the incubation time was 2 hours at 37^0 C. For this reaction, 3 µg of the plasmid, 3 µl of each enzyme, and 7µl of buffer were used. The final volume was increased to 70 µl with water (Table 4).

Table 4. Restriction reaction conditions

Reaction for the pGL3-Basic Vector	
Restriction Enzymes	
***Bgl*II**	3 µl
***Mlu*I**	3 µl
DNA (pGL3- Basic Vector)	3 µg (3 µl)
Buffer 3.1	7 µl
dH_2O	54 µl
Total Reaction Volume	70 µl
Incubation Temperature	37^0 C [1]
Incubation Time	2 hrs

[1]The incubation temperature was the same for both restriction enzymes.

DNA Analysis

After the digestion reactions, the fragments and plasmid were analyzed running a 1.2 % agarose-gel electrophoresis with 1 kb DNA Ladder (N3232L) from NEB as molecular marker. Next, both the fragments and plasmid were purified using Wizard PCR Preps DNA Purification System from Promega. The latter step was necessary because the *Bgl*II-restriction enzyme was not heat inactivable, therefore it needed to be removed prior to the dephosphorylation step.

Dephosphorylation

To prevent the re-circularization of the digested pGL3-Basic Vector during the ligation step, a dephosphophorylation reaction was performed using Shrimp Alkaline Phosphatase (rSAP, #M0371) from NEB (Table 5). This step removes the 5' phosphate of the plasmid and prevents the intramolecular ligation of it. The reaction was performed as follows.

Table 5. Conditions for the Dephosphorylation Reaction

Dephosphorylation Reaction	
Digested pGL3-Basic Vector	1 µg
Cut Smart Buffer (10x)	2µl
Shrimp Alkaline Phosphatase (rSAP) (1Unit/µl)	1 µl
Nuclease-free water	up to total volume 20µl
Incubation	37^{O} C for 30 minutes
Heat Inactivation	65^{O} C for 5 minutes

Vector and Insert Joining (DNA Ligation)

The ligation was performed using an insert:vector molar ratio of 3:1 Before performing the procedure, the master mix was transferred to ice and mixed by finger flicking. The volume of the reaction was adjusted to a final volume of 10 µl, using 5 µl of Instant Sticky-end Ligase Master Mix (NEB #M0370S) and 5 µl of the combination of the insert and vector (volume was adjusted with dH_2O). The total amount of the vector was of 20 ng and of the insert was 12 ng.

After the addition of all components of the reaction the solution was mixed thoroughly by pipetting up and down, 7-10 times. Then the sample was transferred to ice and stored at -20^{0}C. There was not an incubation time given that the kit used is optimized for an instant ligation.

Transformation

NEB 5-alpha Competent *E.coli* (High Efficiency) Cells (NEB #C2987H) were transformed with the ligation reaction. The cells were thawed on ice and 50 µl of them were transferred into a pre-cooled 1.5 ml microcentrifuge tube. Afterward, 2 µl of the ligation reaction was added to the cells and mixed by finger flicking. The cells were incubated on ice during 30 minutes. A heat shock was performed at 42^{O}C for 30 seconds, and the tube was returned to ice for 2 minutes. Next, 950 µl of recovery media (SOC) at room temperature was added to the tube and incubated for one hour at 37^{O}C with shaking (200-250 rpm). Then 100 µl of the outgrowth was spread onto antibiotic selection plates and incubated overnight at 37^{O}C. After incubation, the agar plates were observed and each observed colony was inoculated into a tube containing 1.5

ml of LB medium with ampicillin at a final concentration of 100μg/ml. After that, the bacterial cultures were incubated 16 hours at 37^{O}C, in a rotator shaker. The cell culture was collected and the High Pure Plasmid Isolation Kit (Roche, Cat #11754785001) was used to purify the plasmid.

-Controls used during the transformation

Cells were transformed with:

- An uncut vector to check cell viability.
- A digested, non-dephosphorylated vector plus ligase to check the functionality of the ligase.
- A digested dephosphorylated-vector without ligase to check the efficiency of the digestion with the restriction enzymes.
- A digested dephosphorylated-vector with ligase to check the efficiency of the dephosphorylation reaction.

-Preparation of Antibiotic Selection Plates

The LB Agar plates were prepared with the following ingredients (this concentration was for 500 ml of LB Agar)

- 5 g of NaCl
- 5 g of Tryptone
- 2.5 g of Yeast Extract
- 7.5 g of Agar

After incorporating all the ingredients, the volume was increased up to 500 ml with deionized water. The solution was put on a stirring hot plate and heated to boil for 1 minute. Then it was autoclaved for 20 minutes and allowed to cool down at 55^{O}C. Next, 500 μl of ampicillin at a concentration of 100 μg/ml was added to the LB Agar. Petri dishes of 10 cm were used with approximately 10 ml of LB Agar.

-Preparation of LB Liquid Medium

The LB liquid Medium was prepared using the following ingredients (this concentration was for 1L of LB)

- 10 g of Tryptone
- 5 g of Yeast Extract
- 10 g of NaCl

After incorporating all the ingredients, the volume was increased up to 1L with deionized water. The solution was put on a stirring hot plate and heated to boil for 1 minute. Then it was autoclaved for 20 minutes and allowed to cool down at 55^{O}C. Next, 1 ml of ampicillin was added to the liquid LB to make a final concentration of 50 µg/ml. Then, 1.5 ml of LB was added to a batch of test tubes.

Plasmid Analysis

In order to verify that the selected clones had the recombinant fragments, we performed a restriction enzyme digestion reaction and then the selected plasmids were sent for sequencing. The expected amplicon size for the chimpanzee promoter fragment was 1323 bp and the expected amplicon size for the human promoter fragment was 1303 bp.

Human and Chimpanzee Plasmids

In order to verify if the promoter fragments were incorporated into the pGL3-Basic Vector, we performed a restriction enzyme reaction with *Mlu*I. This reaction was performed as previously mentioned (Table 3) but only with one enzyme. This enabled us to linearize the recombinant plasmid and compare its weight with a pGL3-Basic Vector without the insert. After the restriction enzyme reaction, the digested plasmids were analyzed using a 0.8% agarose-gel electrophoresis with 1 kb DNA Ladder (N3232L) from NEB as the molecular marker.

The recombinant plasmids that demonstrated an increased weight, when visualized in the agarose-gel, were sent for sequencing. Five hundred nanograms of the recombinant plasmid with 1 µl of the corresponding primer (concentration of 10 uM) were sent to Nevada Genomics Center (http://www.ag.unr.edu/genomics/default.html) for DNA sequencing. The primers used are shown in Table 6. The Reporter Vector Primer 3 (RVprimer3) was used to sequence clockwise across the upstream Multiple Cloning Region. The GLprimer2 was used to sequence counterclockwise upstream of the luciferase gene. The primer 5, 8, 9, 10 and 16 were located within the insert, in a clockwise manner. The reported positions for the primers that were located within the insert were retrieved from UCSC genome browser (Mar.2006 (NCB1 36/hg 18 assembly)) and these positions were the first base pair extended by the 3'end of the primer (Table 6). The primer 7 was located within the insert in

a counterclockwise manner. For a better understanding of the location of primers within the pGL3-Basic Vector see the figure 1 (pGL3 Luciferase Reporter Vector circle map).

Table 6. Primers used for sequencing WT recombinant plasmids (cont.)

Primer	***Location***	***GC Content***	***Melting Temperature***	***Length***
Humans				
RV primer3 (Forward) *5'- CTA GCA AAA TAG GCT GTC CC -3'*	Position 4760-4779 within the pGL3-Basic Vector.	50%	53.1	20 bp
GLprimer2 (Reverse) *5'- CTT TAT GTT TTT GGC GTC TTC CA -3'*	Position 89-111 within the pGL3-Basic Vector.	39.1%	54.1	20 bp
Primer 5 (Forward) *5'- GAA CTG AAC CTG CAT AGT AAC TG -3'*	Located at chr7:116,098,509 bp, (Mar.2006 (NCB1 36/hg 18 assembly)).	43.5%	53.4	23 bp
Primer 7 (Reverse) *5'- ACT CGG CTC CGC ATC TGC TCA CAA AGC -3'*	Located at chr7:116,099,721 bp, (NCB1 36/hg 18 assembly)).	59.3%	66.3	27 bp
Primer 16 (Forward) *5'- GGG ACA ATT CGT CCA TCC ACT TC-3'*	Located at chr7:116,099,003 bp, (NCB1 36/hg 18 assembly)).	52.2 %	57.9	23 bp
Chimpanzee				

RV primer3 (Forward) *5'- CTA GCA AAA TAG GCT GTC CC -3'*	Position 4760-4779 within the pGL3-Basic Vector.	50%	53.1	20 bp
Primer 5 (Forward) *5'- GAA CTG AAC CTG CAT AGT AAC TG -3'*	Located at chr7: 118,163,882 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).	43.5%	53.4	23 bp
Primer 7 (Reverse) *5'- ACT CGG CTC CGC ATC TGC TCA CAA AGC -3'*	Located at chr7: 118,165,114 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).	59.3%	66.3	27 bp
Primer 8 (Forward A) *5'- ATC ATT GGG ACA ATT CGT CCA TCC -3'*	Located at chr7: 118,164,392 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).	45.8 %	57.3	24 bp
Primer 9 (Forward B) *5'-TGT GCT AAC TTC AGA CTG CCT GAG C -3'*	Located at chr7: 118,164,655 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).	52%	60.8	25 bp
Primer 10 (Forward C) *5'-GAT TTC CCT CTG GGT GGT GCC AGT C -3'*	Located at chr7: 118,164,821 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).	60%	63.3	25 bp

Mutagenesis

After the incorporation of the human promoter fragment and the chimpanzee promoter fragment into the pGL3- Basic Vector, our goal was to perform a site directed mutagenesis to confer a mutation into our recombinant pGL3-Basic Vector. Given that the human-promoter fragment of *MET* contained a deletion of 23 bp (Figure 3), we wanted to incorporate that deletion into the chimpanzee promoter of the same gene. On the other hand, given that the chimpanzee-promoter fragment of *MET* contained an insertion relative to the humans (Figure 3), we wanted to incorporate that insertion into the human promoter fragment of the same gene. This procedure was accomplished using the Q5 Site- Directed Mutagenesis (E0554S).

Figure 3 A human (hg18_dna) and chimpanzee (panTro4_dna) alignment created by *Geneious*. The deletion of 23 bp is represented by the dotted lines in the human promoter fragment.

PCR Amplification

The first step was a PCR reaction using a master mix formulation provided with the Q5 Hot Start High-Fidelity DNA Polymerase (Table 7) and a set of costumed primers (Table 8). The primers used for the PCR Reaction were designed using NebaseChanger (http://nebasechanger.neb.com/). The melting temperature for the PCR reaction was provided by the NebaseChanger tool. To create the deletion in the chimpanzee promoter, the forward and reverse primers were created in such a way that they flanked the region to be deleted (Figure 4). To create the insertion in the human promoter, we wanted to incorporate 23 bp (CCCGTATGAGACGCTTGAACAAG), in the corresponding area relative to the chimpanzee fragment (Figure 5).

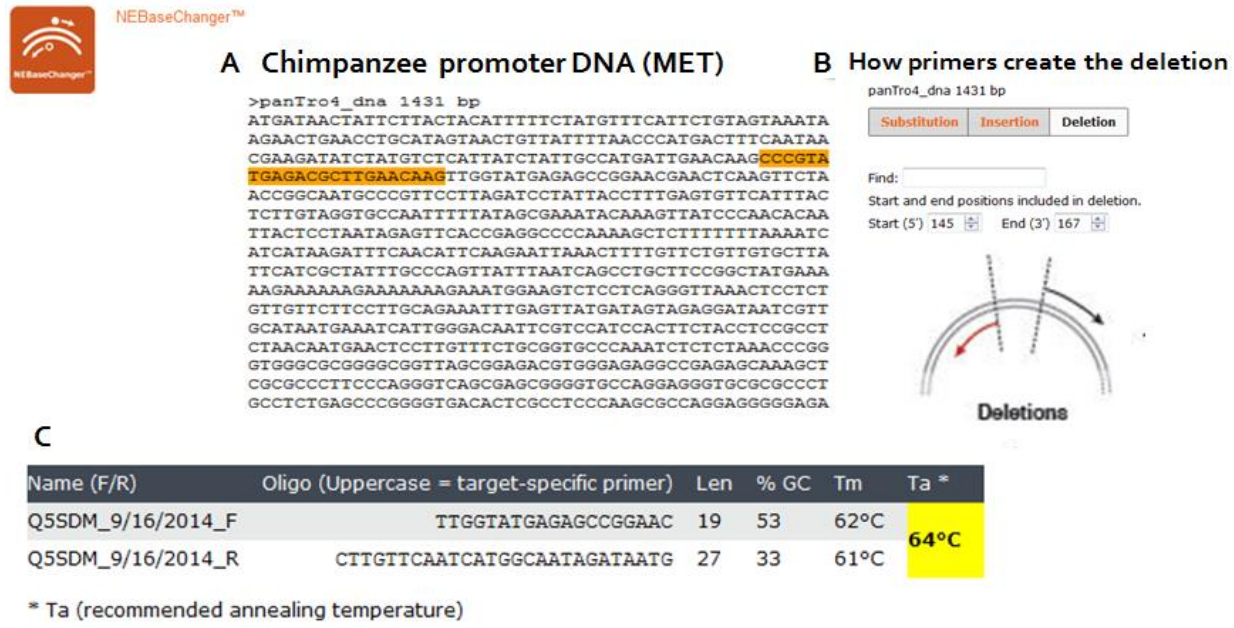


Figure 4 Primer Design for the chimpanzee promoter using Q5 Site- Directed Mutagenesis Kit from NEB. A) The chimpanzee-promoter fragment; the orange area represents the area that we wanted to delete. B) The deletion was created by designing primers that flank both sides of the area to be deleted (black arrow represents the forward primer and red arrow represents the reverse primer). C) Primers generated by the NEBaseChanger tool; the forward and reverse primer.

