TESTIS TRANSCRIPTOME EVOLUTION AMONG HOMINIDS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

EKİN SAĞLICAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOLOGY

JANUARY 2018

Aproval of the thesis:

TESTIS TRANSCRIPTOME EVOLUTION AMONG HOMINIDS

submitted by **EKIN SAĞLICAN** in partial fulfilment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Orhan Adalı
Head of Department, **Biology** _____

Assoc. Prof. Dr. Mehmet Somel
Supervisor, **Biology Dept., METU** _____

Examining Committee Members:

Prof. Dr. Can Bilgin
Biology Dept., METU _____

Assoc. Prof. Dr. Mehmet Somel
Biology Dept., METU _____

Prof. Dr. Tolga Can
Comp. Eng. Dept., METU _____

Prof. Dr. Mesut Muyan
Biology Dept., METU _____

Assist. Prof. Dr. İdil Yet
Bioinformatics. Dept., Hacettepe Uni. _____

Date: 04.01.2018

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : EKIN SAGLICAN

Signature :

ABSTRACT

TESTIS TRANSCRIPTOME EVOLUTION AMONG HOMINIDS

Sağlıcan, Ekin
M.S., Department of Biology
Supervisor : Assoc. Prof. Dr. Mehmet Somel

January 2018, 73 pages

The difference in the relative testis size between humans and their closest extant relatives is remarkable. Relative testis size of humans is more similar to that of gorillas than that of chimpanzees, although chimpanzees are phylogenetically closer relatives of humans. The relative testis size of chimpanzees is larger than those of both humans and gorillas; moreover, it is more similar to that of a more distant relative: the macaque. These differences in testis sizes are thought to be related with the mating behaviour of these species and to have evolved convergently. Specifically, species with single-male mating, humans and gorillas, have relatively small testes, and species with multi-male mating, chimpanzees and macaques, have large testes.

This thesis includes a total of 8 RNA-seq and microarray datasets containing testis transcriptome data of 10 different species; namely, human, chimpanzee, gorilla, macaque, marmoset, mouse of two different species, rat, platypus and opossum. I conduct comparative meta-analyses using these datasets. First, I show that genes showing differential expression in testis between humans and chimpanzees have different levels of correlation with the testis transcriptomes of gorilla and macaques. As in the relationship between testis sizes among these species, this analysis reveals signs of convergent evolution of whole testis gene expression, with higher transcriptome similarity between humans and gorillas, and higher similarity between chimpanzees and macaques.

One possible reason that can explain the divergence in testis transcriptome pro-

files of these species is the relative contribution of cell types present in testis. In the second part of the study, I used genes expressed in isolated cell types of mouse testis to detect the relative contribution of cell types found in the testes of the species used in the analysis. The results of this analysis is consistent with the previous findings: The testis transcriptome profiles of the species with single-male mating behaviour has higher contribution from pre-meiotic and somatic cell types, however the testis transcriptome profiles of the species with multi-male mating behaviour has higher contribution from meiotic and post-meiotic cell types. The proportion of the cell types present in the testis is expected to be changing with development. I therefore tested the hypothesis that single-male species have more immature testes compared to those of multi-male species. Indeed, calculating the levels of correlation of the whole testis transcriptome profiles of different species with testis transcriptome profiles of mice or macaques at different stages of maturation, I found a similar trend with cell type analysis: Single-male species' testis transcriptome profiles similar to those of immature mice and immature macaques.

I then clustered all common genes present in all the datasets into four groups based on their expression profiles. Two of the clusters (about 53% of the genes) showed either increasing or decreasing gene expression profiles in the mouse and macaque testis development datasets. The same genes distinguished single- and multi-male species' profiles as well, indicating that convergent evolution of whole testis transcriptome profiles affects a large proportion of the transcriptome.

To conclude, although a relationship between mating behaviour and testis size was known among hominids, whether such a relationship was also present at the transcriptome level was not known. My work shows that whole testis transcriptomes are affected by cell type proportions and these evolve convergently according to the mating behaviour of species.

Keywords: transcriptome evolution, gene expression, microarray, RNA-seq, testis size, mating behaviour, hominid

ÖZ

HOMİNİDLER ARASI TESTİS TRANSKRİPTOM EVRİMİ

Sağlıcan, Ekin
Yüksek Lisans, Biyoloji Bölümü
Tez Yöneticisi : Doç. Dr. Mehmet Somel

4 Ocak 2018, 73 sayfa

İnsanlar ve yaşayan en yakın akrabaları arasındaki göreli testis büyüklüğü farkı dikkate değer. Her ne kadar şempanzeler insanlara filogenetik olarak daha yakın olsalar da, insanların göreli testis büyüklüğü gorillerinkine daha benzer. Şempanzelerin göreli testis büyüklüğü hem insanlardan hem de gorillerden daha fazla, dahası çok daha uzak bir tür olan makaklara daha benzer. Testis büyüklüğündeki bu farkların bu türlerdeki üreme davranışıyla ilişkili olduğu ve yakınsak olarak evrildiği düşünülmekte. Daha belirgin bir biçimde söylemek gerekirse, insan ve goril gibi tek-erkekli üreme davranışı gösteren türler küçük testislere, şempanze ve makak gibi çok-erkekli üreme davranışı gösteren türler büyük testislere sahip.

Bu tez, içerisinde insan, şempanze, goril, makak, marmoset, iki farklı tür fare, sıçan, ornitorenk ve possum olmak üzere 10 farklı türe ait testis transkriptom datası barındıran, toplamda 8 RNA-dizileme ve mikroçip verisi içermektedir. Bu datasetlerini kullanarak karşılaştırmalı meta-analiz yapmaktayım. İlk olarak insan ve şempanze testislerinde farklı anlatılan genlerin, goril ve makak testis transkriptomlarıyla farklı seviyelerde ilişkilendiklerini gösteriyorum. Bu türler arasındaki testis büyüklüğü ilişkisinde olduğu gibi, bu analiz insanlarla goriller ve şempanzelerele makalar arasındaki yüksek transkripsiyon benzerliğini ortaya koyarak tüm testis gen anlatımında yakınsak evrimin izlerini açığa çıkarmakta.

Bu türlerin testis transkriptom profilleri arasındakı farklılaşmayı açıklayabilecek bir olası sebep, testiste bulunan hücre tiplerinin göreli katkısı olabilir. Çalışmanın ikinci yarısında, testiste bulunan hücre tiplerinin analizde kullanılan türlerin testis-

lerindeki göreli katkısını bulmak için fare testisinden izole edilmiş hücre tiplerinde anlatılan genleri kullandım. Bu analizin sonuçları daha önceki bulgularla tutarlı: Tek-erkekli üreme davranışına sahip türlerin testis transkriptom profilleri, mayoz öncesi ve somatik hücre tiplerinden daha yüksek oranda bir katkıya sahip, diger yandan, çok-erkekli üreme davranışına sahip türlerinki, mayoz ve mayoz sonrası hücre tiplerinden daha yüksek oranda bir katkıya sahip. Testiste bulunan hücre tiplerinin oranının gelişimle birlikte değişmesi beklenir. Bu nedenle tek-erkekli türlerin çok-erkekli türlerle kıyaslandığında daha az gelişmiş testislere sahip olduğu hipotezini test ettim. Gerçekten de, farklı türlerin tüm testis transkriptom profilleri ile olgunlaşmanın değişik aşamalarında olan fare ve makakların testis transkriptom profilleri arasındaki ilişkiyi hesapladığımda, hücre tipi analizinde elde ettiğim sonuçlara benzer sonuçlar elde ettim: Tek-erkekli türlerin testis transkriptom profilleri olgunlaşmamış fare ve makaklara benzerdi.

Daha sonra bütün datasetlerinde bulunan ortak genleri, gen anlatım profillerine göre dört gruba ayırdım. Bu gruplardan ikisi (genlerin yaklaşık %53'ü) fare ve makak testis gelişim datasetlerinde ya azalan ya da artan gen anlatımı profilleri gösterdiler. Aynı genler tek- ve çok-erkekli türlerin profillerinde de ayrışmışlardı. Bu da tüm testis transkriptom profillerinin yakınsak evriminin transkriptoma büyük oranda etki ettiğine işaret etmekte.

Sonuç olarak, hominidler arasında testis büyüklüğü ile üreme davranışı arasında bir ilişki olduğu her ne kadar bilinse de, bu ilişkinin transkriptom seviyesine yansıyıp yansımadığı bilinmemekteydi. Çalışmam tüm testis transkriptomunun hücre tipi oranlarından etkilendiğini ve türlerin üreme davranışına bağlı biçimde yakınsak olarak evrildiğini göstermekte.

Anahtar Kelimeler: transkriptom evrimi, gen anlatımı, mikrodizin, RNA dizileme, testis

This thesis is dedicated to
the salvation of proletariat.

ACKNOWLEDGEMENTS

This thesis is the product of endless hours of enthusiasm and dedication. Besides being a scientific contribution, it is a great lifetime experience and inspection of patience. I have learned so much more than what is written in all these pages.

The both secret and obvious character who made this thesis possible and real is my mentor, Mehmet Somel. He is inexplicable on doing so many things simultaneously, inspiring on so many aspects of life, delightful to brainstorm any time, always able to bring a new aspect to the table with his eccentric perspective and thankfully an helpless optimist not knowing the meaning of giving up.

I would like to thank all my thesis committee members each. Can Bilgin for motivating a thrilled random student years ago and making her aware of the beauty of the enthusiasm. Mesut Muyan for teaching me being prepared for various kind of questions and forcing me to be more familiar to different aspects of my topic. Tolga Can for making me curious for the underlying mechanisms of algorithms and saving me time both for conducting the analyses and interpreting the results. İdil Yet for persuading me to start over every single time and showing me the valuable sides under the surface.

I also would like to thank Philip Khaitovich, Rafik Neme, Diethard Tautz, Haiyang Hu and Melike Dönertaş for kindly sharing their data or codes.

Somel Lab with all the previous and present members deserves blessing for the dawning of this thesis. I especially would like to thank Melike Dönertaş as the principle component of the group for her scientific support and effort in every fraction of this thesis despite our mutually exclusive ways of studying.

My family merits credit for emotional support. It would be so much harder, even impossible to accomplish any academical achievement without them.

I would like to thank Nazlı Somel saving me from confusion by making me remember the principle of parsimony with her line of reasoning, her sister for being

our coffee sponsor, sweet lady Tülin Çetin for her endless curiosity.

This work was partially supported by a TUBİTAK 2232 grant (no: 114C040) and a Science Academy of Turkey through the BAGEP program.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

INDEL	Insertion and Deletion
SRY	Sex-Determining Region Y
RNA-seq	RNA Sequencing
GTF	Gene Transfer Format
GEO	Gene Expression Omnibus
SRA	Sequence Read Archives
GEO	Gene Ensembl Omnibus
TBA	Threaded Blockset Aligner
FPKM	Fragments Per Kilobase of Exon Per Million Fragments Mapped
MAF	Multiple Alignment File
CDF	Chip Definition File
NCBI	National Center for Biotechnology Information
CEL	Affymetrix Data File Format
FDR	False Discovery Rate
PRE	Pre-meiotic and Somatic
POST	Meiotic and Post-meiotic

CHAPTER 1

INTRODUCTION

1.1 Human Evolution

Humans or *Homo sapiens*, which means "wise person" in Latin, are not only the only member of the *Homo* genus still living today but also the only member of hominins, referring to all species that are phylogenetically closer to humans than to chimpanzees. In the last decades, it has been accepted that hominin evolution took place in many adaptive radiation events forming different lineages, all of which except modern humans, are now extinct (Leakey et al., 2001).The most closely related species to humans currently extant are chimpanzees belonging to the *Pan* genus (King & Wilson, 2007). The extent of anatomical and behavioral similarities between these two species are not so surprising, considering the fact that their common ancestor is estimated to have lived approximately 5-8 million years ago (Wood, 2002),(Diogo, Molnar & Wood, 2017).

Although archaeological findings on human ancestors are scarce, the extent of our knowledge about the ancestors of chimpanzees is even less due to the limitations on finding well-preserved archaic chimpanzee fossils (Carroll, 2003)(McBrearty & Jablonski, 2014). This is mainly due to the environmental conditions that chimpanzees' ancestors lived in. Even if it is not currently possible to obtain any genetic material from the common ancestors of humans and chimpanzees living today, it is still possible to extract information about these species' histories from modern genomes, using other species' genomes as out-groups. Since many parts of the genomes are passed along throughout generations with small modifications, it is possible to track changes in each lineage and figure out genetic characteristics underlying unique phenotypic features of either species as well as their common ancestors.

1.1.1 The Similarities and Dissimilarities between Humans and Chimpanzees

The anatomical similarities between humans and their closest relatives are overwhelming. The first thing that comes to the eye is the overall body plan for both males and females of either species, so obvious that no one can identify a major qualitative difference between these lineages. Moreover, arrangement of the internal organs is the same between the two species just like the arrangement of the bone structures. The number of examples can be increased when we consider not only the presence of specific anatomical structures, but also those which both species lack, such as tails.

The behavioural similarities are also striking just like the anatomical ones. Like humans, chimpanzees also live in social groups and are able to communicate with each other using gestures, facial and vocal signals (Pollick & de Waal, 2007). It is shown that tickling-induced laughter in chimpanzees is the same as in humans (Davila Ross, J Owren & Zimmermann, 2009). The evidence on the behavioural similarities between these species is building up day by day and even include characteristics as complex as fairness (Proctor et al., 2012).

All of these anatomical and behavioural similarities are grounded on genetic similarities as expected. Overall genome similarity of humans and chimpanzees is 95%, 3.4% of the remaining 5% is composed of small insertions an deletions (INDELs) and only the 1.2-1.4% can be accounted for single base substitutions which are responsible from the dissimilarities (Britten, 2002) (Cheng et al., 2005). In other words, across 100 alignable bases, only 1 differs on average between humans and chimpanzees, which is about an order of magnitude higher than the difference between two human genomes (Auton et al., 2016).

In order to understand what genetic changes make humans phenotypically different from other species, first thing to do is to compare them with their distant cousins and focus on differences rather than similarities while doing so. Humans have larger brains compared to their closest extant relatives; on the other hand, chimpanzees have greater muscle power than humans (Bozek et al., 2014). Advanced tool making capacity, bipedalism, relative limb length, small canine teeth, reduced

hair cover, presence of a chin can be counted among the traits distinguishing humans (Carroll, 2003). All of these changes are the visible effects of evolution on humans and chimpanzees since the path that they have taken after their ancestral lineages separated millions of years ago.

What are the genetic basis of these unique characteristics? There are two main approaches to determine genes specific to human lineage that may underlie such phenotypes. One of them is to focus on individual genes which have been linked to putatively human-specific phenotypes through medical genetics or functional genomic studies. Examples include *FOXP2*, which is related to human speech (Lai, Fisher, Hurst, Vargha-Khadem & Monaco, 2001)(Enard et al., 2002).

The other approach is focusing on whole genomes and to compare humans with other species. In 2011, McLean *et. al.* have identified complete deletions in the human reference genome by comparing it with the complete chimpanzee genome and found a deletion in close proximity to an enhancer of the *GADD45G* gene, which controls tissue growth, and the authors suggested a possible role for this deletion in the evolutionary expansion of specific regions in the human brain (McLean, 2011). Another study in 2012, used a novel genetic approach to identify missing loci from the reference genome corresponding to human specific gene families and found a *SRGAP2* duplication specific to human lineage (Dennis et al., 2012).

Once the genes possibly underlying the unique characteristics of humans are identified either way, it is possible to focus on the specific functions of these genes, as in the work of Charrier and her colleagues. They have used both *in vitro* and *in vivo* approaches to find out the function of *SRGAP2* and its human paralogs in the neocortex, which is the part of the brain thought to be highly important for human cognitive abilities (Charrier et al., 2013). Another example of mixing these two main strategies is the experimental confirmation of the human specific alterations for *HACNS1* in transgenic mice (Prabhakar et al., 2009). A more recent study conducted in 2015 used transcriptome data of developing mouse and human neocortex and identified a human-specific gene *ARHGAP11B* lacking in mouse. Expressing this gene in mouse induced gyrification in developing mouse brain (Florio et al., 2015).