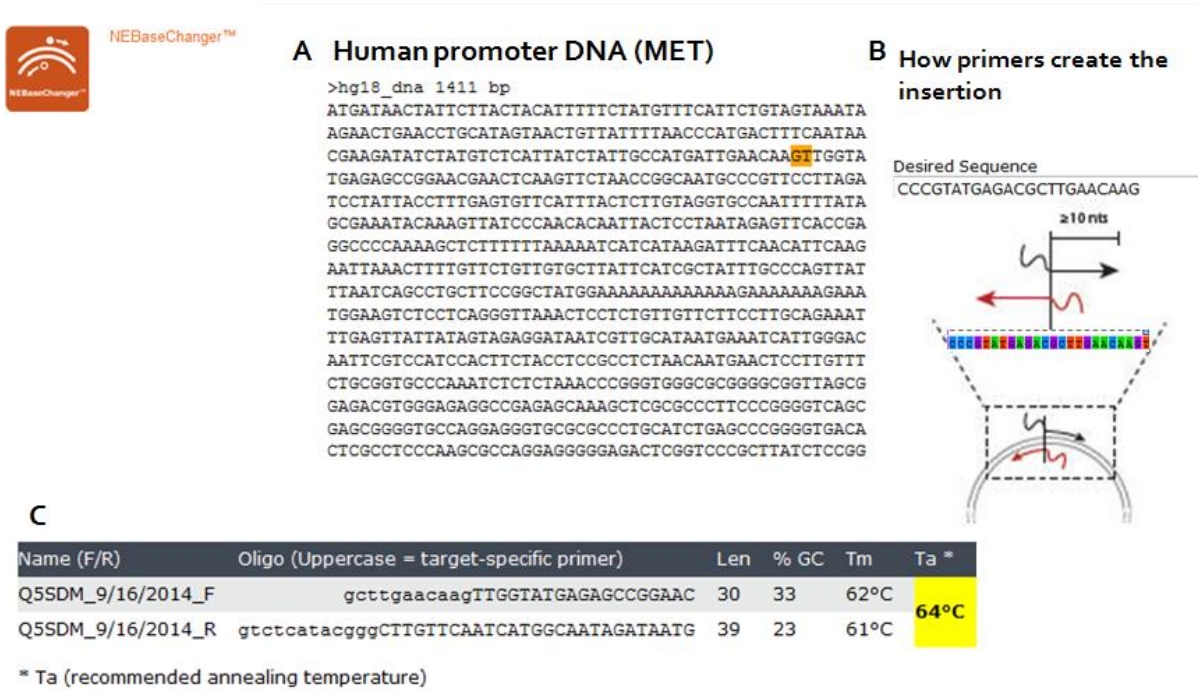


Figure 5 Primer Design for the human promoter using Q5 Site- Directed Mutagenesis Kit from NEB. A) The human-promoter fragment, the human genome 18 was used as a reference. The orange area represents the area in which we incorporated the insertion (between

Guanine and Thymine). B) The insertion of 23 bp was created by incorporating half of the sequence into the 5' end of both primers (black arrow represents the forward primer and red arrow represents the reverse primer). C) Primers generated by the NEBaseChanger tool; the forward and reverse primer. Lower case letters represent inserted sequences.

Table 7. PCR Reaction for Mutagenesis

PCR Reaction		**Final Concentration**
Q5 Hot Start High-Fidelity 2x Master Mix	12.5 µl	1x
10 µM Forward Primer	1.25 µl	0.5 µM
10 µM Reverse Primer	1.25 µl	0.5 µM
pGL3-Basic Vector (1-25 ng/ul)	1 µl	0.4-1 ng/µl
Nuclease-free water	9.0 µl	
Thermocycling Conditions	**Temperature**	**Time**
Initial Denaturation	98^{O}C	30 seconds
	98^{O}C	10 seconds
Annealing (25 cycles)	64 ^{O}C	30 seconds
	72^{O}C	80 seconds
Final Extension	72^{O}C	2 minutes
Hold	4^{O}C	∞

Table 8. Primers used for the mutagenesis assays

	Length	GC Content	NEBasechanger TM
Primers used for the incorporation of the deletion in the chimpanzee promoter			
Forward TTGGTATGAGAGCCGGAAC	19 bp	53%	64^{O}C
Reverse CTTGTTCAATCATGGCAATAGATAATG	27 bp	33%	
Primers used for the incorporation of the insertion in the human promoter			
Forward gcttgaacaagTTGGTATGAGAGCCGGAAC	30 bp	33%	64 ^{O}C
Reverse gtctcatacgggCTTGTTCAATCATGGCAATAGATAATG	39 bp	23%	

*Treatment and Enrichment (Kinase, Ligase and Dpn*I*)*

The second step was an incubation of 5 minutes, at room temperature, of 1 µl of the PCR product with 5 µl of the 2x KLD Reaction Buffer (final concentration of 1x), 1 µl of the 10x KLD Enzyme Mix (final concentration of 1x) and 3 µl of Nuclease-free water. The enzyme mix contained a kinase, a ligase and *Dpn*I. In order to enrich the reaction with the plasmid containing the mutations, we used *Dpn*I. The purpose of this restriction enzyme was to digest the plasmid without the mutation, because they were previously generated by Dam^{+}/DCM^{+} competent cells (NEB #C2987H) in the previous transformation. Therefore, *Dpn*I only digested the wild type plasmid which had fully methylated G $^{MET^{\wedge}}$ATC sequences. The purpose of the kinase was to phosphorylate and the ligase enabled the circularization of the PCR product (plasmid with mutations).

Transformation

NEB 5-alpha Competent *E.coli* (High Efficiency) Cells (NEB #C2987) were transformed with the KLD mix of the previous step. The cells were thawed on ice and 50 µl of them were transferred into a pre-cooled 1.5 ml microcentrifuge tube. After that, 5 µl of the KLD mix was added to the cells and mixed by finger flicking.

The cells were incubated for 30 minutes on ice, then a heat shock was applied for 30 seconds, at 42^{O}C. Next, the cells were incubated 5 minutes on ice, and 950 µl of SOC at room temperature was added to the tube and incubated for one hour at 37^{O}C with shaking (200-250 rpm). After incubation, 100 µl of the outgrowth was spread onto the appropriate selection plates and incubated overnight at 37^{O} C. After incubation, the agar plates were observed and each observed colony was inoculated into a tube containing 1.5 ml of LB medium with ampicillin and incubated for 16 hours at 37^{O}C, in a rotator shaker. The cell culture was collected and the High Pure Plasmid Isolation Kit (Roche) was used to purify the plasmid.

Plasmid Analysis

In order to verify that the selected clones had the desired mutations, we performed a restriction enzyme digestion reaction and then the selected plasmids were sent for sequencing. The expected length of the chimpanzee promoter fragment with the corresponding deletion was 1,302 bp. Meanwhile, the human promoter fragment with the corresponding insertion was 1,327 bp.

Chimpanzee plasmid

To confirm the incorporation of the deletion into the chimpanzee-promoter fragment, we did a digestion with *Mlu*I as previously described but with only one restriction enzyme (Table 3). This enabled us to linearize the mutated plasmids and compare their weight with a pGL3-Basic Vector without any mutation. After the restriction enzyme reaction, the digested mutated-plasmids were analyzed running a 1.2% agarose-gel electrophoresis with 1 kb DNA Ladder (N3232L) from NEB. The chimpanzee plasmids that demonstrated a decreased weight, when visualized in the agarose-gel, were sent for sequencing. Five hundred nanograms of the mutated plasmids with 1 µl of the primers (Table 9) (concentration of 10 µM) were sent to Nevada Genomics Center for DNA for sequencing (http://www.ag.unr.edu/genomics/default.html).

Table 9. Primers used for sequencing mutagenized recombinant plasmids (cont.)

Primer	***Location***	***GC Content***	***Melting Temperature***	***Length***
Chimpanzee				
RV primer3	Position 4760-	50%	53.1^{O}C	20 bp

(Forward) *5'- CTA GCA AAA TAG GCT GTC CC -3'*	4779 within the pGL3-Basic Vector.			
Primer 5 (Forward) *5'- GAA CTG AAC CTG CAT AGT AAC TG -3'*	Located at chr7: 118,163,882 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).	43.5%	53.4^{O}C	23 bp
Primer 8 (Forward A) *5'- ATC ATT GGG ACA ATT CGT CCA TCC -3'*	Located at chr7: 118,164,392 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).	45.8 %	57.3^{O}C	24 bp
Primer 9 (Forward B) *5'-TGT GCT AAC TTC AGA CTG CCT GAG C -3'*	Located at chr7: 118,164,655 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).	52%	60.8^{O}C	25 bp
Primer 10 (Forward C) *5'-GAT TTC CCT CTG GGT GGT GCC AGT C-3'*	Located at chr7: 118,164,821 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).	60%	63.3^{O}C	25 bp
Primer 13 (Forward) *5'-TAA CTT CAG ACT GCC TGA GCT GG 3'*	Located at chr7: 118,164,658 bp (Feb. 2011 (CSAC 2.1.4/panTro4)).	52.2 %	58.7^{O}C	23 bp
Primer 14	Located at chr7:	60 %	60^{O}C	20 bp

(Forward) *5'-AGG CAG ACA GAC ACG TGC TG-3'*	118,164,893 bp, (Feb. 2011 (CSAC 2.1.4/panTro4)).

Human plasmid

To confirm the incorporation of the insertion into the human-promoter fragment, we did a digestion with two different restriction enzymes. The reaction was performed using 0.5 µl of *Bsa*XI (2,000 U/ ml), 0.5 µl *Eco*RV *HF* (20, 000 U/ml) and 2.5 µl of CutSmart Buffer (all components were from NEB) in a total volume of 25 µl. The incubation time was 40 minutes at 37^{O}C. After the restriction enzyme reaction, the digested mutated-plasmids were analyzed running a 3% agarose-gel electrophoresis with 100 bp DNA Ladder (N3231L) from NEB. The incorporation of the insertion into the human plasmid was expected to produce a fragment of 203 bp, when digested with the previous enzymes. The human plasmids that presented a band near 203 bp were sent for sequencing. Five hundred nanograms of the mutated plasmids with 1µl of the corresponding primer (concentration of 10 µM) were sent to Nevada Genomics Center (http://www.ag.unr.edu/genomics/default.html).

Future Studies: MET 5' promoter luciferase Assays

Overview

The objective of this assay will be to measure the luciferase expression of our wild type and mutated recombinant plasmids in humans and chimpanzee backgrounds (Figure 6). The Luciferase belongs to a family of enzymes that are naturally produced by different species, mainly in the genus *Lampyridae* (firefly) (Sarah, 2011). This enzyme catalyzes the oxidation of a small substrate called luciferin, bringing it to an excited electronic state. When luciferin returns back to its ground state, it emits luminescence. Therefore, this feature will be exploited, in order to study the transcriptional activity of the *MET*-promoter fragments from human and chimpanzee lineages.

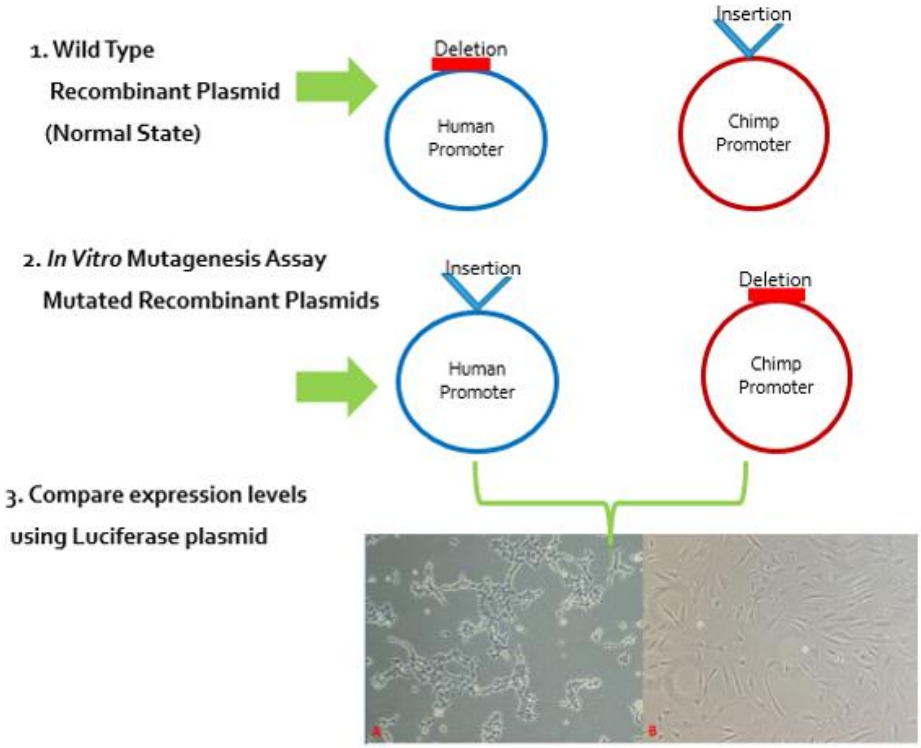


Figure 6 Construction of Recombinant Plasmids. 1) Wild Type Recombinant Plasmids: the human promoter fragment and chimpanzee promoter fragment were incorporated into the pGL3-Basic Vector. 2) Mutated Plasmids: the chimpanzee indel (insertion) was incorporated into the human promoter fragment and the human indel (deletion) was incorporated into the chimpanzee promoter fragment. 3) Comparison of expression levels: Human Embryonic Kidney cells (HEK 293, panel A) and Fibroblasts cells from chimpanzee (panel B) will be transfected transiently with the wild type recombinant plasmids and the mutated recombinant plasmids. The relative light units generated by the luciferase reporter gene will be measured.

Results

We incorporated the human promoter fragment of *MET* (chr7: 116,098,490 to 116,099,792) (Mar.2006 (NCB1 36/hg 18 assembly)) and its orthologous sequence from chimpanzee (chr 7: 118,163,863 to 118,165,185 (Feb. 2011 (CSAC 2.1.4/panTro4))) into a luciferase vector (pGL3- Basic Vector). In addition we created the corresponding deletion into the chimpanzee promoter fragment (Figures 6), as previously described. However, when we were analyzing the human plasmid with the chimpanzee insertion the recombinant fragment was not successfully incorporated. When performing the *in-silico* analysis of this mutated recombinant plasmid, we found an extensive deletion of the human promoter fragment within the

vector. Therefore, we sent the human plasmid to GenScript (http://www.genscript.com/), with instructions to insert the indel of 23 bp into the recombinant plasmid, using the restriction sites for *MluI* and *BglII*.

Discussion

We successfully incorporated the human promoter fragment and its orthologous sequence from chimpanzee into the luciferase vector (pGL3 Basic Vector). In addition, we performed the corresponding deletion into the chimpanzee promoter fragment. As previously mentioned, given that the mutated-human recombinant plasmid was not constructed effectively with the described procedures, we ordered it from GenScript. We also analyzed this mutated recombinant plasmid and confirmed that the insertion was correctly incorporated.

After performing the construction of the different recombinant plasmids, we analyzed the DNA promoter fragment from both human and chimpanzee. We started our analysis comparing the human reference genome (Mar.2006 (NCB1 36/hg 18 assembly)) against the chimpanzee reference genome (Feb. 2011 (CSAC 2.1.4/panTro4)) (Table 10).

Table 10. Comparison between the human reference genome and chimpanzee reference genome (cont.)

Nucleotide observed in the human reference genome ((NCB1 36/hg 18)	**Location at the reference human genome (NCB1 36/hg 18)**	**Position relative to the transcription start site (NCB1 36/hg 18)**	**Nucleotide observed in the chimpanzee reference genome (CSAC 2.1.4/ panTro4 Assembly)**	**State in humans**	**Ancestral allele in the human lineage**
Adenine	Chr7:116, 098, 755	-940 bp	Thymine	rs38839	Thymine
Guanine	Chr7:116, 098, 859	-836 bp	Deletion	Fixed	Deletion
Adenine	Chr7: 116, 098, 860	-835 bp	Deletion	rs34939991	Adenine
Adenine	Chr7: 116,098, 866	-829 bp	Guanine	Fixed	Guanine
Thymine	Chr7: 116,098,	-751 bp	Guanine	Fixed	Thymine

	944				
Guanine	Chr7: 116, 099, 126	-569 bp	Adenine	Fixed	Guanine
Adenine	Chr7: 116,099, 166	-529 bp	Cytosine	rs184953	Cytosine
-	Chr7:116,099, 461 + 1	-233 bp	Cytosine	Fixed	Cytosine
Adenine	Chr7: 116, 099, 473	-222 bp	Guanine	Fixed	Guanine
Adenine	Chr7: 116, 099, 506	-189 bp	Guanine	Fixed	Guanine
Guanine, Guanine	Chr7: 116, 099, 665-116, 099, 666	-30 bp, -29 bp	-, -	fixed	Guanine, Guanine

Human reference genome versus chimpanzee reference genome

The first nucleotide substitution was located at chr7:116, 098, 755 (human genome 18) and was identified as the single nucleotide polymorphism (SNP) rs38839, as reported by the SNP database from NCBI. As expected, the ancestral allele for this SNP was a thymine (Figure 7). We also observed an insertion of two nucleotides (GA) in the human reference genome relative to the chimpanzee reference genome. The adenine that was located at chr7: 116, 098, 860 was identified as the SNP rs34939991 and, interestingly, the ancestral allele was adenine. It is important to note, that the deletion of two nucleotides that was observed in the chimpanzee lineage was predicted by the Multiz Alignments of 44 Vertebrates Track (Conservation Track) from UCSC Genome Browser. The third substitution was observed at chr7: 116,098, 866. We did not find any reported SNP for this position, however the guanine observed in the chimpanzee lineage was also predicted by the conservation track from UCSC Genome browser. The fourth difference was located at chr7: 116,098,944, and for this one, we did not find any reported SNP. The fifth difference was a guanine found at chr7: 116, 099, 126 in the human reference genome relative to the adenine observed in the chimpanzee

reference genome. There were no SNP reported, however the adenine observed in the chimpanzee lineage was confirmed by the conservation track provided by the UCSC Genome browser. The sixth difference was located at chr7: 116, 099, 166 in the human reference, relative to the chimpanzee reference. This substitution corresponds to the SNP rs184953, as reported by the SNP database from NCBI. As expected, the ancestral allele is cytosine, which was observed in the chimpanzee reference. Interestingly, a deletion was also found in the human reference genome relative to an insertion of cytosine in the chimpanzee lineage. The insertion in the chimpanzee lineage was located between the coordinates; chr7: 116, 099, 461 and 116, 099, 462 (coordinates from human reference genome). The insertion in the chimpanzee lineage that was predicted by the conservation track of the UCSC genome browser is ancestral. The eighth substitution was found at chr7: 16, 099, 473. In this position an adenine was observed in the human lineage relative to a guanine in the chimpanzee lineage. There were not reported SNP for this position, however the guanine observed in the chimpanzee lineage was also confirmed by the conservation track from genome browser and was ancestral. Likewise, we observed an adenine in the human lineage at chr7: 116,099, 506, relative to a guanine in the other primate. The guanine in the chimpanzee lineage was also confirmed by the conservation track provided by UCSC. Lastly, we observed an insertion of two nucleotides in the human lineage relative to the chimpanzee lineage. The insertion in the human lineage was located at chr7: 116, 099, 665 – 116, 099, 066. Interestingly, the deletion in the chimpanzee lineage was not confirmed by the conservation track from the genome browser, but was observed in the chimpanzee reference genome (Feb.2011 (CSAC 2.1.4/ panTro4 Assembly)).

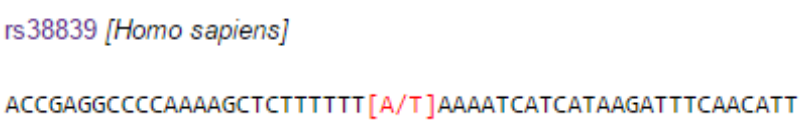


Figure 7 Single Nucleotide Polymorphism (rs38839) located upstream the 5' of *MET*. A thymine was found in the human promoter fragment of *MET* that was incorporated in the pGL3-Basic plasmid.

In summary, we found that three of the differences that were observed, were reported as SNPs and eight differences were observed to be fixed in the human lineage (Figure 8). Interestingly, some differences are located nearby the rs18518830 (located -20 bp, relative to the transcription start site), which have been previously shown to be involved in the severity of the communication and social feature in individuals with ASD.

Figure 8 Differences located near the transcription of *MET*. Summarized representation of the differences observed among the human reference genome (Mar.2006 (NCBI 36/hg 18 assembly)) and the chimpanzee reference genome (Feb. 2011(CSAC 2.1.4/panTro4 Assembly)). The blue stars represent the SNPs that were found in the promoter region. All the positions are relative to the transcription start site of *MET*.

Human reference genome versus inserted human fragment

After analyzing both promoter fragments (reference genomes), we compared the human reference genome (Mar.2006 (NCB1 36/hg 18 assembly)) against the inserted human fragment. When performing the analysis we only found three differences (Table 11); two nucleotide substitutions and one indel. The first nucleotide substitution was located at chr7:116, 098, 755 (human genome 18) and was identified as the single nucleotide polymorphism (SNP) rs38839, as previously mentioned (Figure 7). Interestingly, the ancestral allele for this SNP was a thymine, and this one was found in the inserted promoter fragment. Likewise, the deletion that was found at chr7: 116,098,872 corresponds to the SNP rs36222678 (also known as rs34939991) as reported by the SNP database from NCBI (Figure 9). The ancestral allele for this SNP was adenine. The last nucleotide substitution that was found in the inserted fragment was located at chr7: 116, 099, 166, and corresponds to the SNP rs184953. The ancestral allele for this SNP was cytosine, the one observed in the inserted promoter fragment.

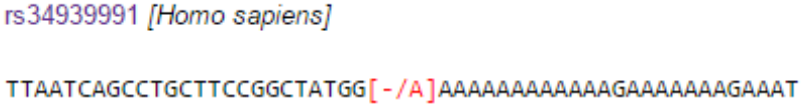


Figure 9 Single Nucleotide Polymorphism (rs34939991) located upstream the 5' of *MET*. A deletion was found in the human promoter fragment of *MET* that was incorporated in the pGL3-Basic plasmid.

Table 11. Differences found between the human reference genome (NCBI 36/ hg 18) and the inserted promoter fragment of *MET*.

Nucleotide observed in the human reference genome ((NCB1 36/hg 18)	Location at the reference human genome (NCB1 36/hg 18)	Position relative to the transcription start site (NCB1 36/hg 18)	Variation found in the inserted promoter fragment	Ancestral allele in the human lineage
Adenine	chr7: 116,098,755	-940 bp	Thymine	Thymine
Adenine	chr7: 116,098,872	-823 bp	Deletion	Adenine
Adenine	chr7: 116, 099, 166	-529 bp	Cytosine	Cytosine

Chimpanzee reference genome versus inserted chimpanzee fragment

We also compared the chimpanzee reference genome from UCSC (Feb.2011 (CSAC 2.1.4/ panTro4 Assembly)) against the inserted chimpanzee-DNA fragment. When performing the *in-silico* analysis we only found an insertion of two nucleotides in the inserted DNA fragment. It is noteworthy to mention that this insertion of two base pairs (GG) was also observed in the human reference genome (chr7: 116, 099, 665-116, 099, 666), as previously mentioned, and in its corresponding inserted fragment. Therefore, in the inserted chimpanzee fragment we found an insertion that was observed in the human reference genome (human genome 18) and also in the inserted human fragment.

Inserted human fragment versus inserted chimpanzee fragment

We also compared the human promoter fragment against the chimpanzee promoter fragment (both inserted in the pGL3-Basic Vector). When analyzing both fragments, we found 7 differences (Table 12). The first one was located at chr7: 116, 098,859 (NCB1 36/hg 18). In this position, a guanine (human genome) was observed relative to a deletion in the chimpanzee fragment. The deletion in the chimpanzee fragment was predicted by the conservation track provided by UCSC. Another difference was located at chr7:116,098,860 (NCB1 36/hg 18), and this one was previously identified as the SNP rs62469050, as

reported by the SNP database from NCBI (Figure 10). The ancestral allele for this SNP was adenine, which was observed in the inserted chimpanzee-fragment of *MET*. A third variation was found in the human lineage relative to the chimpanzee fragment. The guanine observed in the chimpanzee fragment was located at chr7: 118, 164, 338 (Feb.2011 (CSAC 2.1.4/ panTro4 Assembly)). Even though this guanine was observed in the chimpanzee reference genome (pan Tro4 Assembly) it was not predicted by the conservation track provided by UCSC. Similarly, we found an adenine located at chr7: 118,164, 520 (Feb.2011 (CSAC 2.1.4/ panTro4 Assembly)) in the chimpanzee fragment relative to the human fragment. This variation in the chimpanzee fragment was predicted by the conservation track provided by UCSC. An insertion of 1 bp was also found in the chimpanzee fragment located at chr7: 118, 164, 856 (Feb.2011 (CSAC 2.1.4/ panTro4 Assembly)) relative to a deletion in the human fragment. The deletion found in the human fragment was also confirmed by the Chain track, provided by UCSC. This track performs DNA alignments using the human genome 19, ((Feb. 2009 (GRCh37/hg19)) as a reference. Therefore, we reject the idea that this deletion was as a result of an error during the plasmid manipulation. Another difference was observed in the chimpanzee fragment at chr7: 118,164, 868 (Feb.2011 (CSAC 2.1.4/ panTro4 Assembly)) relative to an adenine in the human fragment. In this position a guanine was observed in the chimpanzee, and conversely an adenine was observed in the human fragment. The difference observed in the chimpanzee lineage was also predicted by the conservation track from UCSC. Finally, the last difference was a guanine in the chimpanzee fragment located at chr7: 118, 164,901(Feb.2011 (CSAC 2.1.4/ panTro4 Assembly)), relative to an adenine in the human fragment. The guanine observed in the chimpanzee lineage was also predicted by the conservation track from UCSC.

In summary, we found that in addition to the indel of 23 bp, the human and chimpanzee fragments both have 7 differences located near the transcription start of *MET*, that could possibly cause a differential expression in human and chimpanzee lineages (Figure 11).

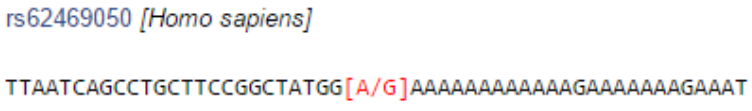


Figure 10 Single Nucleotide Polymorphism (rs62469050) located upstream the 5' of *MET*. The ancestral allele was observed in the chimpanzee fragment, while the human fragment presented guanine at chr7: 116, 098, 860 (NCBI 36/ hg 18).

Table 12 Differences found between the inserted human fragment and the inserted chimpanzee fragment of *MET*.

Nucleotide observed in the human promoter fragment of *MET*	**Location at the reference human genome (NCB1 36/hg 18)**	**Position relative to the transcription start site (NCB1 36/hg 18)**	**Nucleotide found in the chimpanzee promoter fragment of *MET***	**Variation found in the chimpanzee lineage**	**Variation found in the human lineage**
Guanine	Chr7: 116, 098,859	-836 bp	Deletion	Ancestral	Fixed
Guanine	Chr7: 116, 098,860	-835 bp	Adenine	Ancestral	SNP
Thymine	Chr7: 116, 098, 944	-751 bp	Guanine	Derived	Fixed
Guanine	Chr7: 116, 099, 126	-569 bp	Adenine	Ancestral	Fixed
Deletion	Chr7: 116,098, 461 +1	-233 bp	Cytosine	Ancestral	Fixed
Adenine	Chr7: 116, 099, 473	-222 bp	Guanine	Ancestral	Fixed
Adenine	Chr7: 116,099, 506	-189 bp	Guanine	Ancestral	Fixed

Figure 11 Differences located near the transcription start site of *MET*. Summarized representation of the differences observed between the human inserted fragment and the chimpanzee inserted fragment. The blue star represents the SNP (rs62469050) that was found in the promoter region. All the positions are relative to the transcription start site of *MET*.

Conclusion

We created 4 different recombinants plasmids; one with the wild type promoter fragment of *MET* from human (chr7: 116,098,490 to 116,099,792) (Mar.2006 (NCB1 36/hg 18 assembly)) and the other with its orthologous sequence from chimpanzee (chr 7: 118,163,863 to 118,165,185 (Feb. 2011 (CSAC 2.1.4/panTro4))). Also, we created a recombinant plasmid with the promoter fragment of chimpanzee and the corresponding deletion of 23 bp from human and another one with the promoter fragment of human and corresponding insertion of 23 bp from chimpanzee. Therefore, we created 'hybrid' recombinant plasmids with an indel of one species but the sequence of the other.

When performing the *in-silico* analysis (DNA-alignments), we found that besides the indel of 23 bp, both inserted promoter fragments differ by 7 positions (Table 12) (Figure 11). However, when we compared the inserted human fragment with the human reference genome, we only found 3 nucleotide variations. All of the nucleotide variations that were present in the inserted human fragment were already reported as SNPs.

Similarly, when we compared the inserted chimpanzee fragment with the chimpanzee reference genome we only found one difference; an insertion of two consecutive nucleotides.

In conclusion, given that the human and chimpanzee promoter fragments, both have 7 differences in addition to the indel of 23 bp, *MET* could possibly have a differential gene expression because these differences are located nearby the transcription start site. Therefore, in order to gain more insights regarding the implications of the transcriptional activity of *MET*, a luciferase reporter assay could be used to bring new light about this subject.