Although there are numerous examples of research on human-specific characteristics, there is only a little known about the genetic basis underlying them. This is mainly due to the noisy nature of the genome, difficulties in the experimental process and the scarcity of support for such scientific research from the society and the government. In order to answer the question asked above, theoretical and experimental research must expand.

1.2 Testis Development and Evolution

Testes or *testicles*, which are homologous to ovaries in females are crucial elements of male reproductive system by being the place where sperm production takes place. They also act as an endocrine gland being able to produce male-related hormones. The terms testes and testicles are derived from the Latin word *testis*, meaning "witness". One theory suggests that, in Biblical times, men held their testicles as a witness in courts or when making promises, which can also be the source of the phrase "swearing upon the testicles" (Anderson, Hicks & Holmes, 2002).

The typical number of testes in vertebrates is two and many mammals have external testes. In mammalian embryonic stages, the gonads have the potential to form either ovaries or testes. The presence of the sex-specific *SRY* on Y chromosome is the decision center having the capacity to regulate many downstream pathways for sex determination such that the differentiation of Sertoli cells (Piprek, 2010), having an essential role in initial differentiation and development of testes (Palmer & Burgoyne, 1991).

Inside the testis, there are seminiferous tubules, in which spermatogenesis takes place from puberty till death. Through this sperm production process, spermatogonia having 2n chromosomes form primary spermatocytes which also carry 2n chromosomes via mitosis. These spermatocytes then form spermatids bearing 1n chromosomes through the process called meiosis. Haploid spermatids develop into polar spermatozoa with tails, and are released into the epididymis where they gain the ability to move and fertilize the ova (Pansky, 1982).

Among seminiferous tubules, there are immature and mature Leydig cells. Mature

Leydig cells are the cells mainly responsible for the production of testosterone, an essential hormone for sexual development. Functions of testosterone include controlling testes size, controlling timing of puberty, and driving the development of secondary sexual characteristics such as facial hair.

1.2.1 Evolution of the Testis in Mammals

Testis is one of the tissues where humans and chimpanzees display conspicuous differences. The one highly striking difference between humans and chimpanzees is the difference between their testis sizes. Although having smaller body weights than humans, chimpanzees have three times larger testicles. Meanwhile, humans have nearly 1.5 times larger testicles than gorillas that are huge animals being 2.5 times larger than humans (Harcourt, Harvey, Larson & Short, 1981).

This difference in testis size is explained by the mating behaviour of these species (Harcourt et al., 1981). Female chimpanzees mate with multiple males during their estrous, which makes sperm competition between individuals possible. On the other hand, humans have monogamy or rather female monandry as their mating type; therefore, sperm competition is within-individual competition (Kramer & Russell, 2015) (Marlowe, 2000). Thus, producing faster sperm in higher quantities is more advantageous for polygamous, more specifically, polyandrous or fully promiscuous (multi-male) type of mating.

When we take gorilla and chimpanzee for comparison as closely related species to humans, a study conducted in 2011 have identified clear histological differences between the testes of these two species. To begin with; seminiferous epithelium in gorilla is thin, whereas in chimpanzee it is thicker. Moreover, there are many spermatocytes and spermatids in chimpanzee seminiferous epithelium although these cell types are only sparsely scattered within gorilla testes. When we look at the interstitial tissue, Leydig cells are abundant in gorilla; on the other hand, this tissue is loose in chimpanzee and there are only a few Leydig cells. The chance of observing sperms or mature spermatids in gorilla testes is significantly lower than observing them in chimpanzee testes. The total length of seminiferous tubules is also shorter in gorillas when compared to chimpanzee (Fujii-Hanamoto, Matsubayashi, Nakano, Kusunoki & Enomoto, 2011).

Traits directly affecting reproduction are under strong selection as illustrated by many examples and can thus evolve rapidly (Wu, Johnson & Palopoli, 1996) (Ting, Tsaur, Wu & Wu, 1998) (Nurminsky, Nurminskaya, de Aguiar & Harti, 1998)- which is also expected to be the case for male reproductive traits in humans and chimpanzees (Wyckoff, Wang & Wu, 2000). It is therefore not surprising that the testis have evolved dramatically different phenotypes between these closely related species, once their mating types diverged.

This testis size-mating type relation can also be observed in other closely related species. For example, gorillas have harem systems as their mating behavior, in which a male mates with multiple females (Harcourt et al., 1981). Therefore, there is a little opportunity for competition between males. Accordingly, male gorillas, with much larger bodies than humans, have even smaller testicles than humans. A more distant species, such as rhesus macaques mate in a polygamous way, and rhesus macaque males have large testicles like those of chimpanzees. To be more specific, % testes weight (g) / body weight (kg) ratios of macaque and chimpanzee are 0.5 and 0.27 respectively; on the other hand, the same ratios for gorilla and human are 0.02 and 0.06 respectively (Harcourt et al., 1981).

The fact that the species showing more monandrous (single-male) type of mating behaviour like humans and gorillas have smaller testes, and more polyandrous/-promiscuous (multi-male) ones such as chimpanzees and macaques have larger testes, implies convergent evolution. Convergent evolution means the evolution of the same trait multiple times in different lineages independent from one another. We can imagine two convergent evolution scenarios explaining the observation that humans and gorillas have small, and chimpanzees and macaques have large testicles: Either the common ancestor of all these species had small testicles, and chimpanzees and macaques later evolved larger ones due to strong selection caused by multi-male mating; or the common ancestor had large testicles, and relaxation of selection on testis size due to more monogamous mating behaviours of humans and gorillas led to the evolution of smaller testes in these lineages. In both scenarios, we need to assume convergent evolution, because neither humans and gorillas, nor chimpanzees and macaques, are sister species in this phylogeny constructed based on NCBI Taxonomy Database (Federhen, 2015)(**Figure 1.1**).

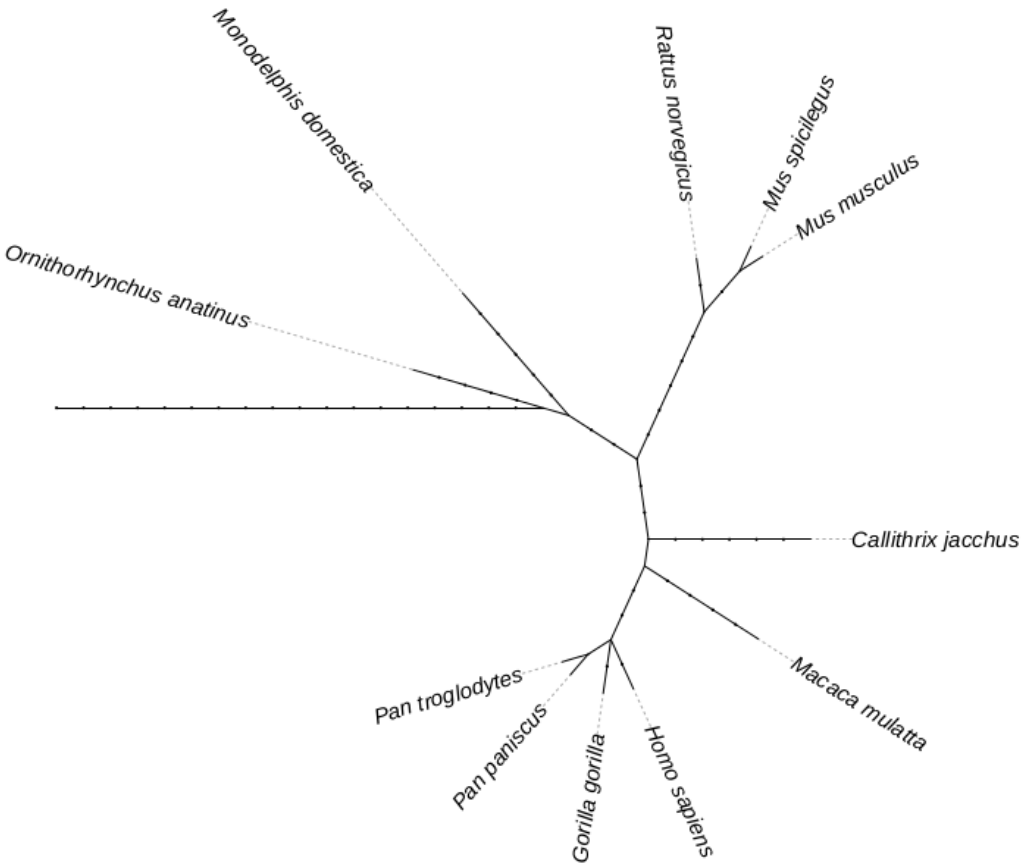

Figure 1.1: The phylogeny of the species used.

Such a huge difference in phenotype and evolution of the same trait more than once in a relatively short period of time in course of evolution, within approximately 25 million years (Rogers et al., 2005), could possibly be explained by small changes in the genome having large effects. For example, a change in the timing of the expression of a transcription factor controlling the expression of many genes, can have widespread effects on the tissue level.

A study published in 2006 found signs of selection on many genes related to fertility and reproduction in various human populations. Some examples of their findings include *RSBN1*, a gene involving in the basic protein structure of sperm in East Asians and Yoruba; the genes *SPAG4* in Europeans and East Asians, and *ODF2* in Europeans having functions in sperm motility; the genes *ACVR1* in Europeans and

CPEB2 in Yoruba affecting sperm and egg viability; and *TGM4* regulating female immune response to sperm in Europeans. Some of these genes the authors identified also show indications of sustained selection over long timescales (Voight, Kudaravalli, Wen & Pritchard, 2006).

Traits affecting reproductive success are shown to have evolved faster in a wide range of species; from fruit flies to humans as mentioned earlier in this section. Accordingly convergent evolution in these type of traits is not so unexpected. If a trait strongly increases the chance of having progeny, it can be selected multiple times in different lineages independent from one another. For the relaxation of selection scenario, if the effect of the selection on the trait is lifted due to changes in the social behaviour of the species, the very same anatomical changes can be observed multiple times in different lineages. These anatomical changes can be the result of a mutation that happened more than once in both lineages, or different mutations on the same gene, or different mutations on different genes.

Anatomical changes can readily evolve by altering the developmental process: changes in the rate and duration of growth (cell division) in one tissue relative to the body average through a process called heterochrony (Raff, Wray & Biology, 1989). In the case of humans and chimpanzees, human development is in general about 1.5-2 times longer than it is in chimpanzees (Wood & Collard, 1999). However, if specific organs have evolved at higher or lower growth rates compared to this average, their mature forms can gain evolutionarily novel anatomic proportions.

For example, heterochrony has been used to explain various features of human brain evolution (Rice, 2002) . There are also findings that support the idea that the human skull shows paedomorphosis. In one of these studies, the authors found out that human growth is accelerated at first, and then this acceleration is followed by a strong decrease in the pace of the growth compared to the chimpanzee (Penin, Berge & Baylac, 2002). Heterochrony altering developmental process is not limited to brain evolution as in the previous examples. For instance, it is shown that heterochronic development of the gonads of *Hynobius retardatus* suggests neotenic reproductive characteristics in this species as a response to their changing environment (Kanki & Wakahara, 2001).

1.2.2 Comparative Transcriptome Analyses and Testis Evolution

Although nearly all somatic cells of an individual share the very same genome in principle, a different set of genes is expressed in each cell specific for the function of each tissue. Even in the same tissue, there is compartmentalization and every cell type shows a different pattern of gene expression. The genes that are expressed are represented in the transcriptome: the whole mRNA content of a cell at a specific time in a specific tissue. In other words, the mRNA molecules present in a cell continuously change in developmental stages, in different tissues, even during the course of a day.

In contrast to the genome of a cell, the transcriptome varies by both internal and external factors. Hence, genomes are virtually static, whereas transcriptomes are dynamic providing the connection between the genes and their functions. Any change in the environment such as heat, pH or presence of a pathogen or a hormone can be counted among the environmental factors that are affecting the transcriptome of a cell. The internal mechanism that enables fundamental changes in the transcriptome composition is the act of transcription factors controlling the transcription of many target genes. Moreover; the amount, durability or timing of the expression of a transcription factor can change the whole transcriptome instantaneously or gradually, by affecting the downstream pathways. The effect can be tremendous if the transcription factor of interest has a role in development. For example, *Foxa* subfamily of transcription factors having functions in multiple developmental stages starting from early development might lead to lethality if absent (Friedman & Kaestner, 2006).

Microarrays or DNA chips were the first widely used method for obtaining transcriptomic data, since the early 2000's (Pease et al., 1994). Commercially available and commonly used Affymetrix chips contain 20-25 bp-long oligonucleotide sequences called probes that are specifically designed to match with known genes' sequences (Trevino, Falciani & Barrera-Saldana, 2007). One of the requirements of microarrays essential for detection is the preparation of the sample as isolating the mRNA content of cells, converting it into cDNA or cRNA, amplifying the molecules, and labelling them with a fluorescent dye. The intensity of the light emitted after hybridization of the array with the labelled target sequences followed by the

incubation and washing steps allows the quantification of the mRNAs transcribed in the cells via a confocal scanner (Trevino et al., 2007).

A very efficient way to use microarrays is to compare samples of interest with control group samples such as cancer cells with healthy cells of an individual, or a cell line before and after being exposed to a drug. These type of comparisons can be done by using just one chip referred as two-dye assays, labelling the sample and the control with different types of florescent dyes emitting light at different wavelengths. This technology requires an additional step to transform the readings to a ratio for comparison and more appropriate if a minority of the genes is expected to be changed (Trevino et al., 2007).

The main drawback of DNA chips is that they require information about the gene sequences in order to design probes. The probes on the chips are designed for a specific purpose like the known genes of a species or genes of various bacterial species having pathogenic effect. So it is not possible to detect mRNAs with an unknown sequence.

A more recent technology to obtain transcriptomic data is RNA-sequencing (RNA-seq) based on next (second) generation DNA sequencing rather than hybridization. This technology does not require any prior information about the target sequences; therefore, it can be used for detection and quantification of novel transcripts. With the next generation sequencing technologies, it became possible to study different populations of RNA such as miRNAs to greater extent or events such as alternative splicing.

1.2.3 Transcriptome Analyses of Testis Development and Evolution

What makes human males and females physiologically different from each other is the sex-determining region of Y chromosome (*SRY*). Once the *SRY* region is activated in males, it triggers a series of events that lead to the development of testis from undifferentiated gonadal primordia (bipotential gonad). If *SRY* is not present, or inactive, ovaries form as default. After the formation of either ovaries or testes, gonadal hormones shape further changes. *SRY* expression turns a subset of somatic cells into Sertoli cells. Gonad size starts to increase. Peritubular myoid cells sur-

round Sertoli cells as a flattened single layer, providing structural support as well as helping mature sperm to move through the seminiferous tubules (Wilhelm, Palmer & Koopman, 2007).

In 2003, Schultz and his colleagues studied transcriptomes of developing mouse testis enriched in Sertoli cells and interstitial cells and found out that a remarkable portion of the mouse genome (nearly 4%) is dedicated to be expressed by male germ cells late in development (Schultz, Hamra & Garbers, 2003). A portion of these genes expressed post-meioticly could be a possible explanation for the gene expression difference detected in the testes of human and chimpanzee.

More than 10 years ago, another study investigated expression differences between testes of human and chimpanzee as well as the tissues brain, heart, liver and kidney using microarrays. This showed that genes expressed in the brain have changed the least between these two species. On the other hand, the ratio of expression divergence between species to diversity within species was higher in testis than in any other tissue. According to Khaitovich and his colleagues, the expression changes detected in testis as well as rapid sequence change among the X-chromosomal genes expressed in testes could be evidence of positive selection (Khaitovich et al., 2005) (Khaitovich, Enard, Lachmann & Paabo, 2006). It is known that in mammals, X-linked genes are evolving slightly faster than autosomal ones since the effective population size of X chromosome is less (Johnson & Lachance, 2012).

Another study conducted in 2011 gathered transcriptomic data from six organs across ten species including all major mammalian lineages, allowing them to compare the gene expression evolution among lineages as well as organs. Their findings include that the transcriptome evolution is slower in rodents than it is in apes, and slower in nervous system than it is in testis. The rapid transcriptome change in testis puts gorilla and humans as a group and chimpanzee and bonobos as a separate one in terms of genes expressed in this tissue. The authors point out that this difference is consistent with the evolution of mating patterns among African apes (Brawand et al., 2011).