References

1. Al-Mahruqi, S. H., Zadjali, F., Beja-Pereira, A., Koh, C. Y., Balkhair, A., & Al-Jabri, A. A. (2014). Genetic diversity and prevalence of *CCR2-CCR5* gene polymorphisms in the Omani population. *Genetics and Molecular Biology*, *37*(1), 7–14.
2. Bauernfeind, A. L., Soderblom, E. J., Turner, M. E., Moseley, M. A., Ely, J. J., Hof, P. R., … Babbitt, C. C. (2015). Evolutionary Divergence of Gene and Protein Expression in the Brains of Humans and Chimpanzees. *Genome Biology and Evolution*, *7*(8), 2276–2288. http://136.145.90.90:4562/10.1093/gbe/evv132
3. Burdick, K. E., DeRosse, P., Kane, J. M., Lencz, T., & Malhotra, A. K. (2010). Genetic Variation in the *MET* Proto-oncogene is Associated with Schizophrenia and General Cognitive Ability. *The American Journal of Psychiatry*, *167*(4), 436–443. http://doi.org/10.1176/appi.ajp.2009.09050615
4. Campbell, D. B., Sutcliffe, J. S., Ebert, P. J., Militerni, R., Bravaccio, C., Trillo, S., … Levitt, P. (2006). A genetic variant that disrupts *MET* transcription is associated with autism. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(45), 16834–16839. http://doi.org/10.1073/pnas.0605296103
5. Carroll, S.B. (2003). Genetics and the making of *Homo sapiens*. *Nature, 422*(6934) 849-57.
6. Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature, 437*(7055), 69-87.
7. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*, 30, 2478-83.
8. Green R. E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S, Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., & Pääbo, S. (2010). A draft sequence of Neandertal genome. *Science, 328*(5979), 710-22. http://doi: 10.1126/science.1188021.

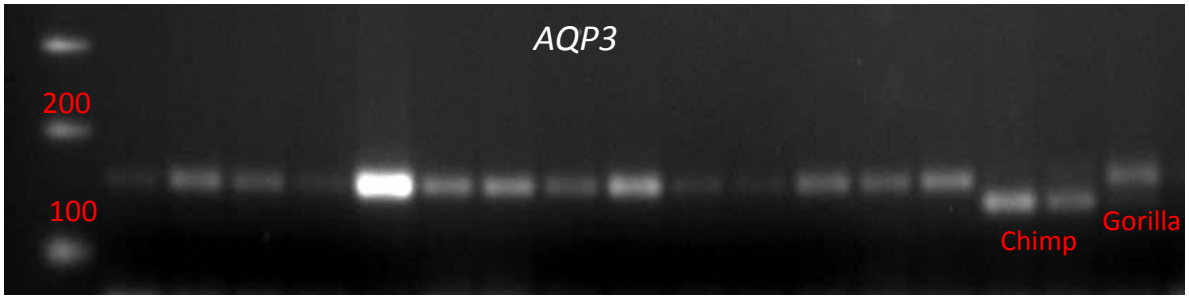
9. Hedrick, A., Lee, Y., Wallace, G. L., Greenstein, D., Clasen, L., Giedd, J. N., & Raznahan, A. (2012). Autism Risk Gene *MET* Variation and Cortical Thickness in Typically Developing Children and Adolescents. *Autism Research : Official Journal of the International Society for Autism Research*, *5*(6), 434–439. http://doi.org/10.1002/aur.1256
10. Judson, M. C., Eagleson, K. L., & Levitt, P. (2011). A new synaptic player leading to autism risk: Met receptor tyrosine kinase. *Journal of Neurodevelopmental Disorders*, *3*(3), 282–292. http://doi.org/10.1007/s11689-011-9081-8
11. Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., … Pääbo, S. (2004). Regional Patterns of Gene Expression in Human and Chimpanzee Brains. *Genome Research*, *14*(8), 1462–1473. http://136.145.90.90:4562/10.1101/gr.2538704
12. King, MC. & Wilson, AC. (1975). Evolution at two levels in humans and chimpanzees. *Science, 188*(4184), 107-116.
13. Mi, H., Muruganujan, A., Casagrande, J.T., & Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols, 8*, 1551-1566. *doi: 10.1038/nprot.2013.092.*
14. Mukamel, Z., Konopka, G., Wexler, E., Osborn, G. E., Dong, H., Bergman, M. Y., … Geschwind, D. H. (2011). Regulation of MET by FOXP2, Genes Implicated in Higher Cognitive Dysfunction and Autism Risk. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(32), 11437–11442. http://doi.org/10.1523/JNEUROSCI.0181-11.2011
15. Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*(R2), R131–R136. http://doi.org/10.1093/hmg/ddq400
16. NCBI. (2015, November). *PRSS12 protease, serine 12 [Homo sapiens (human)]*. Retrieved from http://www.ncbi.nlm.nih.gov/gene/8492
17. NCBI. (2016, February). *SNORA75 small nucleolar RNA, H/ACA box 75 [Homo sapiens (human)]*. Retrieved from http://www.ncbi.nlm.nih.gov/gene/654321
18. NCBI. (2016, January). *DICER1 dicer 1 ribonuclease III [Homo sapiens (human)]*. Retrieved from *Retrieved* from http://www.ncbi.nlm.nih.gov/gene/23405

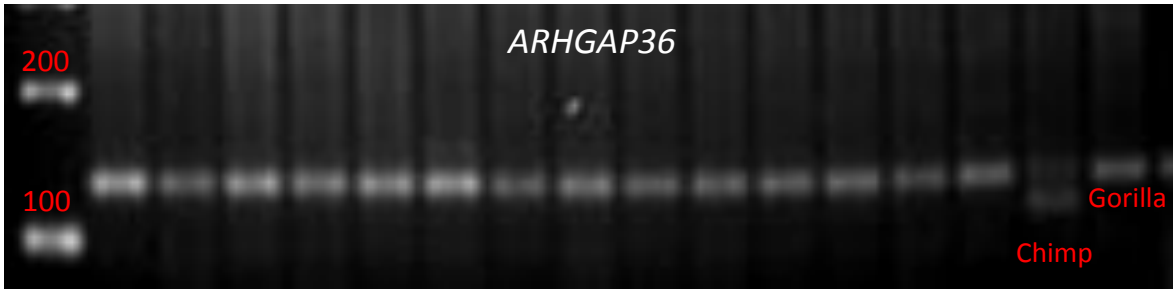
19. NCBI. (2016, January). *MB myoglobin [Homo sapiens (human)]*. Retrieved from http://www.ncbi.nlm.nih.gov/gene/4151
20. NCBI. (2016, January). *MET proto-oncogene, receptor tyrosine kinase [Homo sapiens (human)]*. Retrieved from http://www.ncbi.nlm.nih.gov/gene/4233
21. Noonan, J. P. (2010). Neanderthal genomics and the evolution of modern humans. *Genome Research*, *20*(5), 547–553. http://doi.org/10.1101/gr.076000.108
22. Paixao-Cortes, V., Henriques, L., Mauro, F., Catira, M. & Hunemeier, T. (2013). The Cognitive Ability of Extinct Hominis: Bringing Down the Hierarchy Using Genomic Evidences. *American Journal of Human Biology*, *25*, 702-705.
23. Pattabiraman, S., Baumann, C., Guisado, D., Eppig, J. J., Schimenti, J. C., & De La Fuente, R. (2015). Mouse BRWD1 is critical for spermatid postmeiotic transcription and female meiotic chromosome stability. *The Journal of Cell Biology*, *208*(1), 53–69. http://doi.org/10.1083/jcb.201404109
24. Patterson, N., Ritcher, D., Gnerre, S., Lander, E. & Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature,* 441(29), 1103-1108.
25. Peng, Z., Zhu, Y., Wang, Q., Gao, J., Li, Y., Li, Y., … Shen, L. (2014). Prognostic Significance of MET Amplification and Expression in Gastric Cancer: A Systematic Review with Meta-Analysis. *PLoS ONE*, *9*(1), e84502. http://doi.org/10.1371/journal.pone.0084502
26. Philipps, D. L., Wigglesworth, K., Hartford, S. A., Sun, F., Pattabiraman, S., Schimenti, K., … Schimenti, J. C. (2008). The dual bromodomain and WD repeat-containing mouse protein BRWD1 is required for normal spermiogenesis and the oocyte-embryo transition. *Developmental Biology*, *317*(1), 72–82. http://doi.org/10.1016/j.ydbio.2008.02.018
27. Plummer, J. T., Evgrafov, O. V., Bergman, M. Y., Friez, M., Haiman, C. A., Levitt, P., & Aldinger, K. A. (2013). Transcriptional regulation of the *MET* receptor tyrosine kinase gene by MeCP2 and sex-specific expression in autism and Rett syndrome. *Translational Psychiatry*, *3*(10), e316–. http://doi.org/10.1038/tp.2013.91
28. Polavarapu, N., Arora, G., Mittal, V. K., & McDonald, J. F. (2011). Characterization and potential functional significance of human-chimpanzee large INDEL variation. *Mobile DNA*, *2*, 13. http://doi.org/10.1186/1759-8753-2-13

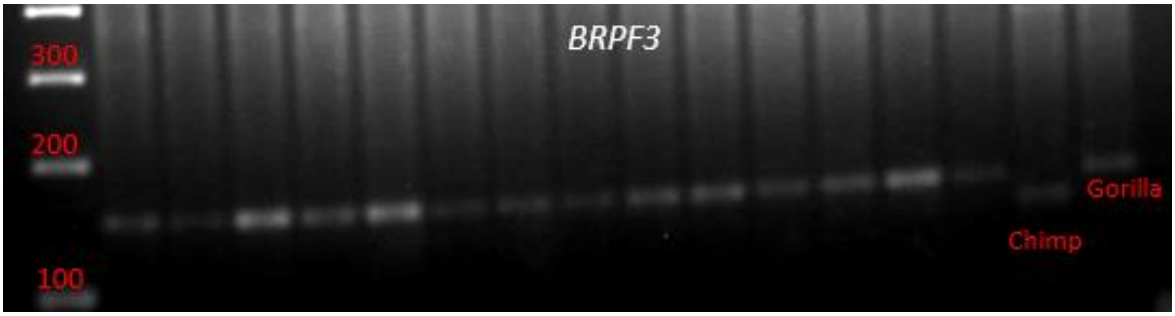
29. Poulain, M., Frydman, N., Tourpin, S., Muczynski, V., Souguet, B., Benachi, A., Habert, R., Rouiller-Fabre, V. & Livera, G. (2014). Involvement of doublesex and mab-3-related transcription factors in human female germ cell development demonstrated by xenograft and interference RNA strategies. *Molecular Human Reproduction*, *20*(10), 960-971. doi: 10.1093/molehr/gau058.
30. Rudie, J. D., Hernandez, L. M., Brown, J. A., Beck-Pancer, D., Colich, N. L., Gorrindo, P., … Dapretto, M. (2012). Autism-Associated Promoter Variant in *MET* Impacts Functional and Structural Brain Networks. *Neuron*, *75*(5), 904–915. http://doi.org/10.1016/j.neuron.2012.07.010
31. Ruiz, J., Blanché, H., Cohen, N., Velho, G., Cambien, F., Cohen, D., … Froguel, P. (1994). Insertion/deletion polymorphism of the angiotensin-converting enzyme gene is strongly associated with coronary heart disease in non-insulin-dependent diabetes mellitus. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(9), 3662–3665.
32. Sjödin, P., Bataillon, T., & Schierup, M. H. (2010). Insertion and Deletion Processes in Recent Human History. *PLoS ONE*, *5*(1), e8650.http://doi.org/10.1371/journal.pone.0008650
33. Sousa, I., Clark, T. G., Toma, C., Kobayashi, K., Choma, M., Holt, R., … International Molecular Genetic Study of Autism Consortium (IMGSAC). (2009). *MET* and autism susceptibility: family and case–control studies.*European Journal of Human Genetics*, *17*(6), 749–758. http://doi.org/10.1038/ejhg.2008.215
34. Sun, T., Gao, Y., Tan, W., Ma, S., Shi, Y., Yao, J., Guo, Y., Yang, M., Zhang, X., Zhang, Q., Zeng, C. & Ling, D. (2007). A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter is associated with susceptibility to multiple cancers. *Nature Genetics*, *39*(5), 605-613. doi:10.1038/ng2030
35. The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. http://doi.org/10.1038/nature11632
36. Trzyna, E., Duleba, M., Faryna, M. & Majka, M. (2009). Abstract A209: Comprehensive analysis of the MET protooncogene promoter regulation. *Molecular Targets and Cancer Therapeutics*, AACR-NCI-EORTC International Conference.
37. Volfovsky, N., Oleksyk, T. K., Cruz, K. C., Truelove, A. L., Stephens, R. M., & Smith, M. W. (2009). Genome and gene alterations by insertions and deletions in the evolution of human and chimpanzee chromosome 22. *BMC Genomics*, *10*, 51. http://doi.org/10.1186/1471-2164-10-51

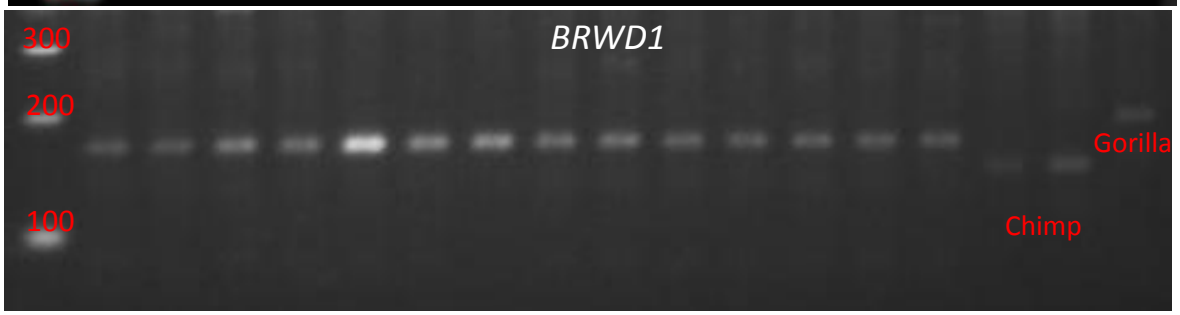
38. Wetternbom, A., Sevov, M., Cavelier, L. & Bergstrom, T. F. (2006). Comparative Genomic Analysis of Human and Chimpanzee Indicates a Key Role for Indels in primate evolution. *Journal of Molecular Evolution,* 63, 682-690.
39. Wray, G.A. Hahn, M. W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M. V. & Romano, L. (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution, 20*(9) 1377-1419. http://doi.org/10.1093/molbev/msg140.

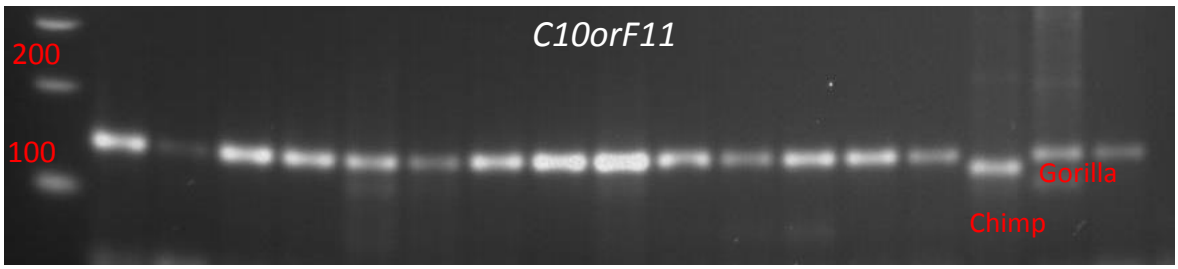
Appendix

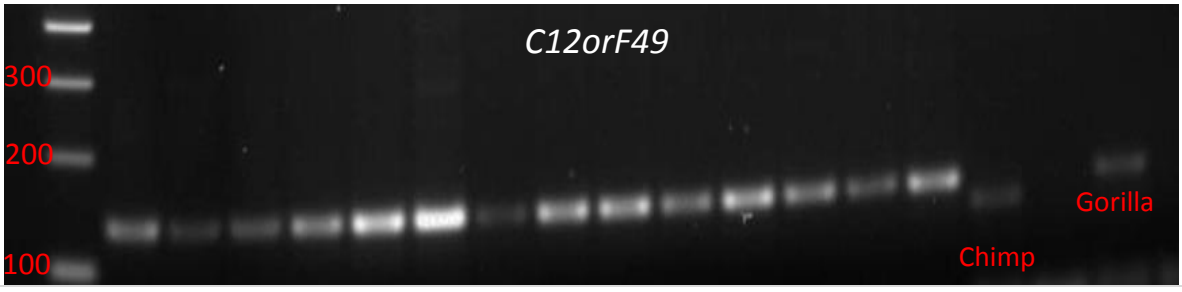

AQP3
200
100
Chimp
Gorilla

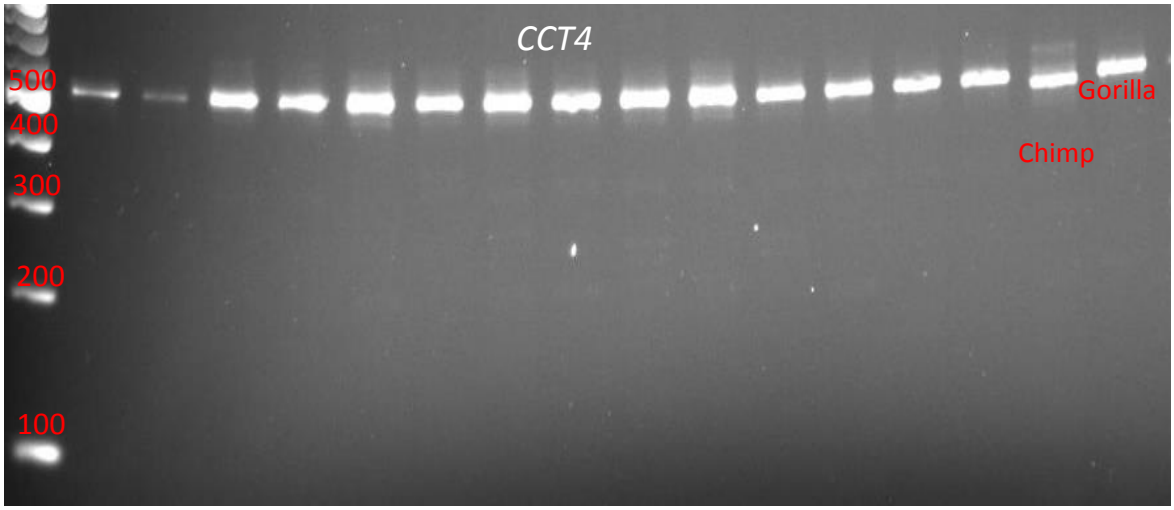

ARHGAP36
200
100
Gorilla
Chimp

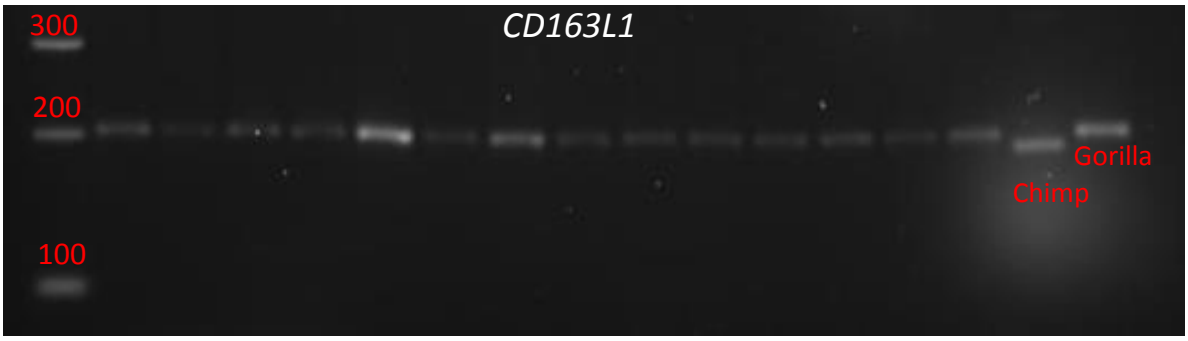

BRPF3
300
200
100
Gorilla
Chimp

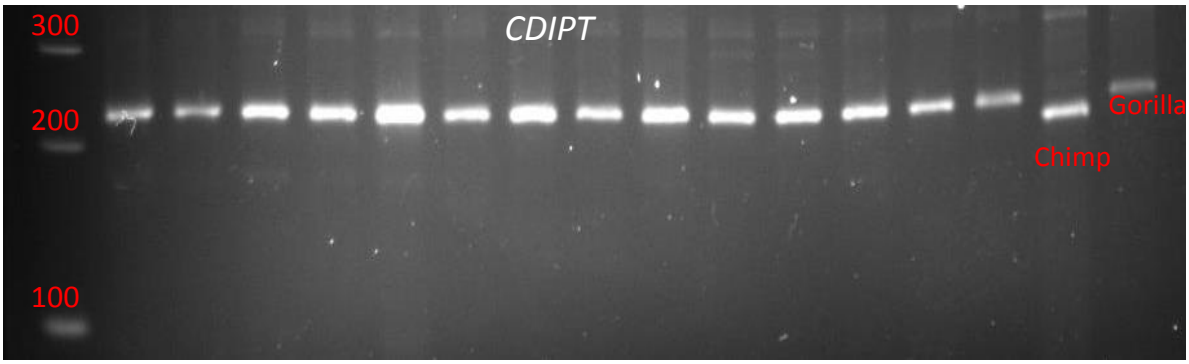

BRWD1
300
200
100
Gorilla
Chimp

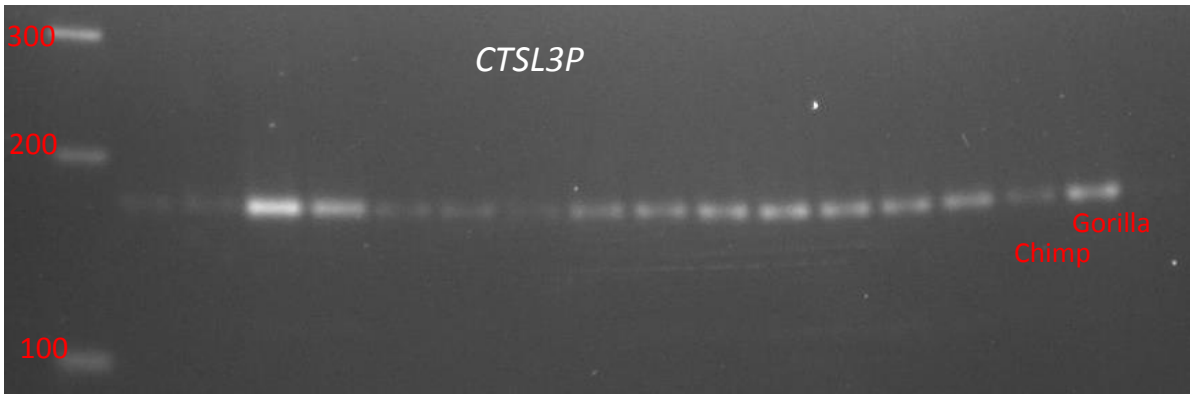

C10orF11
200
100
Gorilla
Chimp

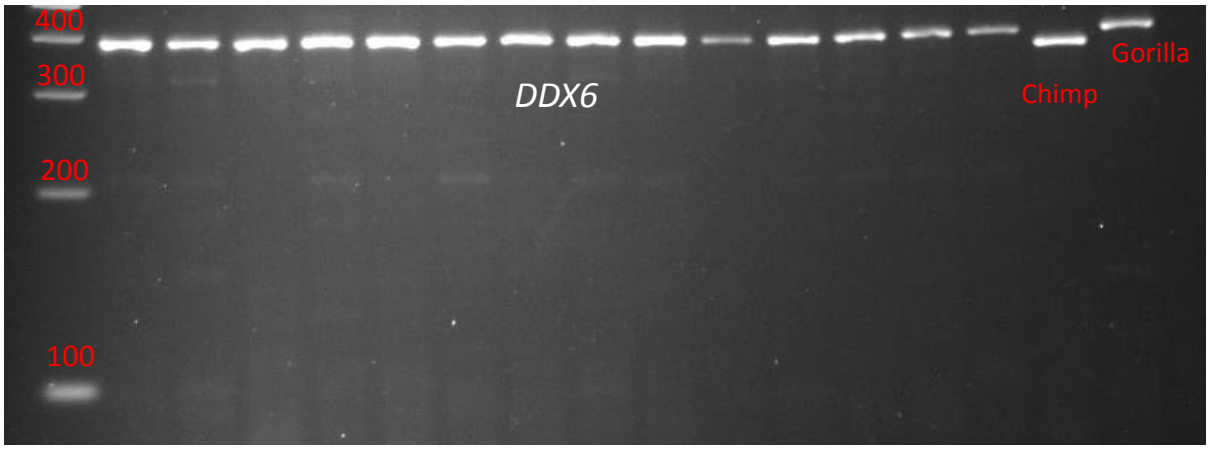

C12orF49
300
200
100
Gorilla
Chimp

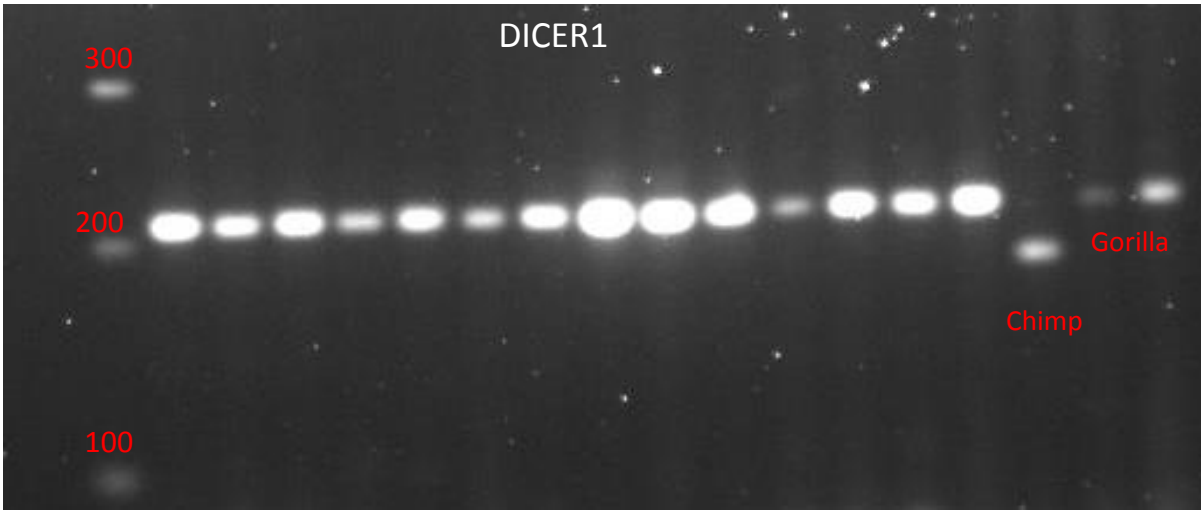

CCT4
500
400
300
200
100
Gorilla
Chimp

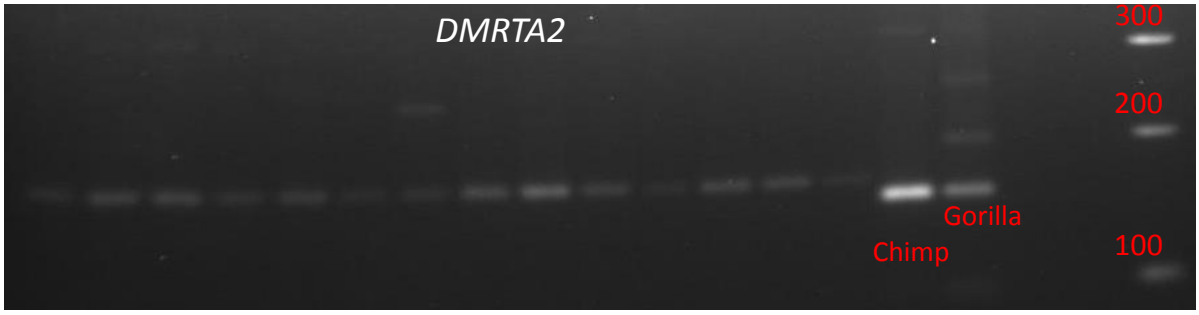

CD163L1
300
200
100
Gorilla
Chimp


CDIPT
300
200
100
Gorilla
Chimp

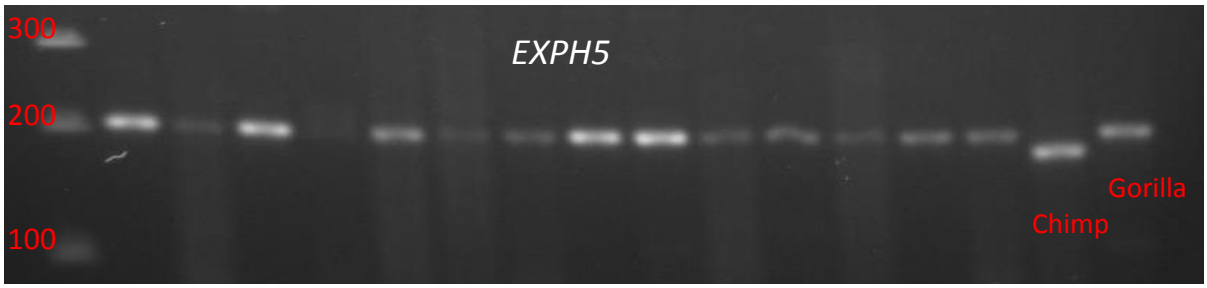

CTSL3P
300
200
100
Gorilla
Chimp

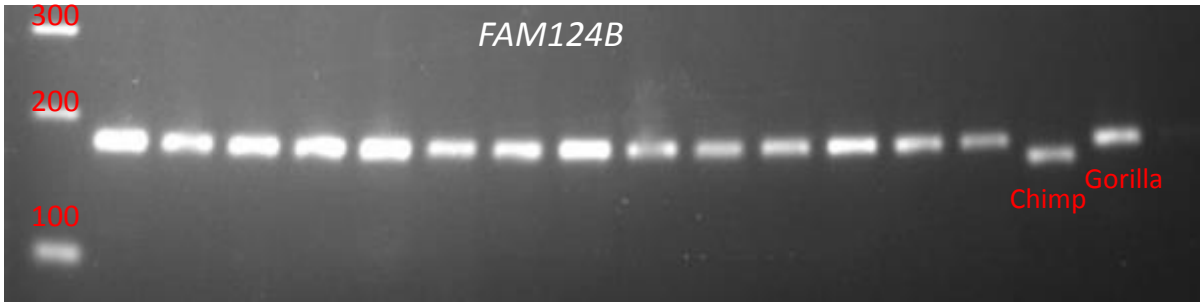

400
300
200
100
DDX6
Chimp
Gorilla

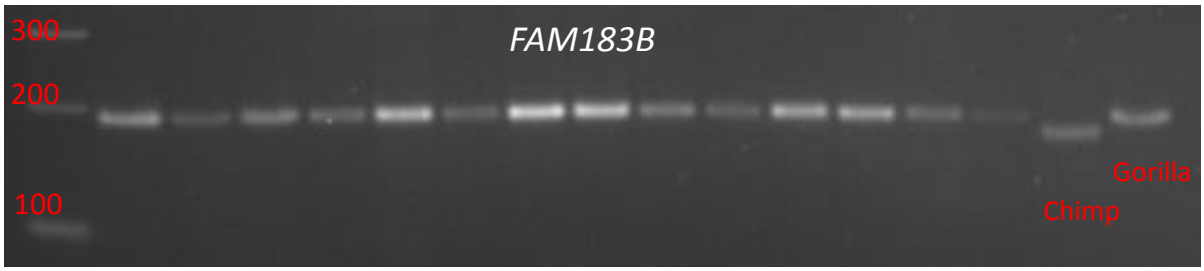

DICER1
300
200
100
Gorilla
Chimp

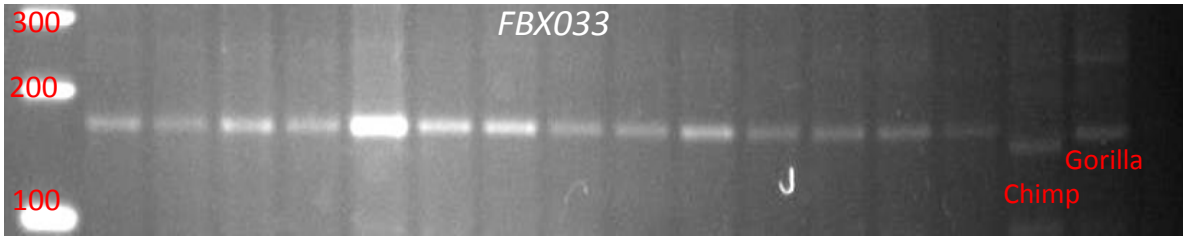