In this study, I have analysed testis gene expression profiles of various species in order to find the reflection of the testis transcriptome on the testis size affected

by the mating behaviour of these species. My hypothesis is that the signs of convergent evolution observed in testis size can also be detected on testis gene expression levels. The high transcriptome complexity of the testis, more specifically the remarkable transcriptome divergence of meiotic spermatocytes and postmeiotic spermatids (Soumillon et al., 2013), could be an explanation for the relationship between testis size and mating behaviour.

CHAPTER 2

MATERIAL AND METHOD

2.1 Pre-processing of the Datasets

RNA-seq datasets and microarray datasets are pre-processed in a similar way as described in the following sections in order to remove bias due to platform differences.

2.1.1 RNA-seq Datasets

After downloading all the RNA-seq data (**Table 2.1**) in sra format, the files are converted to fastq format via "fastq-dump" with split-3 option. The quality of the reads is checked with FASTQC. The reads were aligned to each species genomes obtained from Ensembl version 83 (Cunningham et al., 2015), by using TopHat2 (Kim et al., 2013) with the options: “tophat2 -i 40 -I 1000000 -a 8 -N 1 -g 2 —no-novel-juncs”. No novel junctions was searched during the alignment process. The list of one-to-one orthologous genes (n=14875) of Ensembl version 83 for all the species present in the analysis were downloaded from BioMart (Smedley et al., 2015). Gene Transfer Format Files (GTFs) were filtered only to contain these common genes with an in-house python code (see **Appendix A**) and only the perfectly matched unique reads were used. The reads were quantified with Cufflinks (Trapnell et al., 2013) providing the genomes and the filtered GTFs with the -b and -G options respectively. Genes showing no expression value in none of the subjects for a dataset as well as transcripts with multiple expression values were removed.

Table 2.1: RNA-seq dataset summary.

GEO Acc.	Dataset	Species	SRR Number
GSM752707	Brawand2011	*Homo sapiens*	SRR306857
GSM752708	Brawand2011	*Homo sapiens*	SRR306858
GSM752690	Brawand2011	*Pan paniscus*	SRR306837
GSM752678	Brawand2011	*Pan troglodytes*	SRR306825
GSM752663	Brawand2011	*Gorilla gorilla*	SRR306810
GSM752642	Brawand2011	*Macaca mulatta*	SRR306789
GSM752643	Brawand2011	*Macaca mulatta*	SRR306790
GSM752629	Brawand2011	*Mus musculus*	SRR306775
GSM752630	Brawand2011	*Mus musculus*	SRR306776
GSM752611	Brawand2011	*Monodelphis domestica*	SRR306755
GSM752613	Brawand2011	*Monodelphis domestica*	SRR306756
GSM752583	Brawand2011	*Ornithorhynchus anatinus*	SRR306739
GSM752585	Brawand2011	*Ornithorhynchus anatinus*	SRR306740
SRX335333	Bellott2014	*Callithrix jacchus*	SRR952610
GSM1227961	Cortez2014	*Callithrix jacchus*	SRR975185
GSM1227962	Cortez2014	*Callithrix jacchus*	SRR975186
GSM1227963	Cortez2014	*Callithrix jacchus*	SRR975187*
unpublished	NemeMus	*Mus musculus*	unpublished
unpublished	NemeMus	*Mus spicilegus*	unpublished

Subject with an asteriks (*) is not used in the analysis.

Brawand2011

This dataset (Brawand et al., 2011) downloaded from NCBI GEO database (Edgar, Domrachev & Lash, 2002) with the accession number GSE30352 includes adult gene expression data for the testicles of two humans (*Homo sapiens*), one bonobo (*Pan paniscus*), one chimpanzee (*Pan troglodytes*), one gorilla (*Gorilla gorilla*), two macaques (*Macaca mulatta*), two mice (*Mus musculus*), two platypuses (*Ornithorhynchus anatinus*) and two opossums (*Monodelphis domestica*). The single-

end libraries were constructed using the platform Illumina Genome Analyzer IIx and then the reads were aligned to the genomes of Ensembl version 83 (**Table 2.2**) using TopHat2. Chimpanzee genome was used for bonobo since it was the only available genome in the Ensembl database. The reads were quantified with Cufflinks using GTFs of Ensembl version 83 filtered to contain only one-to-one orthologous genes of all the species used in the analyses. A second version of this dataset was also constructed by only using the one-to-one orthologous genes of the primates used in the analysis. The filtering process was done by using MafFilter version 1.1.4-1 (Dutheil, Gaillard & Stukenbrock, 2014), MAFs constructed by TBA (Threaded Blockset Aligner) (Blanchette et al., 2004) were filtered only to contain exact matches for every species. The FPKM (Fragments Per Kilobase of Exon Per Million Fragments Mapped) values obtained from the preprocessing stage were log2 transformed as log2("FPKM"+1) and quantile normalized for further analysis.

Table 2.2: Genome versions.

Species	**Genome Assembly Version**	**Size**
Homo sapiens	GRCh38.p5	471537 KB
Pan paniscus	CHIMP2.1.4	480099 KB
Gorilla gorilla	gor.Gor3.1	467795 KB
Macaca mulatta	MMUL 1.0	494205 KB
Callithrix jacchus	C_jacchus3.2.1	487721 KB
Mus musculus	GRCm38.p4	484424 KB
Rattus norvegicus	$\text{Rnor}_6.0$	498640 KB
Ornithorhynchus anatinus	OANA5	363223 KB
Monodelphis domestica	monDom5	797959 KB

Cortez2014

This dataset (Cortez et al., 2014) downloaded from GEO database with the accession number GSE50747 includes gene expression data for the testicles of two infant

marmosets (*Callithrix jacchus*). One of the subjects had a technical replicate, one of the replicates was randomly selected (**Table 2.1**) and used in the analysis. Trimmomatic (v35) software and TruSeq3-SE adapter sequence library (Bolger, Lohse & Usadel, 2014) are used in order to remove adapter sequences. The single-end libraries were constructed using the platform Illumina Genome Analyzer IIx and then the reads were aligned to the marmoset genome version of Ensembl release 83 using TopHat2. The reads were quantified with Cufflinks using GTFs of Ensembl version 83 filtered by containing only one-to-one orthologous genes of all the species used in the analyses. A second version of this dataset was also constructed by only using the one-to-one orthologous genes of the primates used in the analyses. The filtering process was done by using MafFilter, MAFs constructed by TBA were filtered only to contain exact matches for every species. The FPKM values obtained from the preprocessing stage were log2 transformed as log2(FPKM+1) and quantile normalized for further analysis.

Bellott2014

This dataset (Bellott et al., 2014) downloaded from NCBI SRA (Leinonen, Sugawara, Shumway, Nucleotide & Database, 2011) with the accession number SRX335333 includes gene expression data for the testicles of an adult marmoset, *Callithrix jacchus*. The paired-end libraries were constructed using the platform Illumina MiSeq and then the reads were aligned to the marmoset genome version of Ensembl release 83 using TopHat2. The reads were quantified with Cufflinks using GTFs of Ensembl version 83 filtered by containing only one-to-one orthologous genes of all the species used in the analyses. A second version of this dataset was also constructed by only using the one-to-one orthologous genes of the primates used in the analysis. The filtering process was done by using MafFilter, MAFs constructed by TBA were filtered only to contain exact matches for every species. The FPKM values obtained from the preprocessing stage were log2 transformed and quantile normalized for further analysis.

NemeMus

This dataset generated in the Department of Evolutionary Genetics, Max-Planck Institute for Evolutionary Biology had already been quality checked and trimmed; downloaded in fastq format including transcriptome data of testes of one *Mus mus-*

culus from Massif Central France and one *Mus spicilegus* from Eastern Europe. The reads were aligned to the mouse genome version of Ensembl release 83 using TopHat2. The reads were quantified with Cufflinks using GTFs of Ensembl version 83 filtered by containing only one-to-one orthologous genes of all the species used in the analyses. A second version of this dataset was also constructed by only using the one-to-one orthologous genes of the primates used in the analysis. The filtering process was done by using MafFilter, MAFs constructed by TBA were filtered only to contain exact matches for every species. The FPKM values obtained from the preprocessing stage were log2 transformed and quantile normalized for further analysis.

2.1.2 Microarray Datasets

In order to avoid biases, the probes of the microarrays were aligned to the specific species' genomes using Bowtie2 (Langmead & Salzberg, 2012), which is also the alignment algorithm of TopHat2, with the same options used for the RNA-seq data as mentioned in **Section 2.1.1**. In the cases where a microarray designed for one species was used for generating data of another species, the probes were aligned to the sample species' genome and only the perfect matches corresponding to a specific gene were used. Chip Definition Files (CDFs) were filtered as to contain only these desired probes for the quantification process with an in-house python code (see **Appendix B**). Genes showing no expression value in none of the subjects for a dataset were removed.

Schultz2003

This dataset (Schultz et al., 2003) includes testis development data for 15 mice (*Mus musculus*) of ages ranging from newborn to adult (1 day old n=3 mice , 4 days old n=2 mice, 8 days old n=2 mice, 11 days old n=2 mice; 14, 18, 21, 26, and 29 days old n=1 mouse each and an adult mouse – the adult mouse is treated as 42 days old in the analyses). The Affymetrix CEL files (Affymetrix Data File Format) constructed using the chips MGU74 A, B and C were downloaded from NCBI GEO database with accession number GSE640. The CDF containing the probe information (version 20.0.0) for this array was downloaded from Microarray Lab (Dai et al., 2005), Molecular and Behavioral Neuroscience Institute, University of Michigan and the CEL files were read in R (Ihaka & Gentleman, 1996)

Table 2.3: Microarray dataset summary.

Accession Number	Dataset	Species	Microchip
GSE640	Schultz2003	*Mus musculus*	Affymetrix MGU74 A,B,C
E-AFMX-11	Khaitovich2005	*Homo sapiens* *Pan troglodytes*	Affymetrix Human HGU133Plus2
GSE4193	Namekawa2006	*Mus musculus*	Affymetrix Mouse 430.2
E-TABM-130	Chalmel2007mus	*Mus musculus*	Affymetrix Mouse 430.2
E-TABM-130	Chalmel2007rat	*Rattus norvegicus*	Affymetrix Rat 230.2
unpublished	Khaitovich Testis Development Data	*Macaca mulatta*	Affymetrix Human Gene1.0ST

using Bioconductor "affy" package (Gautier, Cope, Bolstad & Irizarry, 2017). The expression values were background corrected, log transformed and normalized via "exprs" and "rma" functions present in the "affy" package. Only the one-to-one orthologous genes for all species used in the analysis, namely; human, chimpanzee, gorilla, macaque, marmoset, mouse, rat, platypus and opossum were selected and then quantile normalized using "normalize.quantiles" function found in "preprocessCore" package (Bolstad, 2016).

Khaitovich2005

This dataset (Khaitovich et al., 2005) includes adult gene expression data of the testicles of 6 humans (*Homo sapiens*) and 5 chimpanzees (*Pan troglodytes*) downloaded from EBI ArrayExpress (Parkinson et al., 2005) with accession number E-AFMX-11. Affymetrix Human HGU133Plus2 microarrays were used for both species. CDF for this microarray chip was downloaded from Microarray Lab and the probe sequences were aligned to the genomes of Ensembl version 83 of the

sample species' genomes using Bowtie2. Only the perfectly matched uniquely mapped probes which are also common for both humans and chimpanzees were used for the quantification of the gene expression using the modified CDF mentioned in (**Subsection 2.1.2**). The CEL files were read in R using Bioconductor "affy" package using the filtered chip definition files. Then the data was background corrected, log transformed and normalized via "exprs" and "rma" functions present in the "affy" package. Only the one-to-one orthologous genes for all species used in the analysis were selected and then quantile normalized using "normalize.quantiles" function found in "preprocessCore" package. A second version of this dataset was also constructed by only selecting the one-to-one orthologous genes of the primates used in the analysis.

Namekawa2006

This dataset (Namekawa et al., 2006) for cell types of mouse (*Mus musculus*) testicles including spermatogonia, spermatocytes and spermatids of mouse generated via the microarray Affymetrix Mouse 430.2, was downloaded from NCBI GEO database with the accession number GSE4193. The CDF containing the probe information for this array was downloaded from Microarray Lab and the CEL files were read in R using Bioconductor "affy" package. The expression values were background corrected, log transformed and normalized via "exprs" and "rma" functions present in the "affy" package. Only the one-to-one orthologous genes for all species used in the analysis were selected and then quantile normalized using "normalize.quantiles" function found in "preprocessCore" package.

Chalmel2007mus

The part of this dataset [Chalmel2007] that is used includes testis cell type data of 8 pooled samples of mouse (Mus musculus) downloaded from ArrayExpress under the accession number E-TABM-130. The data for Sertoli cells, spermatogonia, spermatocytes and spermatids of this species was generated using the microarray Affymetrix Mouse 430.2. The CDF containing the probe information for this array was downloaded from Microarray Lab and the CEL files were read in R using Bioconductor "affy" package. The expression values were background corrected, log transformed and normalized via "exprs" and "rma" functions present in the "affy" package. Only the one-to-one orthologous genes for all species used in the

analysis were selected and then quantile normalized using “normalize.quantiles” function found in “preprocessCore” package.

Chalmel2007rat

The part of this dataset (Rolland et al., 2007) includes testis gene expression data of 2 pooled sample of adult rat (*Rattus norvegicus*) downloaded from ArrayExpress under the accession number E-TABM-130. The data was generated using the microarray Affymetrix Rat 230.2. The CDF containing the probe information for this array was downloaded from Microarray Lab and the CEL files were read in R using Bioconductor “affy” package. The expression values were background corrected, log transformed and normalized via “exprs” and “rma” functions present in the “affy” package. Only the one-to-one orthologous genes for all species used in the analysis were selected and then quantile normalized using “normalize.quantiles” function found in “preprocessCore” package.

Khaitovich Testis Development Data

This unpublished microarray dataset includes testis development data of n=12 rhesus macaques (*Macaca mulatta*) of ages ranging from 16 days to 26 years (16, 20, 215, 471, 739, 1135, 1205, 1487, 2355, 2570, 8104 and 9518 days old) collected from Suzhou Experimental Animal Center (Suzhou, China). Affymetrix Human Gene1.0ST microarrays were used for the generation of the data. The CDF containing the probe information for this array was downloaded from Microarray Lab and the probe sequences were aligned to the macaque genome of Ensembl version 83 using Bowtie2. Only the perfectly matched probes were used for further analysis. The CEL files were read in R using Bioconductor “affy” package using the filtered CDFs mentioned in (**Subsection 2.1.2**). Then the data was background corrected, log transformed and normalized via “rma” function present in the “affy” package.

2.2 Combining the Datasets

Two versions of the datasets are constructed. First one referred as "Primate" data involves one-to-one orthologous genes of human, chimpanzee, gorilla, macaque and marmoset. Second one referred as "All" involves species as mouse, rat, platypus and opossum additional to the species used for constructing primate data.

2.2.1 Primate Data

Primate-version of the datasets are used.

The microarray dataset Khaitovich2005 consisting of 6 humans and 5 chimpanzees and the RNA-seq dataset Brawand2011 consisting of 2 humans, 2 chimpanzees (1 chimpanzee + 1 bonobo), 2 macaques and 1 gorilla are merged only to contain common genes (n=7308) found in both datasets. Different normalization strategies were followed for merging two different platforms namely, microarray and RNA-seq (explained below).

For the 1st normalization strategy, both datasets were scaled separately as to have a mean of 0 and standard deviation of 1 for each gene in each dataset. The mean of all humans and chimpanzees found in both datasets before the scaling process was added to the scaled version of the dataset in order to retain information on expression level of per gene.

In the 2nd normalization strategy, the means of humans and chimpanzees of each dataset was subtracted from its own dataset and the mean of all humans and chimpanzees found in both datasets before the scaling process, which was also used in the 1st normalization strategy, was added.