DMRTA2
300
200
100
Gorilla
Chimp

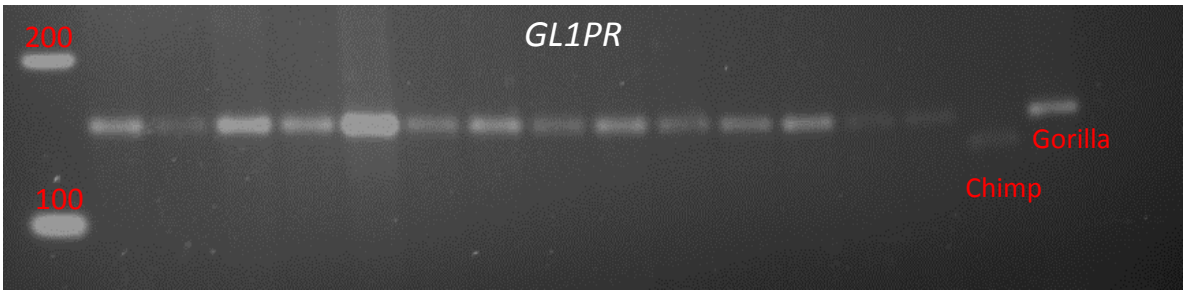

200
100
EIF6
Gorilla
Chimp

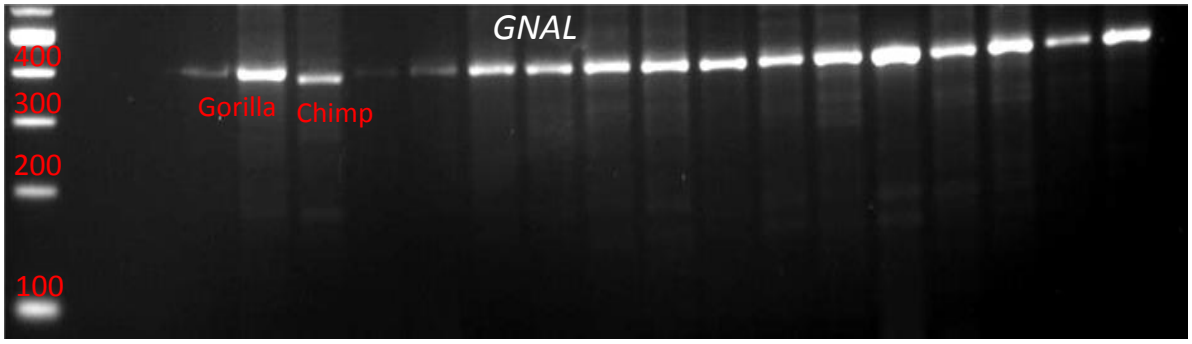

300
200
100
EXPH5
Gorilla
Chimp

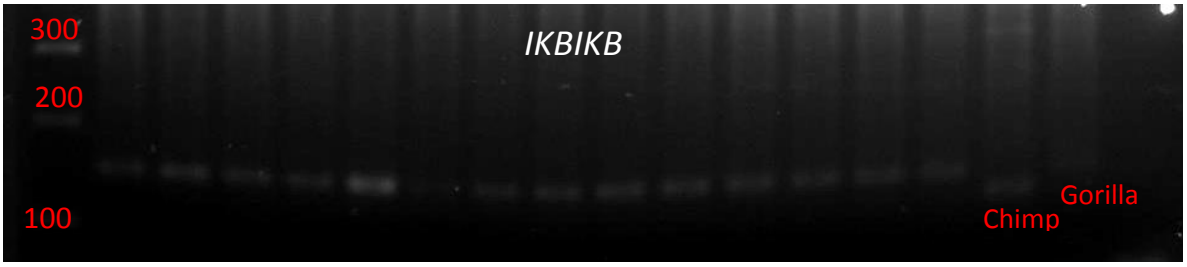

300
200
100
FAM124B
Chimp
Gorilla

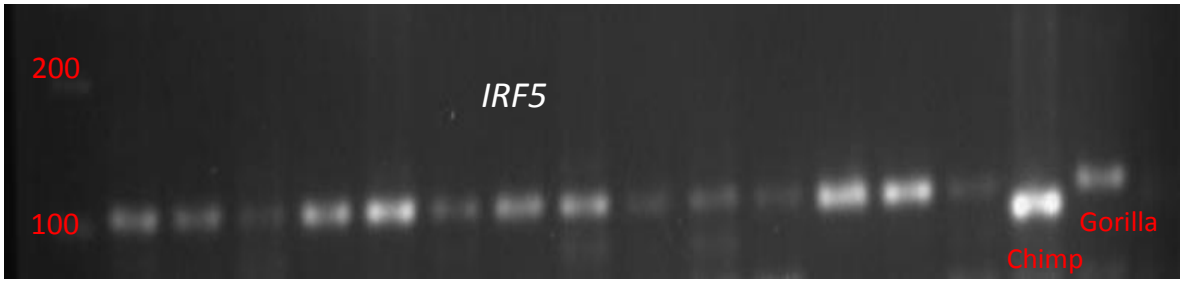

300
200
100
FAM183B
Gorilla
Chimp

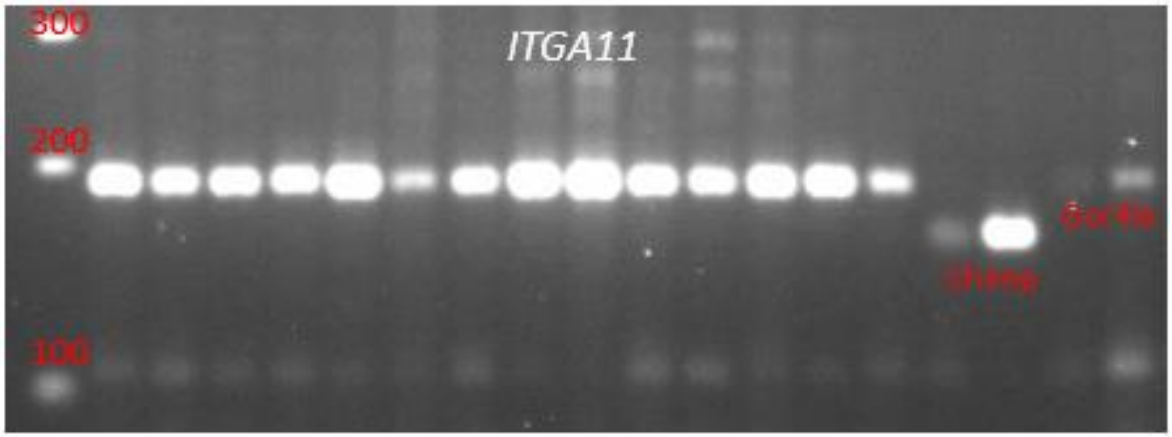

300
200
100
FBX033
Gorilla
Chimp

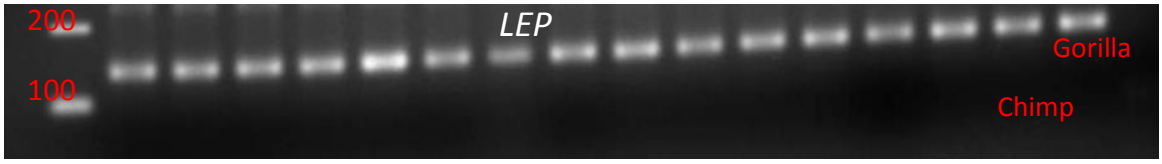

200
100
GL1PR
Gorilla
Chimp

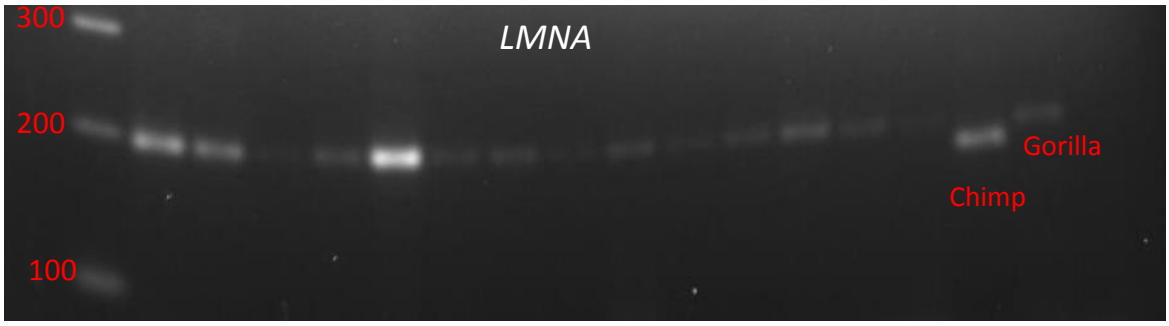

GNAL
400
300
200
100
Gorilla
Chimp

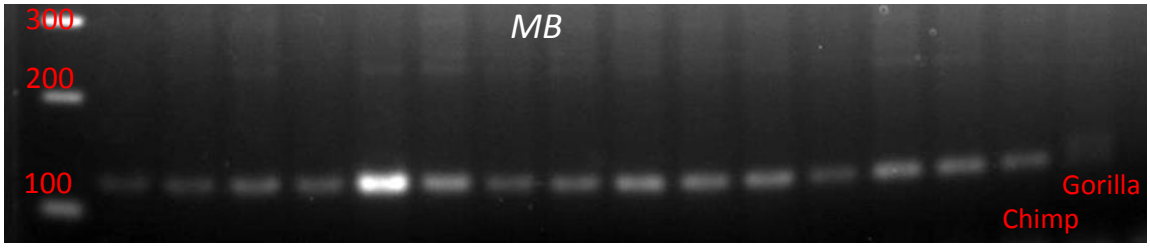

IKBIKB
300
200
100
Chimp
Gorilla

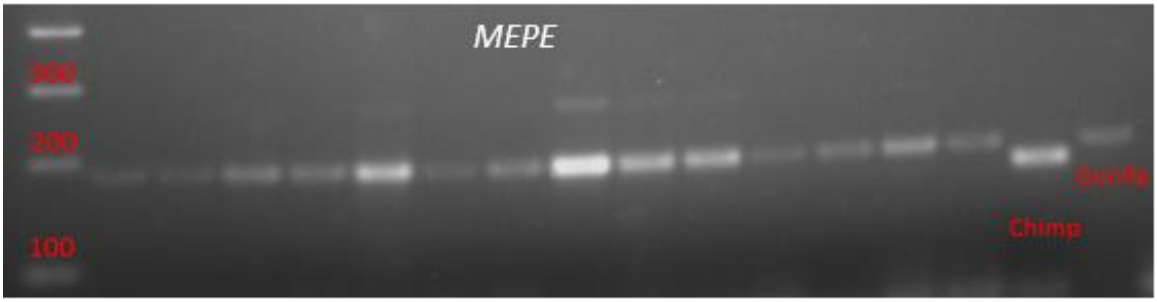

IRF5
200
100
Gorilla
Chimp

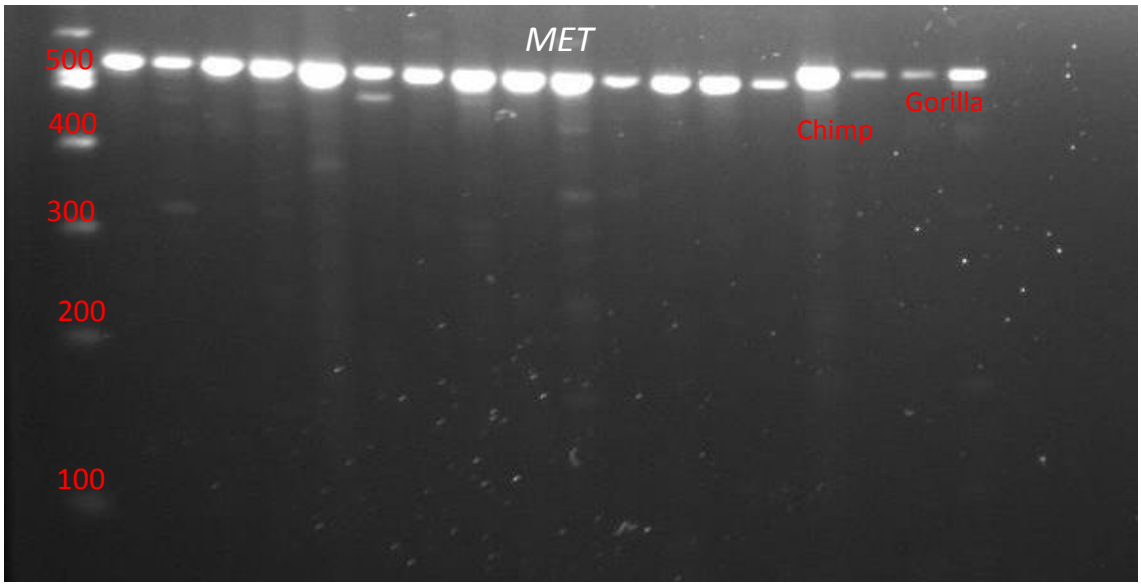

ITGA11
300
200
100

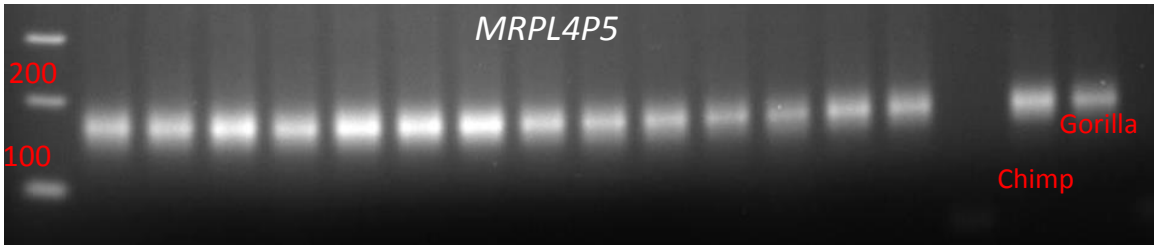

LEP
200
100
Gorilla
Chimp

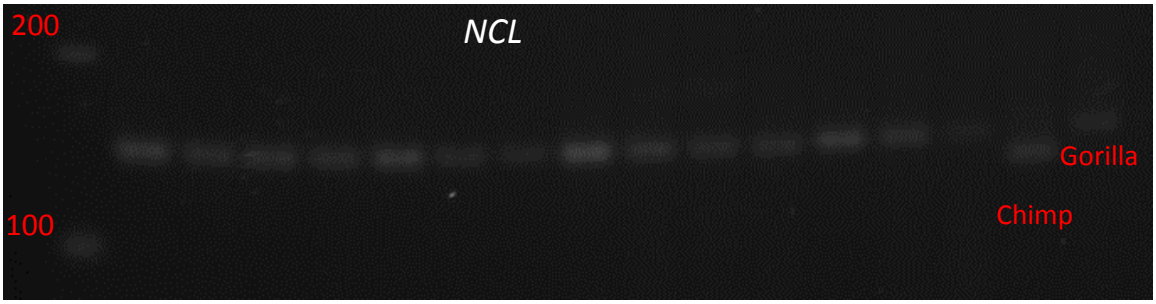

LMNA
300
200
100
Gorilla
Chimp