In the 3rd normalization strategy (also called 'combat normalization' from now on), "ComBat" function that is found in Bioconductor "sva" package (Leek et al., 2016) designed to remove unknown sources of noise as batch effects is used. The batch effect was two different platforms in this case.

The merged data for Brawand2011 and Khaitovich2005 constructed as following the 2nd strategy was used further in the analysis (see **Section 3**). RNA-seq data of an adult marmoset (Bellott2014) and 2 infant marmosets (Cortez2014) was simply merged with the previously constructed primate data (n=6418).

2.2.2 All Data

All-version of the datasets were used.

First steps of preparing all-data is parallel with the primate-dataset preparation as described in **Section 2.2.1**. Additionally; 2 mice (*Mus musculus*), 2 platypuses (*Ornithorhynchus anatinus*) and 2 opossums (*Monodelphis domestica*) in Brawand2011; cell type data of mouse (*Mus musculus*) in Namekawa2006 and Chalmel2007; microarray data of 2 rat samples (*Rattus norvegicus*) in Chalmel2007; mouse testis development data of Schultz2003 and macaque testis development data of Khaitovich as well as 2 mice RNA-seq data (1 *Mus musculus* and 1 *Mus spicilegus*) were added to the dataset for further analysis (n=4149).

2.3 Analyses of the Datasets

This section includes a series of computational and statistical analyses.

2.3.1 Transcriptome-wide correlations between species

Only the primate-data was used for determining the genes showing differential expression between humans and chimpanzees. Two-sided Student's t-test was used, which also does Welch modification by default assuming that variance is not equal between groups. The obtained p-values showing the significance of the test for each gene are corrected via BH (Benjamini & Hochberg) method in order to reduce Type-1 errors when conducting multiple comparisons. The genes showing differential expression between humans and chimpanzees, meaning having FDR (False Discovery Rate) < 0.10 were used to calculate Pearson correlation between other primates namely gorilla, macaque and marmosets in the analysis to humans and chimpanzees separately. To detect whether there was a difference in the correlation coefficients of each primate in the analysis to humans or chimpanzees, two-sample Kolmogorov-Smirnov test was used.

2.3.2 Cell Type Analysis

The testis cell types spermatogonia and Sertoli cells were grouped as "PRE" (premeiotic and somatic) and spermatocytes and spermatids are grouped as "POST" (meiotic) for further analysis to detect the relative contribution of the gene expression of the cell types to the overall gene expression value of other whole testis samples. The means of the PRE and POST cell types were calculated and a linear model was constructed between these values and remaining part of the data with

the following R-code. Resulting POST/PRE ratio on logarithmic base 2 was done for visualisation purposes.

```
adult_means_lm1 =
   apply( adult_means[XXX,], 2, function(y) {
      lmx = lm ( y ~ PRE + POST )
      lmx$coef })
adult_means_log2 = log2(( POST_lm1 / PRE_lm1 ))
```

lm : Linear Model

$coef$: Coefficient

The ratio of the POST interception point of each species to the linear model to PRE was calculated and transformed into logarithmic scale of base 2. Then, those values were used to test whether single-male species had less POST-meiotic cell type contribution shaping their overall gene expression pattern more than multi-male species via one-sided Wilcoxon rank sum test. Same analysis was done for each species in both primate-dataset and all-dataset separately.

When the POST/PRE ratio received a negative value, log2 version of this ratio assigned as "-4", which was a random number smaller than all the other log2(POST/PRE) results for resolution. This assigned value was only used for visualization of the data, it was not used in any part of the statistical analysis.

2.3.3 Mouse Testis Development Analysis

The genes showing a correlated expression pattern during the development of mouse testis of mouse ages ranging from day 1 to adult were determined via Spearman correlation, having an FDR < 0.10. Then, these genes were used to obtain a regression (Y = ax + b) between gene expression and ages of mice and further to predict hypothetical expression values of missing ages. 43 hypothetical expression values were interpolated via built-in functions of R, namely "predict" and "loess". Further, the correlation between those predicted mouse testis development data and the mean gene expression value of all the other species used in the analysis was calculated with Spearman correlation method.

2.3.4 Macaque Testis Development Analysis

The genes showing a correlated expression pattern during the development of macaque testis of macaque ages ranging from day 16 to day 9518 on a logarithmic scale of 2 were determined via Spearman correlation, having an FDR < 0.10. Then, these genes were used to obtain a regression (Y = ax + b) between gene expression and ages of macaques and further to predict hypothetical expression values of missing ages. 20 hypothetical expression values are interpolated via "predict" and "loess" functions of R. Further, the correlation between those predicted macaque testis development data and the mean gene expression value of all the other species used in the analysis was calculated with Spearman correlation method.

2.3.5 K-means Analysis

The data containing all the common genes found in every dataset were clustered into four groups according to their gene expression levels. Increasing the number of groups resulted in either small clusters having less than 30 genes or multiple clusters having same gene expression patterns. The K-means algorithm which is an unsupervised machine-learning algorithm used for clustering data was used for the analysis. Grouping was repeated 500 times and the clusters that were formed more than the others out of 500 were chosen for further analysis.

2.3.6 Transcription Factor Binding Site Analysis

The genes in each cluster formed in **Section 2.3.5** was searched for transcription factor binding sites at their promoter regions using TRANSFAC database (Reuter, Cheremushkin, Kel, Go & Wingender, 2003). Fisher Exact Test was used for selecting transcription factors enriched in a cluster compared to other clusters. The transcription factors having FDR < 0.5 were detected and reported in **Table 3.2**.

CHAPTER 3

RESULTS

In order to minimize the biases arising from platform differences, all the datasets were pre-processed in a similar way as much as possible. Since the aim of the thesis is to focus on the biological differences and similarities among the species, any technical effect might increase the chance of obtaining unreliable results. We used two approaches to avoid technical biases that could arise in species-specific microarray data and in RNA-seq data.

For microarray data, a probe masking step conducted to ensure that genomic differences between species do not influence the measured expression values. In Khaitovich2005 microarray dataset, a micro-chip designed for humans had been used for the quantification of chimpanzee samples. Macaque testis gene expression in the macaque testis development dataset was also measured using human microarrays. Hence, the probes of each microarray were treated as reads in RNA-seq data and aligned to the genomes of the subject species. Only the perfectly and uniquely aligned probes are used for the further quantification process. For the Khaitovich2005 and macaque testis development datasets, in addition to the human genome, the chimpanzee and macaque genomes were used in the alignment process, respectively. The chip definition files (CDFs), used for quantification of raw microarray data, were then filtered only to contain perfectly and uniquely matched probes. This approach is known to reduce species bias considerably (Khaitovich et al., 2005).

For the RNA-seq data, Gene Transfer Format Files (GTFs) were filtered to contain one-to-one orthologous genes of all the species used in the analysis. Only the perfectly matched unique reads were used as in the microarray pre-processing steps. After the quantification step, genes with no expression value, indicating that they are not expressed were removed from the analysis. The reason for that is microarray data could not detect the gene expression as RNA-sequencing if there are no probes on the micro-chip designed specifically for that genes. There were a

couple of genes with multiple expression values as a side effect of the quantification algorithm of Cufflinks. These genes were also removed from the analysis in order to obtain reliable results instead of taking an average value for them.

Since the datasets containing data for humans and chimpanzees have been produced on two different platforms, RNA-seq and microarray, three different methods were tested in order to minimize the platform differences and to obtain better results to detect species' differences in this meta-analysis.

In the first strategy, the datasets are treated as if they have been produced on the same platform and normalized together. In the second strategy, the mean expression value of humans and chimpanzees are subtracted from each dataset separately and then normalized in order to scale each dataset by using human and chimpanzee expression as a common reference. The third strategy uses the "ComBat" function mentioned in **Subsection 2.2.1** designed to remove batch effects and here treated RNA-seq and microarray platforms as different batches of an experiment.

I then used hierarchical clustering algorithm of R using Euclidean distances of the datasets produced by these three different normalization methods to study possible batch effects and species differences, I found that none of the trees reflected the evolutionary relationships of the four species, which is shown in **Figure 1.1**.

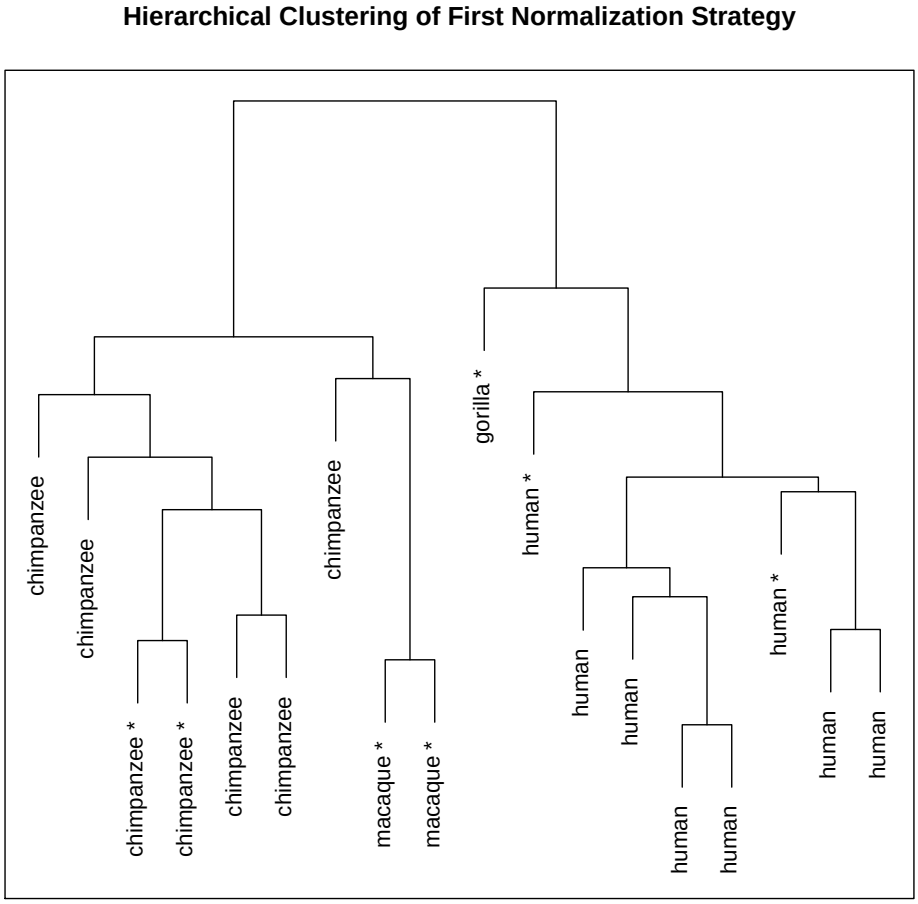

Figure 3.1: The hierarchical clustering of the dataset produced using the first normalization strategy.

The first normalization strategy was suboptimal as it was inefficient in making a full distinction among the species as shown in **Figure 3.1**. One of the chimpanzees from the Khaitovich2005 dataset was grouped with the macaques from the Brawand2011 dataset.

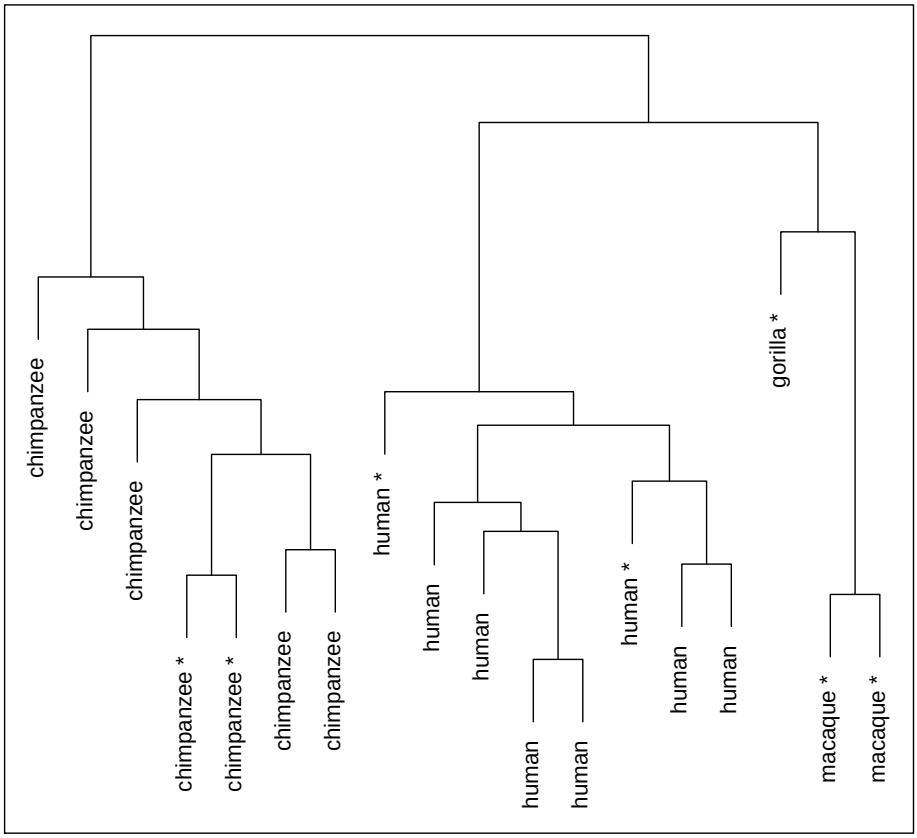

Figure 3.2: The hierarchical clustering of the dataset produced using the second normalization strategy.

The second normalization strategy was able to group species as non-overlapping (monophyletic) groups as shown in **Figure 3.2**. Moreover, having two different platforms had less effect on the topology of the tree, as the two human samples of the Brawand2011 dataset were closer to other human samples in the Khaitovich2005 dataset than to each other.

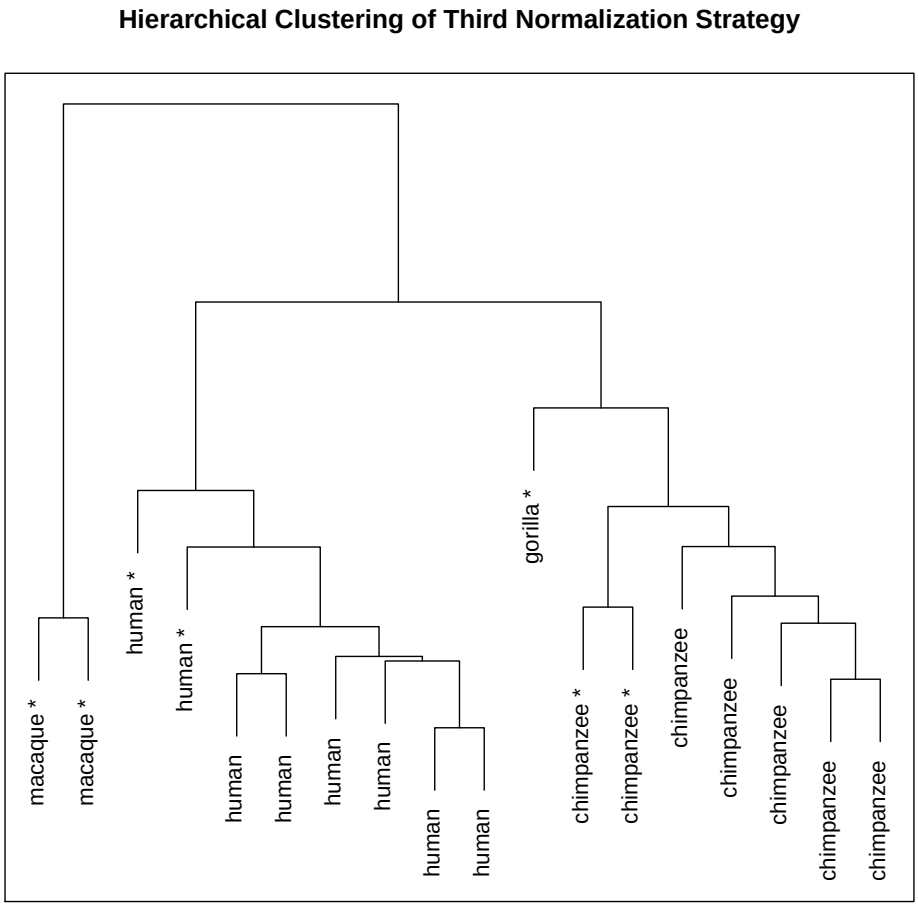

Figure 3.3: The hierarchical clustering of the dataset produced using the third normalization strategy.

The third normalization strategy (ComBat) was successful in grouping species separately, though it was less successful in removing the batch effect, such that the species coming from the same platform grouped together (**Figure 3.3**).

I thus decided to use the second normalization strategy in downstream analysis.

3.1 Transcriptome-wide correlations between species

As mentioned in Introduction, testis size evolves convergently among primates, and testis transcriptomes may also show a similar pattern. Here I tested this latter observation on convergent evolution by using all available primate testis transcriptome datasets listed in **Tables 2.1 and 2.3**. Specifically, I used human and chimpanzee as reference for single-male and multi-male species, respectively, and compared all other primate species with each of the two hominids, with respect to their transcriptome profiles. My hypothesis was that species with similar mating types would show similar expression profiles.

For this, I first identified genes having differential expression between humans and chimpanzees. Out of 6418 genes that are common for all the primates in the analysis including marmosets, more than 50% (3758 genes) showed significant differential expression between humans and chimpanzees (FDR<0.1).

Using these genes showing differential expression between humans and chimpanzees, I then calculated The Pearson correlation coefficient between these 3758 genes' expression profiles in different non-hominid primates and in humans or in chimpanzees (**Figure 3.4**). Here I calculated the mean expression level per gene whenever a non-hominid species had multiple individuals (e.g. macaque). In constrast, I calculated these correlations for each hominid separately.

The results suggest that testis transcriptome profiles partly reflect mating type in gorilla and macaque. It also suggests that human testis transcriptomes are more similar to that of infant marmosets than to that of mature marmosets.

When these inter-species correlations using differentially expressed genes are considered, gorilla showed significantly higher positive correlation with humans rather than with chimpanzees as stated in **Table 3.1**. Similarly, the correlation between infant marmoset and humans was highly significant though adult marmoset was equally similar to both humans and chimpanzees.

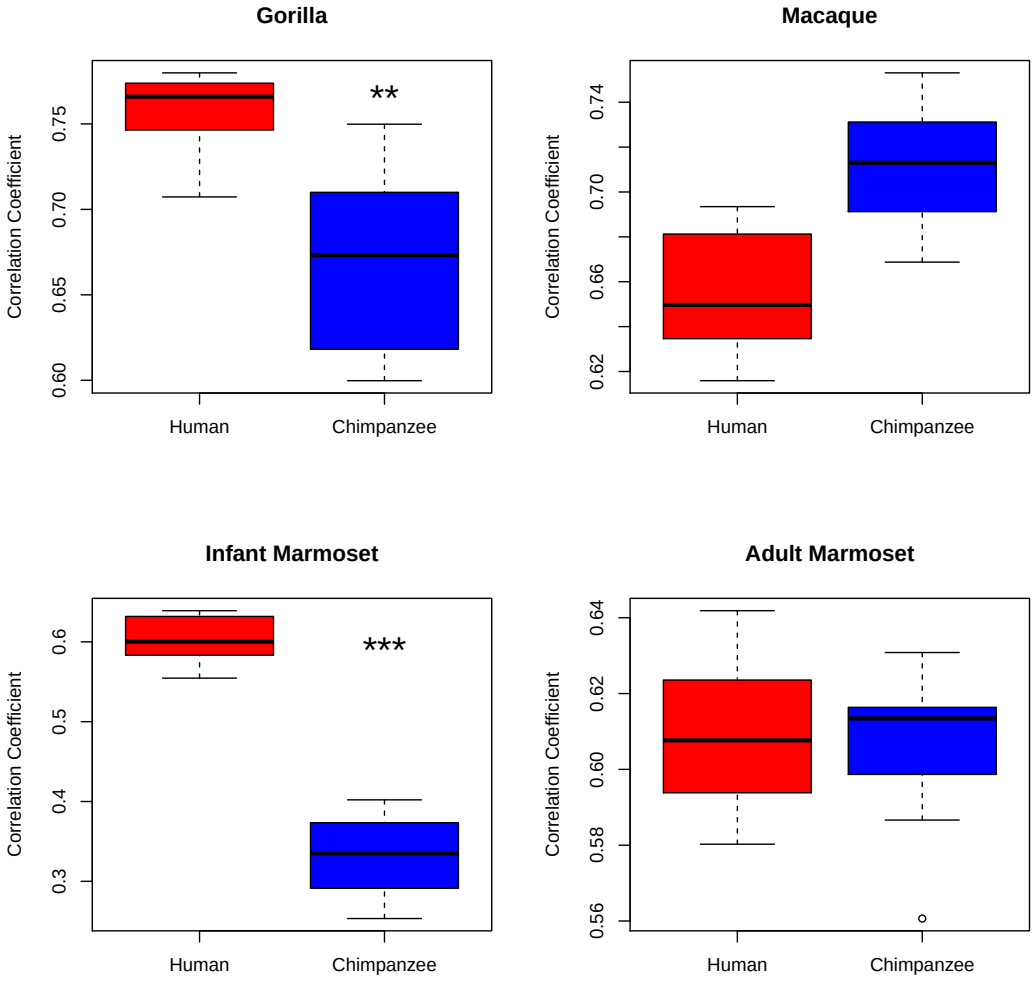

Figure 3.4: Boxplot of Pearson correlation coefficients between transcriptome-wide expression levels of gorilla (n=1), macaque (n=2), infant marmoset (n=1) and adult marmoset (n=1) to humans (n=8) and chimpanzees (n=7). For non-hominid primates where we had more than one individual, we calculated the mean expression for each individual.(* indicates significance based on Ks-test)

Table 3.1: The mean of Pearson correlation coefficients of gorilla, macaque, infant marmoset and adult marmoset to humans and chimpanzees in testis transcriptomes. The last column indicates Kolmogorov-Smirnov test p-values for difference between mean correlation coefficients of humans and chimpanzees.

	Mean cor. coef. of humans	**Mean cor. coef. of chimps**	**KS test p-value**
Gorilla	0.7574423	0.6684533	0.02424**
Macaque	0.6550545	0.711379	0.05594
Adult Marmoset	0.6090281	0.6050009	0.9478
Infant Marmoset	0.6030628	0.3314291	0.0003108***

3.2 Cell type analysis

When the genes showing differential expression between humans and chimpanzees are considered, there is a significant difference in testis transcriptome profiles between species as shown in **Section 3.1**. I here test if this difference is originated from the difference in the proportion of the cell types found in testis. For this purpose, the mouse testis cell types spermatogonia and Sertoli cells were grouped as "PRE" (pre-meiotic and somatic) and spermatocytes and spermatids are grouped as "POST" (meiotic and post-meiotic). Further, these cell types were used to detect the relative contribution of the gene expression of the cell types to the overall gene expression value of other whole testis samples.

POST/PRE ratios of chimpanzees and macaques (multi-male mating behaviour) is significantly higher than those of humans and gorilla (single-male mating behavior) (p-value=2 e-05, one-sided Wilcoxon rank sum test) visualised in **Figure 3.5**.

Among primates excluding marmosets, transcriptomes of chimpanzees had a higher contribution from genes expressed in POST cell types, whereas the transcriptomes of humans had a higher contribution from genes expressed in PRE cell types. The POST/PRE ratios of macaques were in the range of chimpanzees' and the POST/PRE ratio of gorilla in the range of humans'.

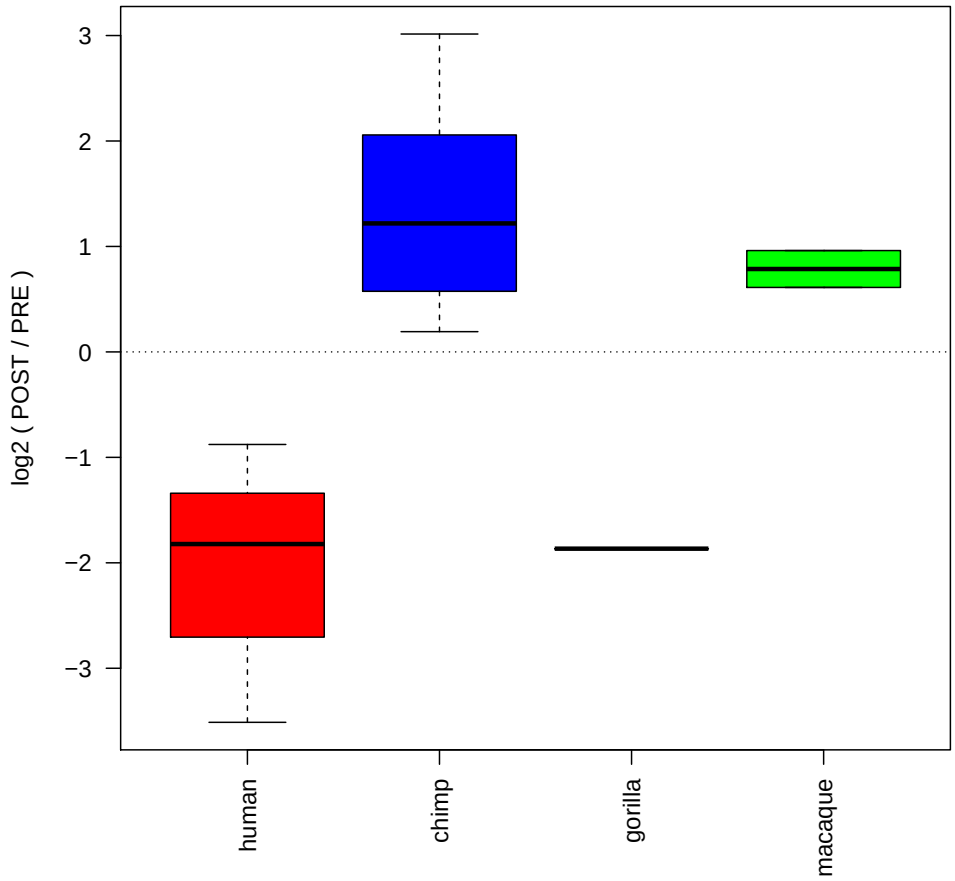

Figure 3.5: Boxplot of cell type ratios of human, chimpanzee, gorilla and macaque.

When all the data (3788 genes) is used for the same analysis (**Figure 3.6**), the transcriptomes with higher contribution from genes expressed in POST cell types were the ones of chimpanzees, macaques, mice of the species *Mus musculus*, adult marmoset, rats and mouse of the species *Mus spicilegus*. The POST/PRE ratio of *Mus spicilegus* was in the range of *Mus musculus*. Remaining data involving humans, gorilla, opossums, platypuses and infant marmoset had a higher contribution from genes expressed more in PRE cell types.

The results obtained is this section is consistent with the idea that the testis transcriptome differences detected in **Section 3.1** might be the result of the relative

contribution of cell types found in the testis.

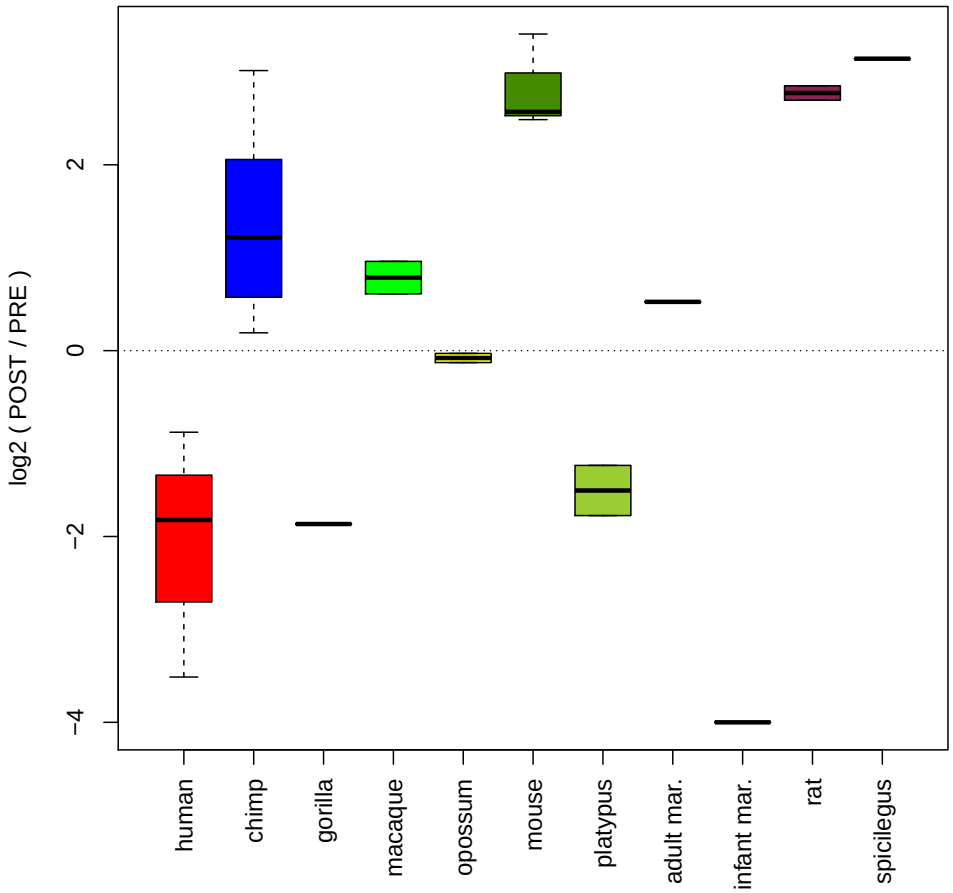

Figure 3.6: Boxplot of cell type ratios of all species used in the analyses.

3.3 Mouse testis development analysis

As shown in the **Section 3.1**, for the genes showing differential expression between humans and chimpanzees, gorilla and infant marmoset showed significantly higher positive correlation to humans. Moreover, the analyses in **Section 3.2** shows that there is a higher contribution of PRE cell types to the testis gene expression profiles of humans, gorilla and adult marmoset. I here test whether there is an effect of the genes changing expression during development on this significantly high positive correlation. For this purpose, I have used mouse testis development data.

Fifteen mouse samples with ages ranging from newborn to adult (taken as 42 days in the analysis) were first used to detect genes differentially expressed during mouse testis development. Out of 3788 genes that are common in all the datasets, 1994 of them (FDR<0.1) showed significantly high positive correlation between gene expression values and the ages of the mice. For these genes, correlation coefficients calculated between the mouse testis development data and the mean expression values of each species as well as PRE and POST cell types are plotted against the mouse ages as shown in **Figure 3.7**.

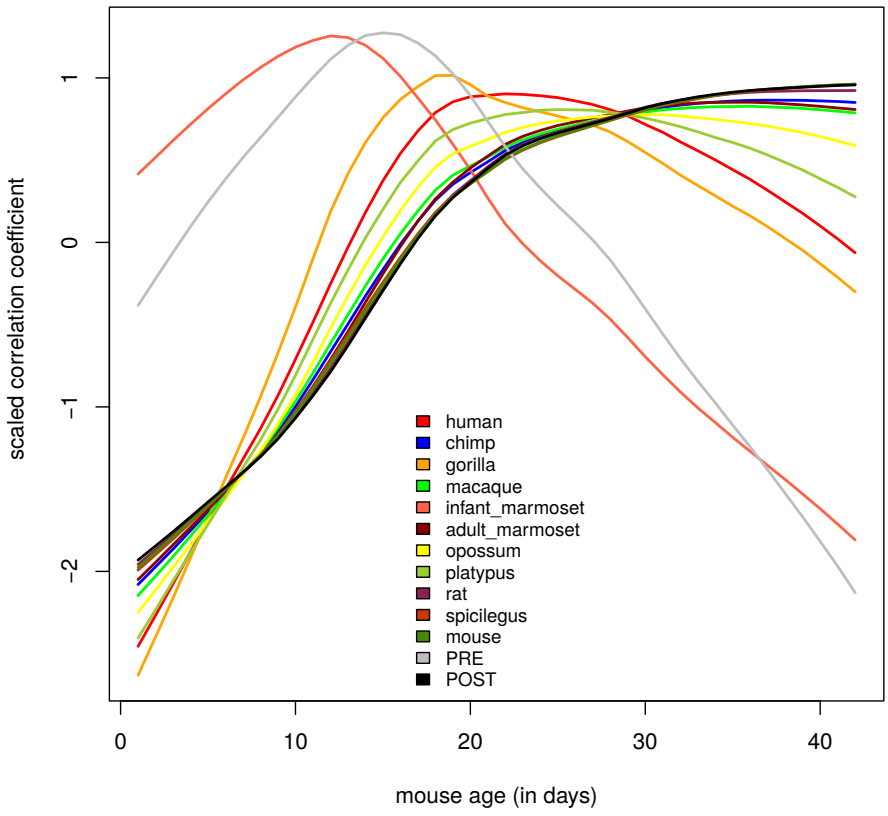

Figure 3.7: Plot of correlation coefficients between different species' whole testis transcriptome profiles and mouse testis transcriptome profiles at different ages, against mouse age (n=1994 genes). Each curve was scaled to mean=0, sd=1 independently.