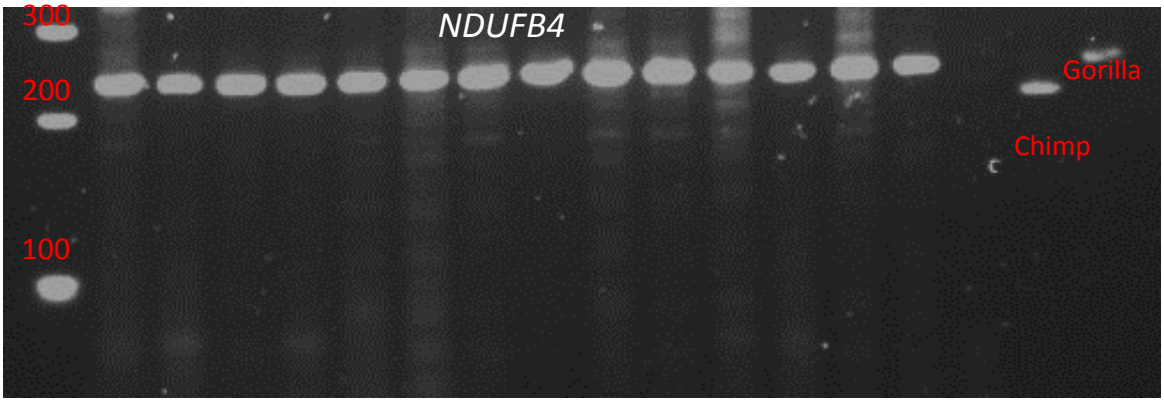

MB
300
200
100
Gorilla
Chimp

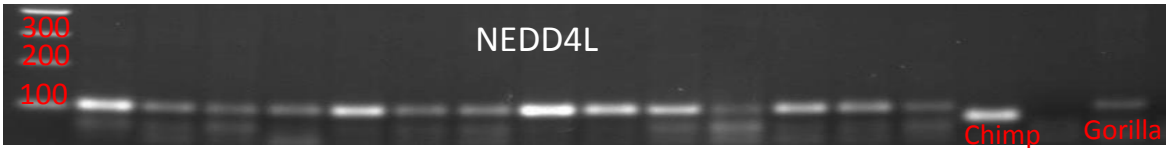

MEPE
200
100
Chimp

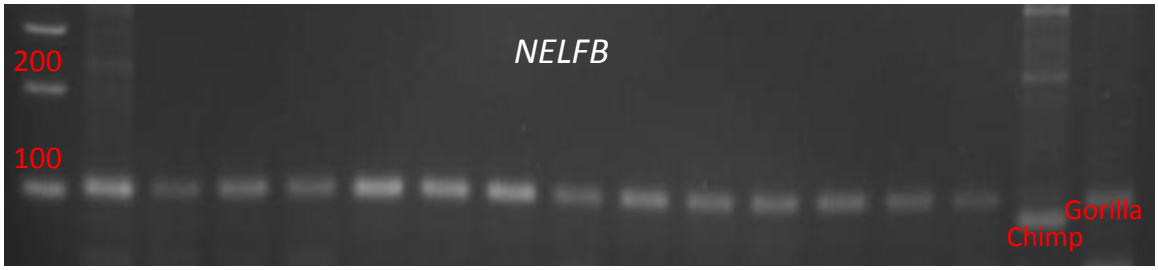

MET
500
400
300
200
100
Chimp
Gorilla

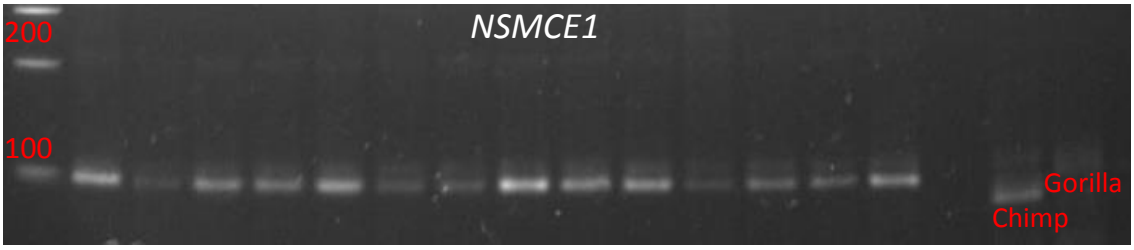

MRPL4P5
200
100
Gorilla
Chimp

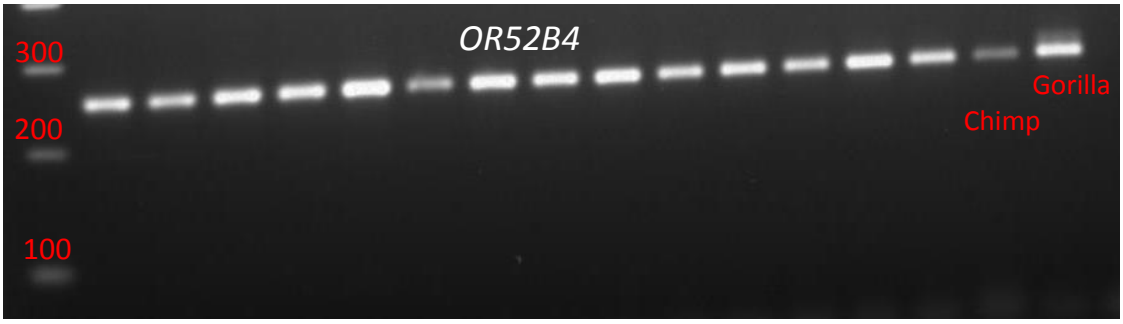

200
NCL
100
Gorilla
Chimp

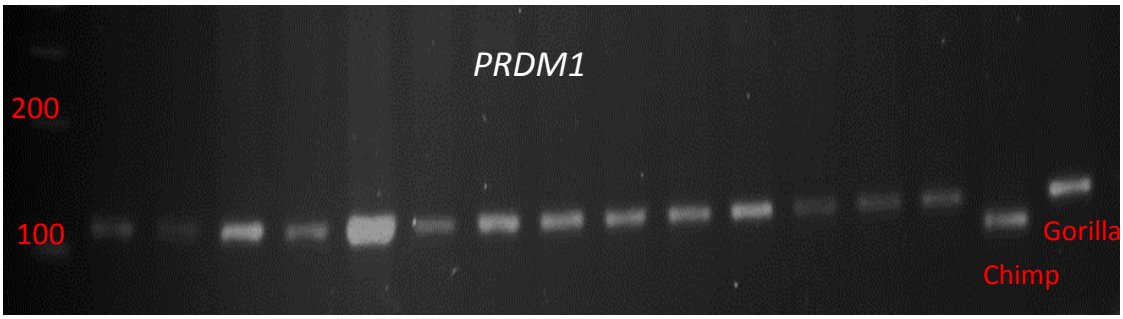

300
200
NDUFB4
100
Gorilla
Chimp

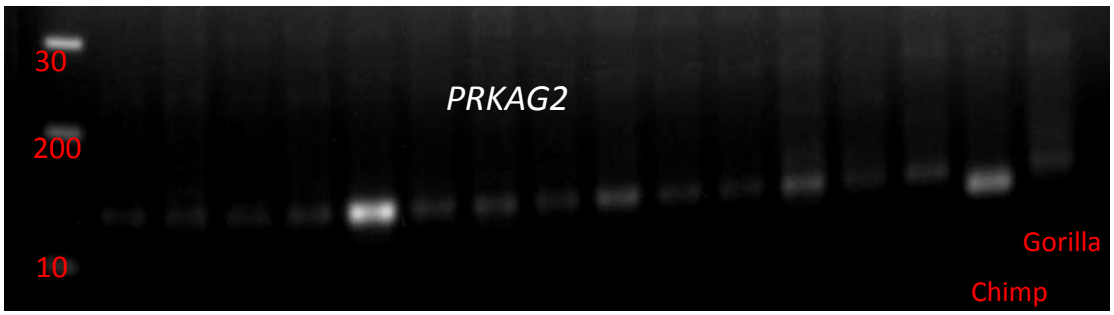

300
200
NEDD4L
100
Chimp
Gorilla

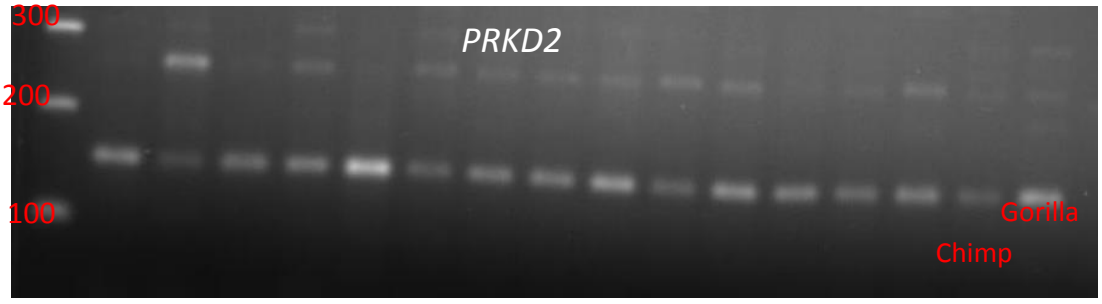

200
NELFB
100
Gorilla
Chimp

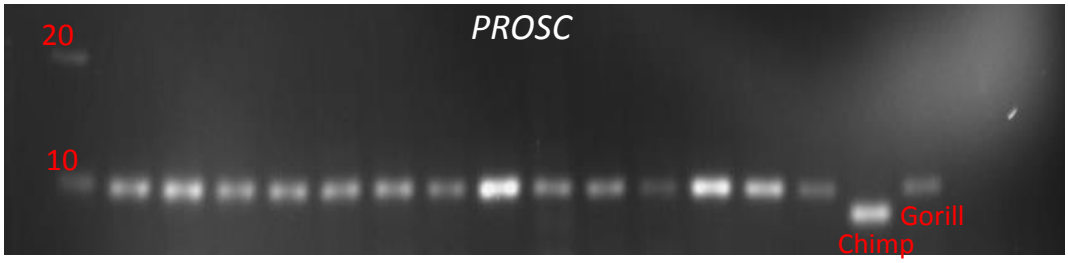

NSMCE1
200
100
Gorilla
Chimp


OR52B4
300
200
100
Gorilla
Chimp

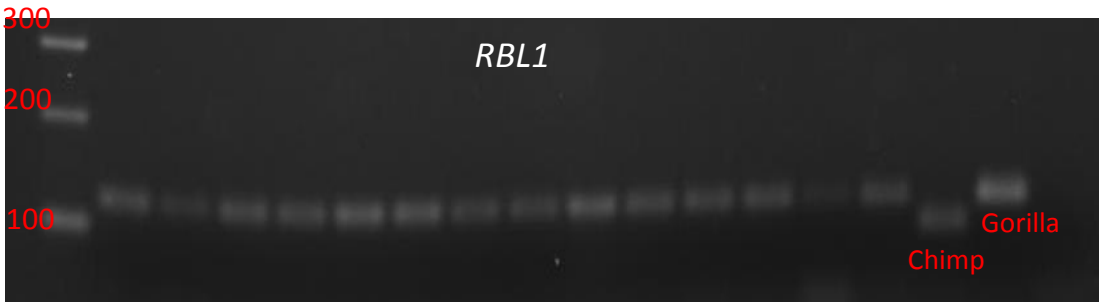

PRDM1
200
100
Gorilla
Chimp

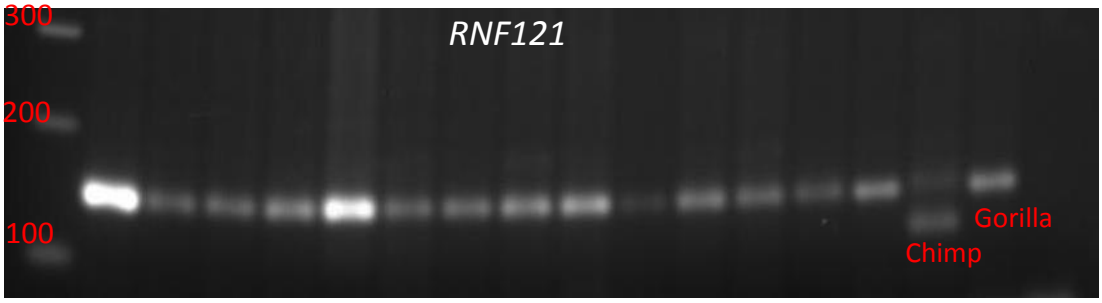

PRKAG2
30
200
10
Gorilla
Chimp

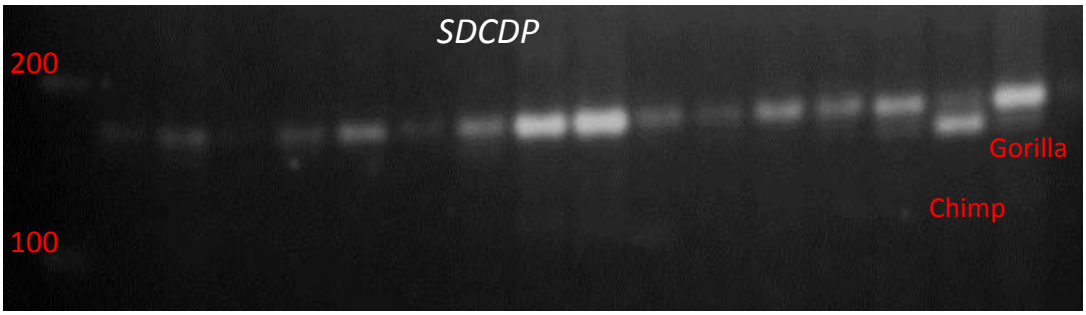

PRKD2
300
200
100
Gorilla
Chimp

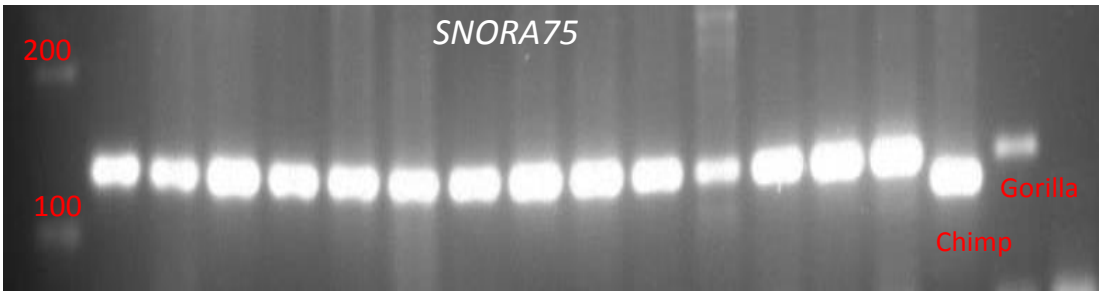

PROSC
20
10
Gorill