Considering the fact that 42 days old mouse represents adult mice, most of the species used in the analyses (including chimpanzee, macaque and rat) showed highest correlation with adult mice in their testis expression profiles. On the other hand, the transcriptome data of the infant marmoset as well as PRE cell types were the most distinct ones from the transcriptome data of adult mice. Humans and the gorilla individual also were most similar to young mice since the correlation coefficient peaks were around the 20 days old mice. The other two species that showed peak correlation with mice younger than 42 days were the evolutionarily more distant species: platypus and opossum.

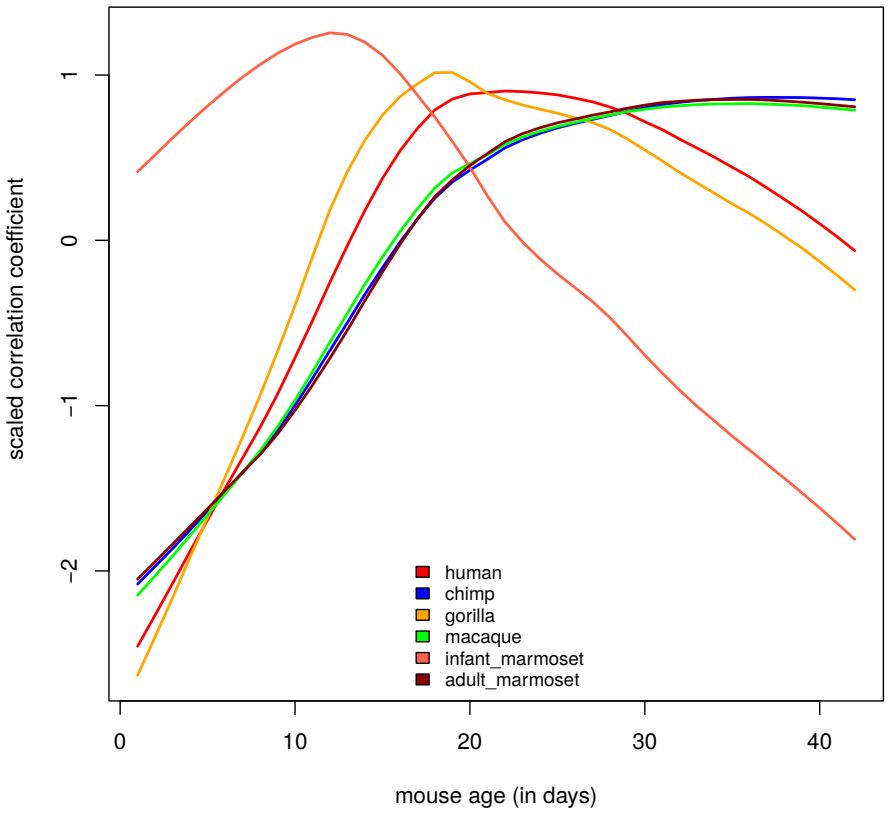

Figure 3.8: Plot of correlation coefficients between primate species' whole testis transcriptome profiles and mouse testis transcriptome profiles at different ages, against mouse age (n=1994 genes). Each curve was scaled to mean=0, sd=1 independently.

The results obtained from the primate data shows that the transcriptome data of infant marmoset, gorilla and human were more similar to immature mice in the ascending order of mouse age (**Figure 3.8**). In other words, among the primates, infant marmoset showed the highest correlation to immature mice of 10 days of age, and humans to immature mice of 20 days of age. The remaining primates, namely chimpanzee, macaque and adult marmoset showed highly similar correlation values among each other throughout the comparisons between the different ages of mice and showed highest correlation to the adult mouse rather than the immature ones.

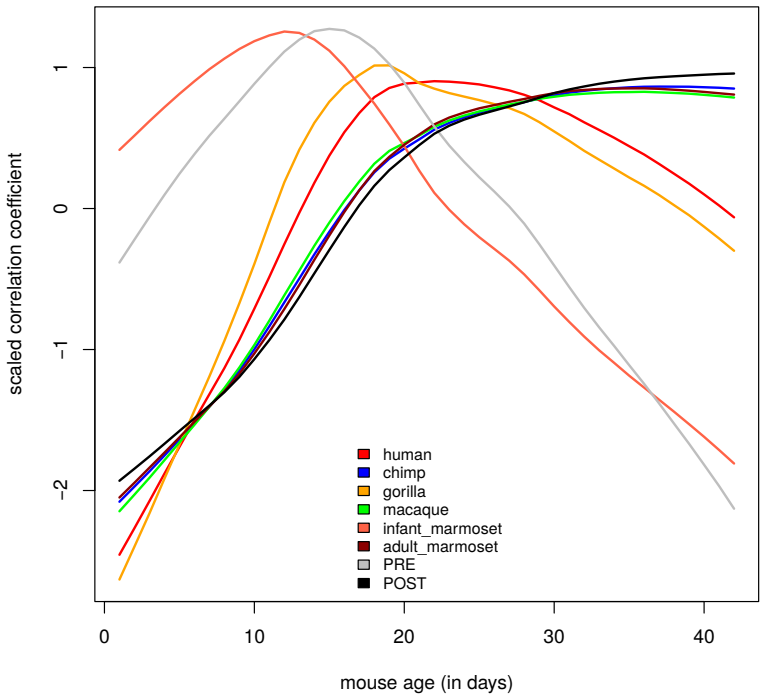

Figure 3.9: Plot of correlation coefficients between primate species' whole testis transcriptome as well as cell types' transcriptome profiles and mouse testis transcriptome profiles at different ages, against mouse age (n=1994 genes). Each curve was scaled to mean=0, sd=1 independently.

When the PRE and POST cell types were added to the analysis together with the primate data (**Figure 3.9**), the correlation coefficient peak of PRE cell types' transcriptome was in between the peaks of infant marmoset and the two single-male

species namely human and gorilla. The correlation coefficient values of POST cell types was almost inseparable from those of chimpanzee, macaque and adult marmoset.

The results of the analysis in this section suggest that the difference in the levels of testis gene expression between single- and multi-male species might be the result of the difference in the relative contribution of the cell types present in the testis.

3.4 Macaque testis development analysis

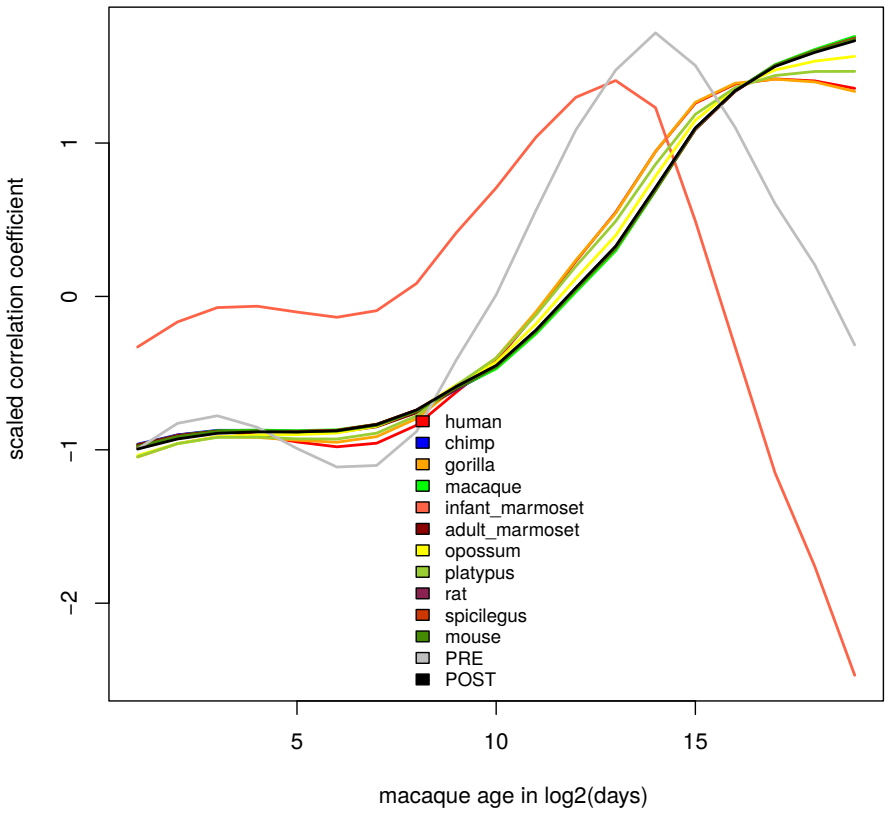

Figure 3.10: Plot of correlation coefficients between different species' whole testis transcriptome profiles and macaque testis transcriptome profiles at different ages, against macaque age (n=2030 genes). Each curve was scaled to mean=0, sd=1 independently.

To confirm the results we obtained with mouse development, I repeated the analysis with a more closely related species. n=12 macaque samples with ages ranging from 16 days to 9518 days were first used to detect genes differentially expressed during macaque testis development. Out of 3788 genes that are common in all the datasets, 2030 of them (FDR<0.1) showed correlation between gene expression values and the ages of the macaques. For these genes, correlation coefficients of the mean expression values of each species as well as PRE and POST cell types were plotted against the macaque testis development data as shown in **Figure 3.10**.

The transcriptomes of most of the species used in the analyses as well as POST cell types showed highest correlation with adult macaque and produced similar correlation values when compared to the mean gene expression values of macaques of varying ages. On the other hand, the transcriptome data of infant marmoset as well as PRE cell types were the most distinct ones from the transcriptome data of adult macaques, relative to other datasets. Humans and gorilla were also more similar to immature macaques since the correlation coefficient peaks were detected at younger ages of macaques. The other two species that showed peaks before the peak of the POST cell types were yet again the evolutionarily more distant species: platypus and opossum (**Figure 3.10**).

The results obtained from the primate data (**Figure 3.11**) shows that the transcriptome data of infant marmoset yielded more similar correlation coefficients to the most immature macaque compared to other primates. The highly similar correlation coefficients of human and gorilla gave a peak around younger macaques. The remaining primates, namely chimpanzee, macaque and adult marmoset showed highly similar correlation values with each other throughout the comparisons between the different ages of macaques and the peak could not be detected in the analysis, suggesting a value higher than the oldest macaque used for the comparisons.

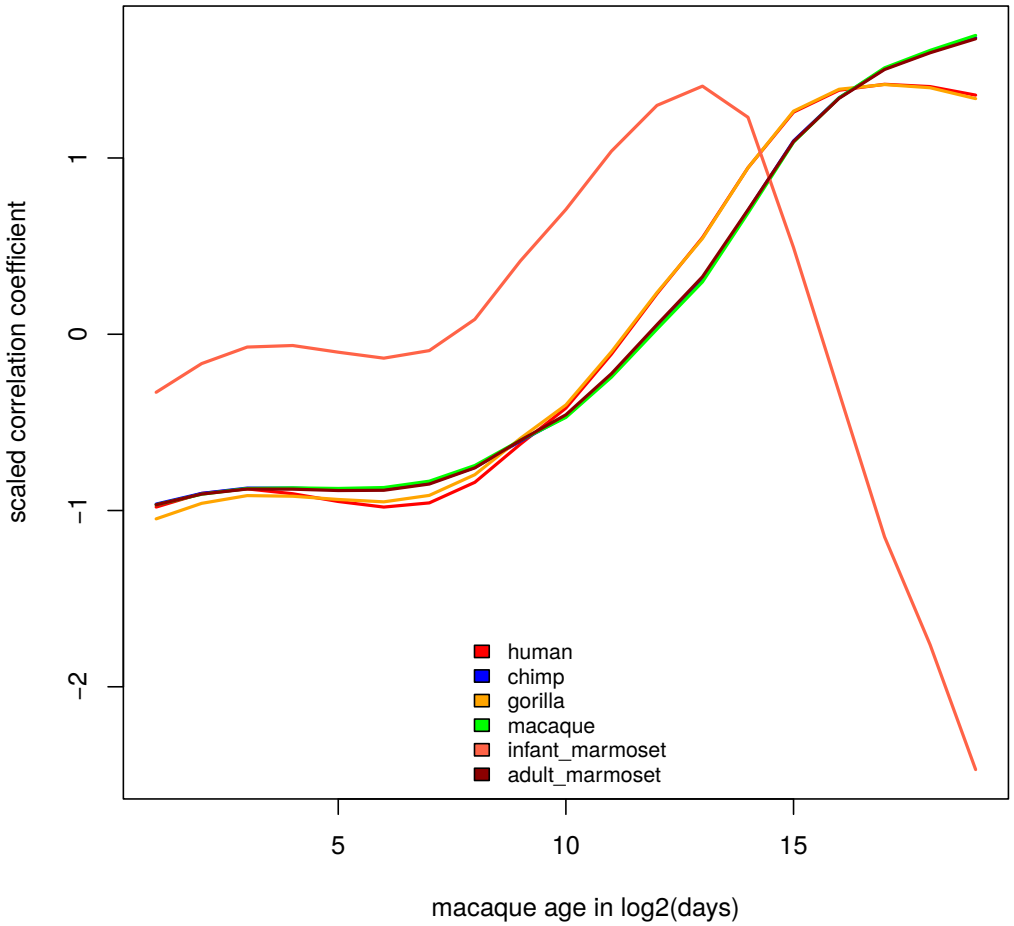

Figure 3.11: Plot of correlation coefficients between primate species' whole testis transcriptome profiles and macaque testis transcriptome profiles at different ages, against macaque age (n=2030 genes). Each curve was scaled to mean=0, sd=1 independently.

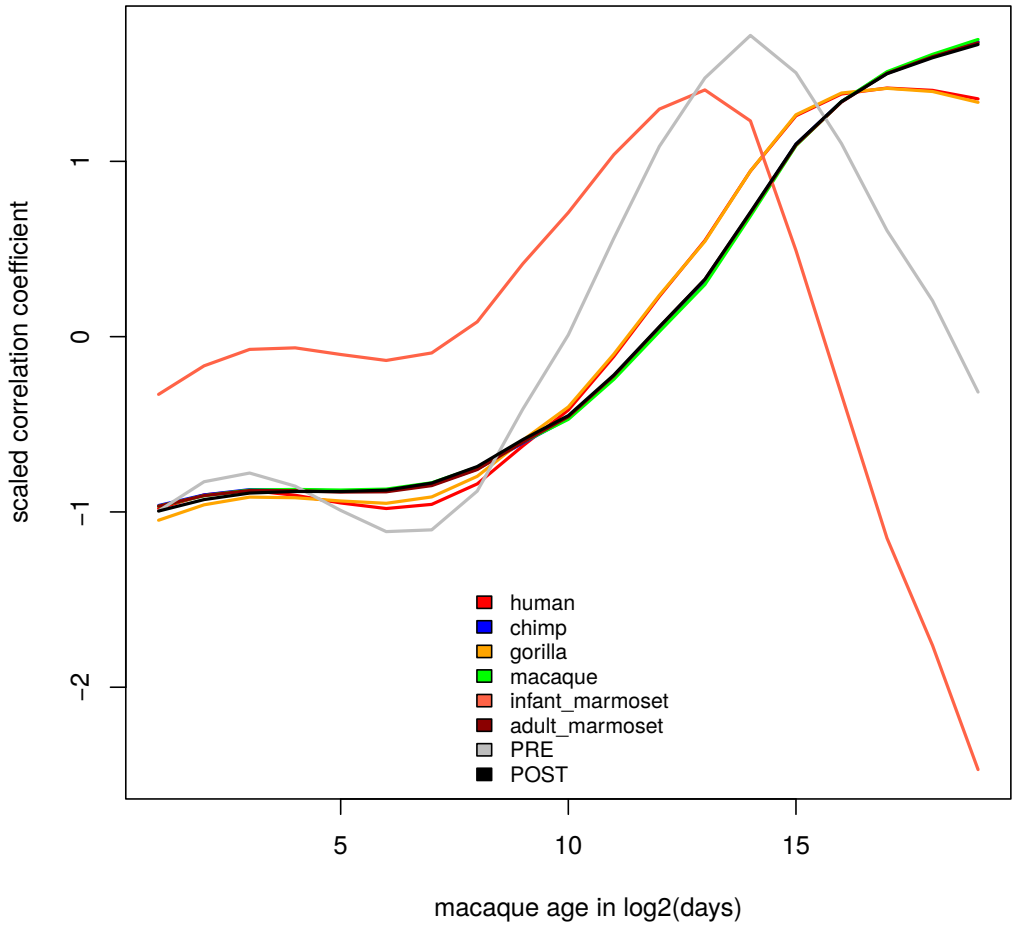

Figure 3.12: Plot of correlation coefficients between primate species' whole testis transcriptome as well as cell types' transcriptome profiles and macaque testis transcriptome profiles at different ages, against macaque age (n=2030 genes). Each curve was scaled to mean=0, sd=1 independently.

The results obtained from the primate data (**Figure 3.12**) shows that the transcriptome data of infant marmoset, gorilla and human were more similar to immature macaque in the ascending order of macaque age in terms of correlation coefficients calculated between the gene expression values of the primates and transcriptome of macaques with different ages. In other words, among the primates, infant marmoset showed the highest correlation to the youngest macaque; human and gorilla having correlation coefficient values very similar to each other to a less immature macaque though not mature. The remaining primates, namely chimpanzee, macaque and adult marmoset showed highly similar correlation values with each other throughout the comparisons between the different ages of macaques and their highest correlation coefficients could not be detected in the analysis suggesting that a higher value would have been obtained if an even older macaque had been used in the comparisons. Their correlation coefficients were increasing with the increasing macaque age.

The results obtained in this section is consistent with the results in the **Section 3.3**. A similar trend, though not as clear as in the mouse testis development analysis can also be detected for macaque testis development data.

3.5 K-means Analysis

The analysis conducted in the previous sections showed the presence of the genes showing differential expression profiles between single- and multi-male species (**Section 3.1**), having different amounts of contribution from PRE and POST cell types trancriptome profiles (**Section 3.2**) and different levels of correlation with developing mouse and macaque testis gene expression profiles (**Sections 3.3 and 3.4**). I decided to group the genes present in all the datasets used in the analysis according to their gene expression profiles. Each gene was scaled for each dataset separately and a total of 3788 genes were grouped into four clusters according to their expression profiles via the k-means algorithm. The resulting four groups contained 1186, 980, 838 and 784 genes and the expression profiles of clusters are shown in **Figure 3.13**.

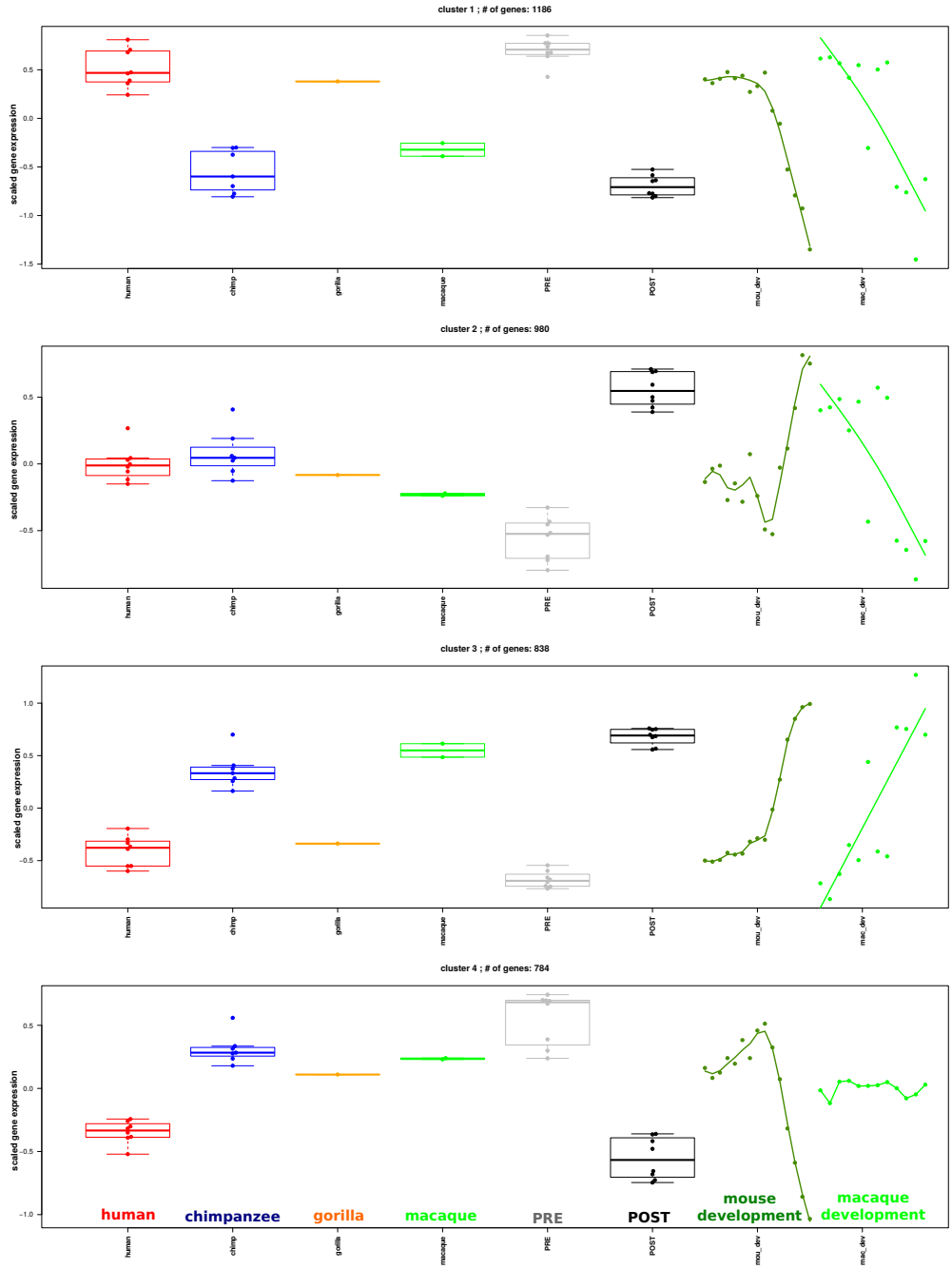

Figure 3.13: Four clusters formed according to the expression profiles of common genes between humans, chimpanzees, gorilla and macaques. The boxplots are shown in order of human, chimpanzee, gorilla, macaque, PRE and POST cell types. Dark green expression line represents mouse development and light green expression line represents macaque development, ordered according to age.

The first cluster containing 1186 genes shows decreasing gene expression in developing mouse and macaque testes. Humans and chimpanzees are completely separated. Gene expression of levels gorilla and PRE cell types overlap with those of humans and they all show similarity to both immature mice and immature macaques having higher gene expression. On the other hand, gene expression of macaques and POST cell types overlaps with the lower levels of gene expression in chimpanzees. These latter three groups' expression levels show similarity to those of mature mice and macaques.

In the second cluster containing 980 genes, gene expression profiles of the four primate species do not separate clearly from one another, whereas the genes present in this cluster show low relatively expression in PRE cell types and relatively high expression in POST cell types. In the mouse development data, the expression shows a steady expression in young mice and later rapidly increases with age towards maturation. The macaque development data shows a continuous decrease.

The third cluster containing 838 genes shows increasing gene expression in developing mouse and macaque testes. Humans and chimpanzees are completely separated. Relatively low gene expression levels of gorilla and human are similar to relatively low expression in PRE cell types and they all show similarity to both mice and macaques of young age having lower gene expression. On the other hand, relatively high gene expression levels of macaques and chimpanzees overlap, similar to relatively high expression in POST cell types. These three groups show similarity to mature mice and mature macaques having relatively high gene expression.

In the forth cluster containing 784 genes, the gene expression levels of humans and chimpanzees separate from one another. The expression levels of gorilla and macaques are similar to that of chimpanzees, having higher levels than humans. On the other hand, the expression of PRE cell types are also higher than in POST cell types, reminiscent of the human profiles. The macaque testes development data shows a steady level of gene expression. The mouse testes development data shows an increasing gene expression in young mice and decreases dramatically after a point.

Overall, 2024 genes (53% of the common genes found in all the datasets) found in

Cluster1 and Cluster3 show an expression profile that could be an explanation for the difference in the transcriptome profiles of single- and multi-male species.

3.6 Transcription Factor Binding Site Analysis

Since the expression of many genes can be controlled by same transcription factors, in this section, I search for common transcription factors enriched in the clusters found in the previous section, **Section 2.3.5**. For this purpose, I have used TRANSFAC database entries and used Fisher's exact test to compare transcription factors regulating the genes in each cluster against the transcription factors enriched in the remaining clusters.

According to the TRANSFAC database, there are a few transcription factors that are enriched in Cluster1, as listed in **Table 3.2**.

Table 3.2: Enriched Transcription Factors in Clusters

Cluster #	Transcription Factor	one-sided Fisher's Exact Test (FDR)
1	**NKX25_02**	0.4317
	CEBPGAMMA_Q6	0.4317
	FOX_Q2	0.1939
	RFX1_02	0.4317
2	**SZF11_01**	0.2319
	CACD_01	0.3067
	PAX4_03	0.1367
	PAX4_01	0.2982
	PAX6_01	0.4269
	MYOGNF1_01	0.3067
	ZF5_B	0.0325
	HOX13_01	0.3870
	CP2_02	0.3247
	ATF6_01	0.3860
	DR1_Q3	0.2319
	GATA4_Q3	0.4212
	ZIC2_01	0.3870
	SP1_Q2_01	0.3247
	LRF_Q2	0.2442
	SRY_02	0.4212

ER_Q6	0.3067
SPZ1_01	0.1367
PAX5_01	0.3312
CEBP_Q3	0.2319
AP2_Q6_01	0.3870
EGR1_01	0.1367
CREB_02	0.0578
AP1_Q4_01	0.3009
GRE_C	0.4336
LEF1TCF1_Q4	0.4375
ETF_Q6	0.0762
PAX3_B	0.4145
SREBP1_Q6	0.4336
CREB_Q4_01	0.3720
P300_01	0.4432
AP2ALPHA_01	0.3351
ETS_Q6	0.4715
NKX25_Q5	0.4774
CEBPA_01	0.4336
MAZ_Q6	0.1367
WT1_Q6	0.2507
AP2_Q6	0.0762
GEN_INI3_B	0.4317
PAX5_02	0.3870
SREBP_Q3	0.4145
DBP_Q6	0.3009
AP1_Q2_01	0.1367
TAXCREB_01	0.3093
TAXCREB_02	0.1554
CRX_Q4	0.4382
AR_Q2	0.4212
HIC1_02	0.3698
SP3_Q3	0.4317
AHRHIF_Q6	0.4212
MINI19_B	0.0791
HIF1_Q3	0.4336
KROX_Q6	0.3009
SF1_Q6	0.4336
PPARA_02	0.3312

	COUP_DR1_Q6	0.3698
	VDR_Q3	0.4212
	STAT1_01	0.2696
	HNF3B_01	0.3870
	AHR_Q5	0.1367
	AHRARNT_01	0.2583
	PEBP_Q6	0.3093
	PPAR_DR1_Q2	0.4336
	IRF2_01	0.0578
	AP4_01	0.4852
	MTF1_Q4	0.3698
	SRF_C	0.3067
3		
4	**CETS1P54_03**	0.4350
	OCT4_02	0.0783
	STAT_Q6	0.4580
	PAX6_01	0.0928
	XVENT1_01	0.4350
	TST1_01	0.0007
	RFX_Q6	0.0682
	CETS1P54_01	0.4350
	NKX25_02	0.0598
	SRY_02	0.0598
	PAX4_04	0.4350
	NCX_01	0.0598
	VJUN_01	0.4610
	CDPCR1_01	0.3210
	HMGIY_Q6	0.4350
	OCT_Q6	0.4350
	CHOP_01	0.4350
	FOXJ2_02	0.1459
	OCT1_02	0.0883
	PLZF_02	0.3158
	POU3F2_02	0.3210
	BRCA_01	0.1202
	SOX9_B1	0.4350
	POU3F2_01	0.4350
	YY1_Q6_02	0.0009

	PAX4_02	0.0089
	TBP_Q6	0.4350
	CART1_01	0.1757
	TEL2_Q6	0.4350
	HNF3ALPHA_Q6	0.4350
	STAF_02	0.2514
	CDXA_02	0.0928
	DEAF1_02	0.4350
	KROX_Q6	0.4490
	HOXA7_01	0.3621
	YY1_Q6	0.0113
	NFY_01	0.4350
	POU6F1_01	0.2143
	GFI1_Q6	0.4350
	ZNF219_01	0.3158

CHAPTER 4

DISCUSSION

This thesis includes 8 different mammalian testis gene expression datasets produced on two different platforms; to be more specific, 3 RNA-seq and 5 microarray. A total of 10 species is investigated in this study and compared. Two of the microarray datasets consist of testis development data of mice and macaques having different ages. My aim was to test a possible relationship between the testis sizes of these species and the genes expressed in their testicles.

My analysis included two multi-species primate datasets that were the focus of my analysis. One of these was a microarray dataset, Khaitovich2005, comprising 6 humans and 5 chimpanzees, and the other an RNA-seq dataset, Brawand2011, comprising 2 humans, 2 chimpanzees (1 common chimpanzee + 1 bonobo), 2 macaques and 1 gorilla. These two datasets were combined using three different strategies. The reason of combining gene expression data from these two different datasets was that they both contain expression data for humans and chimpanzees, which are considered single- and multi-male species, respectively.

Through the above-mentioned pre-processing and combining steps, I produced two versions of multi-species datasets: one includes the one-to-one orthologous genes for all the species used in the analysis (2.2), and the other one includes one-to-one orthologous genes for primates only. Since there are distant species used in the analysis such as opossum, the orthologous genes number drops to 3788 from 6418 when all the species is used instead of only using primate orthologs. The two dataset versions thus served different functions: The large one includes more species and thus provides a wider phylogenetic perspective in my analysis, while the smaller primate only dataset contains more genes and thus has higher power to identify expression divergence patterns.

3758 genes out of 6418 show significant differences between humans and chimpanzees when these species were used as representatives of single-male and multi-

male mating types respectively.

First, I tested the hypothesis that mating type influences mammalian testis expression levels, using humans and chimpanzees as representatives of single-male and multi-male mating types respectively. Although this was previously proposed by (Brawand et al., 2011), this was the first systematic test of this notion.

Among primates, gorilla and infant marmoset showed significant correlation to humans rather than chimpanzees. Only macaque showed significantly higher correlation with chimpanzee. This is consistent with the mating type hypothesis and that expression profiles have evolved convergently in primates. The mating type of marmosets are less clear, but they are generally considered to be single-male (Digby, 1999), although I could not find a significant correlation between testis transcriptome of adult marmoset to either the transcriptomes of single-male species or multi-male species. Intriguingly, the finding that the infant marmoset has human-like expression might suggest that single-male species, as humans, have expression profiles that are "immature" relative to those of multi-male species.

The observed convergent expression patterns might be the result of changes in the proportion of the cell types found in testes.