Chimp

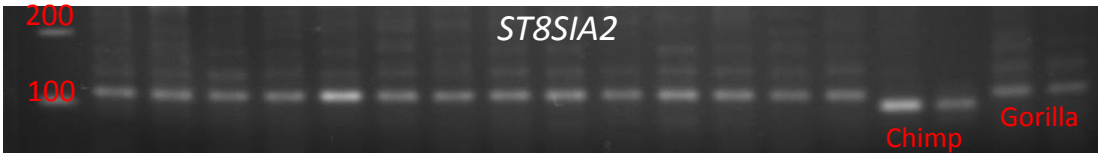

300
200
100
PRSS12
Gorill
Chimp

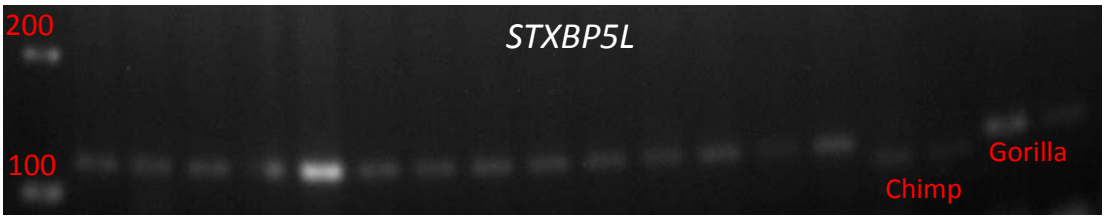

300
200
100
RBL1
Gorilla
Chimp

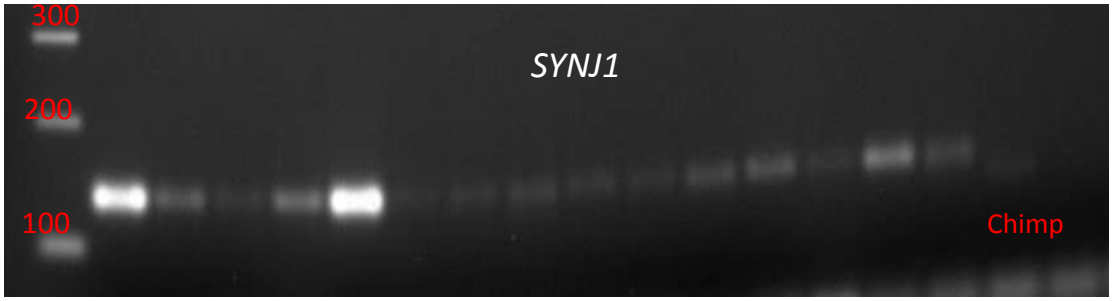

300
200
100
RNF121
Gorilla
Chimp

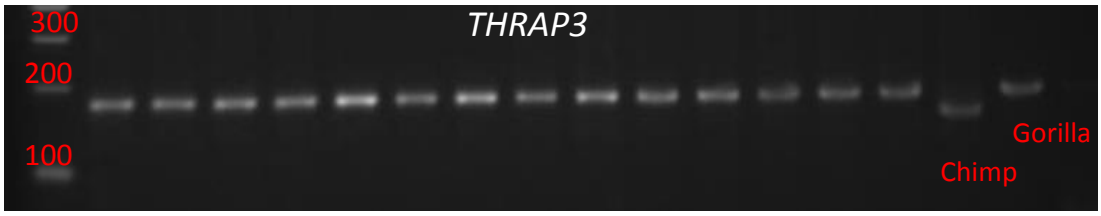

SDCDP
200
100
Gorilla
Chimp

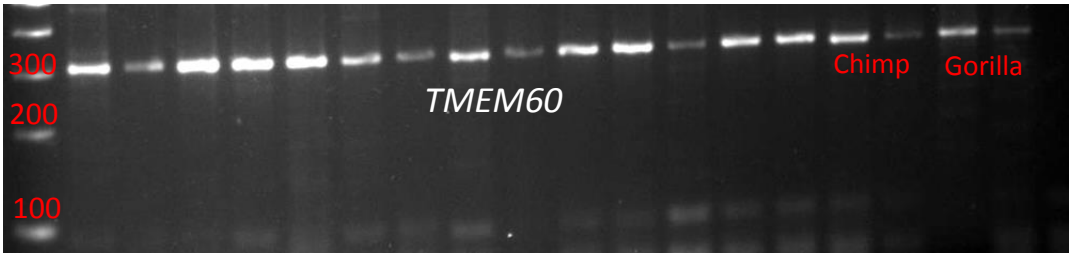

SNORA75
200
100
Gorilla
Chimp

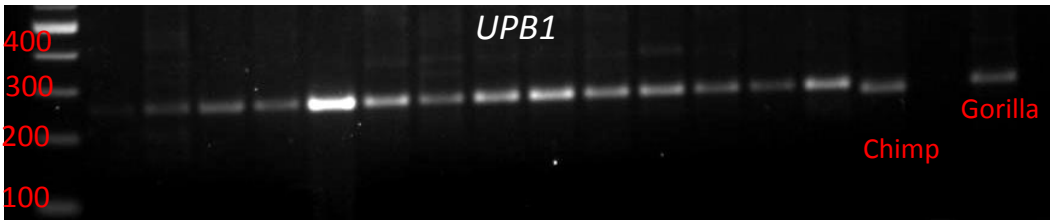

ST8SIA2
200
100
Chimp
Gorilla

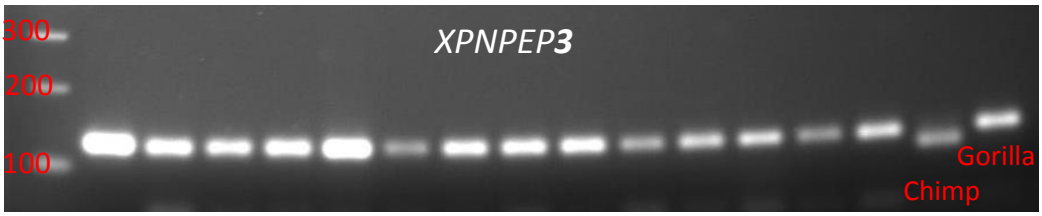

STXBP5L
200
100
Gorilla
Chimp

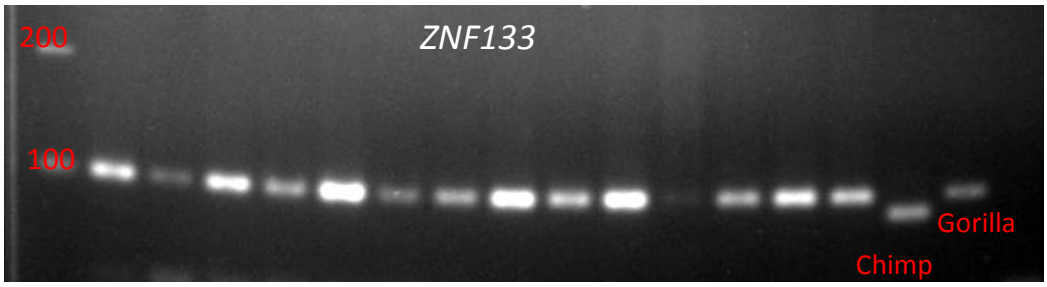

SYNJ1
300
200
100
Chimp

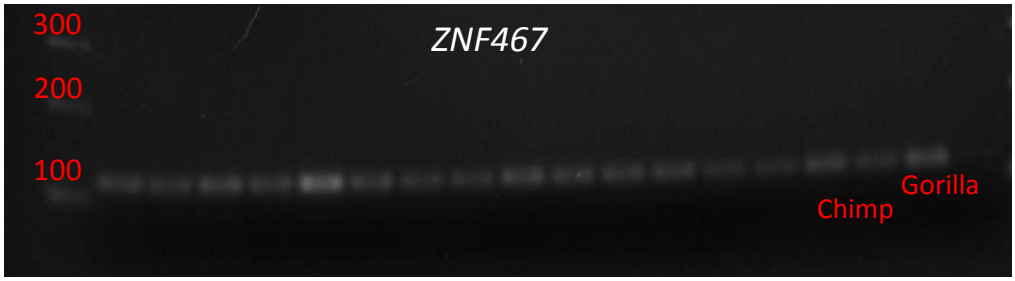

THRAP3
300
200
100
Gorilla
Chimp


300
200
100
TMEM60
Chimp
Gorilla

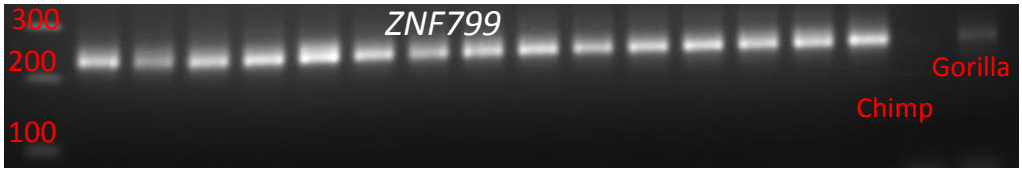

400
300
200
100
UPB1
Gorilla
Chimp


300
200
100
XPNPEP3
Gorilla
Chimp


200
100
ZNF133
Gorilla
Chimp


300
200
100
ZNF467
Chimp
Gorilla


200
100
ZNF623
Gorilla
Chimp


300
200
100
ZNF799
Gorilla
Chimp