As shown in the **Figure 3.5**, chimpanzee and macaque testis transcriptomes are estimated to have a higher contribution from genes expressed in POST cell types (higher regression coefficients), that is germ-line cells, whereas humans and gorilla testis transcriptomes are estimated to have a higher contribution from genes expressed in PRE cell types, including somatic cells. The cell type proportions are changing not only from species to species, but also in the same species as development proceeds. This can be seen in **Figure 3.6**: The testis transcriptome of infant marmoset is dominated with the genes expressed in PRE cell types, as in single-male species; on the other hand, the genes expressed in POST cell types have more influence on the transcriptome of adult marmoset, as in multi-male species.

This leads to the idea that single-male species' testes may represent an immature, neotenic state in testis development, relative to multi-male species. The transcriptome profiles of testis changing with the developmental stages of mouse and

macaque and the affinity of different adult mammals' testis profiles to these developmental stages are investigated in **Section 3.3** and **Section 3.4**. The transcriptome profiles of infant marmoset, PRE cell types, gorilla and humans showed higher correlation to the testis transcriptome of mice at early developmental stages, shown in **Figure 3.9**. A similar trend is also detected for the macaque testis development data, though not explicit as in the case of mouse development data (**Figure 3.12**). The testis gene expression profiles of other species in these analyses, namely; chimpanzee, macaque, and adult marmoset, show same high level of correlation as the transcriptome profile of POST cell types to the gene expression profiles of both mature mouse and macaque testes.

When the shared genes between human, chimpanzee, gorilla, macaque, PRE and POST cell types together with the developmental data of mouse and macaque were grouped into four clusters according to their mean gene expression levels, two of the clusters (Cluster1 and Cluster3, in total comprising more than 50% of the genes involved) separated human, chimpanzee, gorilla and macaque into two groups consistent with their mating behaviours.

The genes represented in Cluster1 showed decreasing gene expression with age in development datasets of mouse and macaque testes. Human, gorilla and PRE cell type samples show higher gene expression similar to mice and macaques of young ages. On the other hand, genes represented in the Cluster3 have the reversed expression levels in Cluster1. This again supports the idea that convergent evolution of whole testis transcriptome profiles affects a large proportion of the transcriptome.

Meanwhile, the other two gene clusters did not follow an obvious mating type-related pattern. These genes' expression profiles could possibly represent other cellular processes such as regulation of metabolic processes, response to an external stimuli, RNA transport, DNA repair, organelle organization, cell-cell signalling, etc.

Since trancription factors are capable of controlling expression of a group of genes, it is plausible to infer that the genes in these two clusters are controlled by the same transcription factors. Transcription factor binding site analysis results (see **Section**

3.6) reveal a couple of transcription factors that show a weak trend for enrichment in Cluster1 (with FDR<0.5), while none are found at the same cut-off in Cluster3 (see **Table 3.2**).

The transcription factors that showed an enrichment trend (at a very relaxed cut-off for multiple testing correction) included *NKX25*, *CEBPGAMMA*, along with another member of FOX transcription factor family. These have been shown to have common functions in the immune system and in cancer development. Their functions are reported to be related with regulation of cell division and DNA replication (Hu & Gallo, 2010). Moreover, the other transcription factor enriched in this cluster, *RFX1* was shown to be differentially expressed during spermatogenesis (Kistler, Horvath, Dasgupta & Kistler, 2009), and was shown to have important functions related to spermatogenesis such as maintaining testis cord integrity in mouse embryos (Wang et al., 2016).

4.1 Limitations and Possible Improvements

- This thesis includes 8 different species, some closely related and some distantly related. Adding more datasets including other species would improve the results and provide more information about the relationship between testis transcriptome profiles and mating behaviour.

- This thesis includes the transcriptome of the testis only. Adding transcriptome profiles of other tissues such as brain would show that the differences that we detect are indeed correlated with mating type and do not reflect ubiquitous differences among the species used in the analysis.

- While calculating the similarity levels of gene expression, using the phylogenetic distance of the species used would give more reliable results.

- The functions of the genes showing differential expression between humans and chimpanzees as well as changing gradually in the testis development data of mouse and macaque could be investigated more deeply both computationally and experimentally.

- The genes in Cluster1 and Cluster3 could be pooled to search for common transcription factors controlling them since they show reversed levels of gene

expression, whereas here I tested them separately.

- The featured transcription factors detected in **Section 3.6** or their target genes could be knocked-out in transgenic mouse to see their effects on the size of the testis.
- In addition, Gene Ontology analysis could be conducted to study the common functions of the genes in the clusters.

CHAPTER 5

CONCLUSION

It is a highly accepted theory that there is a relationship between testis size and mating behaviour. To be more specific, when females evolve polyandrous type of mating, males having larger testicles capable of producing higher volumes of sperm and more efficient sperm gain advantage to pass their genes to the next generation. However, the underlying developmental and genetic mechanisms explaining this observation is still largely obscure. Here I analysed transcriptomic data of various species with known testis sizes and mating behaviours, in order to gain insight about the molecular basis of this phenomenon, which is also the main aim of the research elaborated in this thesis. The results suggest that testis gene expression profiles are possible candidates for explaining the relationship between testis size and mating behaviour, as they also evolve convergently. Furthermore, I predicted that testes of single-male mating species contain more pre-meiotic cell types, and are in a relatively immature, neotenic state, compared to the testes of multi-male mating species. These changes most likely explain the bulk of the observed convergent transcriptome evolution patterns.

Based on my results, further studies using comparative transcriptomics as in this thesis can be used for:

- predicting the mating behaviour of a species hard to track in nature in a more reliable way than to study anatomical traits such as testis size, and
- detecting candidate genes responsible for disorders affecting regular development of the testis.

Overall, expanding transcriptome data allows us to investigate similar or distinct biological characteristics on various levels.

REFERENCES

Anderson, W. R., Hicks, J. A. & Holmes, S. A. V. (2002). The testis: What did he witness? *BJU International*, *89*(9), 910–911. doi: 10.1046/j.1464-410X.2002.02783.x

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E., Kang, H., Korbel, J. O., ... Abecasis, G. R. (2016). HHS Public Access. *Nature*, *526*(7571), 68–74. doi: 10.1038/nature15393.A

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T.-J., ... Page, D. C. (2014, apr). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature*, *508*(7497), 494–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24759411 doi: 10.1038/nature13206

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., ... Miller, W. (2004). Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research*, *14*, 708–715. doi: 10.1101/gr.1933104.6

Bolger, A. M., Lohse, M. & Usadel, B. (2014). Original paper. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Bolstad, B. (2016). *preprocessCore: A collection of pre-processing functions.* Retrieved from https://github.com/bmbolstad/preprocessCore

Bozek, K., Wei, Y., Yan, Z., Liu, X., Xiong, J., Sugimoto, M., ... Khaitovich, P. (2014, may). Exceptional Evolutionary Divergence of Human Muscle and Brain Metabolomes Parallels Human Cognitive and Physical Uniqueness. *PLoS Biology*, *12*(5), e1001871. Retrieved from http://dx.plos.org/10.1371/journal.pbio.1001871 doi: 10.1371/journal.pbio.1001871

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., ... Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. [SupMat]. *Nature*, *478*(7369), 343–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22012392 doi: 10.1038/nature10532

Britten, R. J. (2002). Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(21), 13633–13635. Retrieved from http://www.pnas.org/content/99/21/13633.full doi: 10.1073/

pnas.172510699

Carroll, S. B. (2003). Genetics and the making of Homo sapiens. *Nature*, *422*(6934), 849–857. doi: 10.1038/nature01495

Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J.-E., Lambert, N., de Marchena, J., ... Polleux, F. (2013). NIH Public Access. *Cell*, *149*(4), 923–935. doi: 10.1016/j.cell.2012.03.034.Inhibition

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., ... Pa, S. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, *437*(September), 88–93. doi: 10.1038/nature04000

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., ... Kaessmann, H. (2014, apr). Origins and functional evolution of Y chromosomes across mammals. *Nature*, *508*(7497), 488–93. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24759410 doi: 10.1038/nature13151

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., ... Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Research*, *43*(Database), 662–669. doi: 10.1093/nar/gku1010

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., ... Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, *33*(20). doi: 10.1093/nar/gni179

Davila Ross, M., J Owren, M. & Zimmermann, E. (2009). Reconstructing the Evolution of Laughter in Great Apes and Humans. *Current Biology*, *19*(13), 1106–1111. Retrieved from http://dx.doi.org/10.1016/j.cub.2009.05.028 doi: 10.1016/j.cub.2009.05.028

Dennis, M. Y., Nuttle, X., Sudmant, P. H., Antonacci, F., Tina, A., Nefedov, M., ... Eichler, E. E. (2012). NIH Public Access. *Cell*, *149*(4), 912–922. doi: 10.1016/j.cell.2012.03.033.Human-specific

Digby, L. (1999). Sexual Behavior and Extragroup Copulations in a Wild Population of Common Marmosets (Callithrix jacchus). *Folia Primatologica*, *70*, 136–145.

Diogo, R., Molnar, J. L. & Wood, B. (2017). Bonobo anatomy reveals stasis and mosaicism in chimpanzee evolution, and supports bonobos as the most appropriate extant model for the common ancestor of chimpanzees and

humans. *Scientific Reports*, *7*(1), 1–8. Retrieved from http://dx.doi.org/10.1038/s41598-017-00548-3 doi: 10.1038/s41598-017-00548-3

Dutheil, J. Y., Gaillard, S. & Stukenbrock, E. H. (2014). MafFilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics*, *15*(53).

Edgar, R., Domrachev, M. & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210.

Enard, W., Przeworski, M., Fisher, S., Lai, C. S. L., Wiebe, V., Kitano, T., ... Pääbo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, *418*(6900), 869–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12192408 doi: 10.1038/nature01025

Federhen, S. (2015). Type material in the NCBI Taxonomy Database. *Nucleic Acids Research*, *43*(D1), D1086–D1098. doi: 10.1093/nar/gku1127

Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., ... Huttner, W. B. (2015). Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science*, *347*(6229), 1465–1471.

Friedman, J. R. & Kaestner, K. H. (2006). Review The Foxa family of transcription factors in development and metabolism. *Cellular and Molecular Life Sciences*, *63*, 2317–2328. doi: 10.1007/s00018-006-6095-6

Fujii-Hanamoto, H., Matsubayashi, K., Nakano, M., Kusunoki, H. & Enomoto, T. (2011). A Comparative Study on Testicular Microstructure and Relative Sperm Production in Gorillas, Chimpanzees, and Orangutans. *American Journal of Primatology*, *73*(February), 570–577. doi: 10.1002/ajp.20930

Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. (2017). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)*, *20*(3), 307–315. doi: 10.1093/bioinformatics/btg405

Harcourt, A. H., Harvey, P. H., Larson, S. G. & Short, R. V. (1981). *Testis weight, body weight and breeding system in primates.* (Vol. 293) (No. 5827). doi: 10.1038/293055a0

Hu, Z. & Gallo, S. M. (2010). Identification of interacting transcription factors regulating tissue gene expression in human. *BMC genomics*, *11*(49), 1–15.

Ihaka, R. & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/10618600

.1996.10474713 doi: 10.1080/10618600.1996.10474713

Johnson, N. A. & Lachance, J. (2012). NIH Public Access. *Annals of the New York Academy of Sciences*, *1256*, 1–32. doi: 10.1111/j.1749-6632.2012.06748.x .The

Kanki, K. & Wakahara, M. (2001). The possible contribution of pituitary hormones to the heterochronic development of gonads and external morphology in overwintered larvae of Hynobius retardatus. *International Journal of Developmental Biology*, *45*, 725–732.

Khaitovich, P., Enard, W., Lachmann, M. & Paabo, S. (2006). Evolution of primate gene expression. *Nature reviews Genetics*, *7*, 693–702. doi: 10.1038/nrg1940

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., ... Pa, S. (2005). Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees. *Science*, *309*(September), 1850–1855.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. Retrieved from http://genomebiology.com/2013/14/4/R36 doi: 10.1186/gb-2013-14-4-r36

King, M.-C. & Wilson, A. (2007). Evolution at Two Levels in Humans and Chimpanzees. *Science*, *188*(4184), 107–116.

Kistler, W. S., Horvath, G. C., Dasgupta, A. & Kistler, M. K. (2009). NIH Public Access. *Gene Expression Patterns*, *9*(7), 515–519. doi: 10.1016/j.gep.2009 .07.004.Differential

Kramer, K. L. & Russell, A. F. (2015). Was Monogamy A Key Step on the Hominin Road? Reevaluating the Monogamy Hypothesis in the Evolution of Cooperative Breeding. *Evolutionary Anthropology*, *83*(24), 73–83. doi: 10.1002/evan.21445

Lai, C. S. L., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. & Monaco, A. P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, *413*(6855), 519–523. doi: 10.1038/35097076

Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *NATURE METHODS*, *9*(4), 357–360. doi: 10.1038/nmeth.1923

Leakey, M. G., Spoor, F., Brown, F. H., Gathogo, P. N., Kiarie, C., Leakey, L. N. & McDougall, I. (2001). New hominin genus from eastern Africa shows diverse middle Pliocene lineages. *Nature*, *410*(6827), 433–440. doi: 10

.1038/35068500

Leek, J., Johnson, W., Parker, H., Fertig, E., Jaffe, A., Storey, J., ... Torres, L. (2016). *sva: Surrogate Variable Analysise.*

Leinonen, R., Sugawara, H., Shumway, M., Nucleotide, I. & Database, S. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(November 2010), 2010–2012. doi: 10.1093/nar/gkq1019

Marlowe, F. (2000). Paternal investment and the human mating system. *Behavioural Processes*, *51*, 45–61.

McBrearty & Jablonski, N. G. (2014). First fossil chimpanzee. *Nature*, *437*(October 2005), 105–108. doi: 10.1038/nature04008

McLean, C. Y. (2011). Re: Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*, *471*, 216–219. Retrieved from http://dx.doi.org/10.1038/nature09774 doi: 10.1016/j.eururo.2011.08.037

Namekawa, S. H., Park, P. J., Zhang, L.-F., Shima, James, E., McCarrey, J. R., Griswold, M. D. & Lee, J. T. (2006). Report Postmeiotic Sex Chromatin in the Male Germline of Mice. *Current Biology*, *16*, 660–667. doi: 10.1016/j.cub.2006.01.066

Nurminsky, D. I., Nurminskaya, M. V., de Aguiar, D. & Harti, D. L. (1998). Selective sweep of a newly. *Nature*, *396*, 572–575.

Palmer, S. J. & Burgoyne, P. S. (1991). In situ analysis of fetal, prepuberal and adult XX—XY chimaeric mouse testes: Sertoli cells are predominantly, but not exclusively, XY. *Development*, *112*(Palmer1991), 265–268.

Pansky, B. (1982). No Title. In *Review of medical embryology* (chap. 3). CA: Embryome Sciences Incorporation.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., ... Brazma, A. (2005). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, *33*(553-555). doi: 10.1093/nar/gki056

Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P. & Fodor, S. P. A. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *National Academy of Sciences of the United States of America*, *91*(May), 5022–5026.

Penin, X., Berge, C. & Baylac, M. (2002). Ontogenetic study of the skull in modern humans and the common chimpanzees: Neotenic hypothesis reconsidered with a tridimensional procrustes analysis. *American Journal of Physical An-*

thropology, *118*(1), 50–62. doi: 10.1002/ajpa.10044

Piprek, R. P. (2010). Molecular and cellular machinery of gonadal differentiation in mammals. *The International Journal of Developmental Biology*, *786*(September 2009), 779–786. doi: 10.1387/ijdb.092939rp

Pollick, A. S. & de Waal, F. B. M. (2007). Ape gestures and language evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(19), 8184–9. Retrieved from http://www.pnas.org/content/104/19/8184.short doi: 10.1073/pnas.0702624104

Prabhakar, S., Visel, A., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., ... Noonan, J. P. (2009). Response to Comment on "Human-Specific Gain of Function in a Developmental Enhancer". *Science*, *323*(5915), 714d–714d. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.1166571 doi: 10.1126/science.1166571

Proctor, D., Williamson, R. A., Waal, F. B. M. D., Brosnan, S. F., de Waal, F. B. M. & Brosnan, S. F. (2012). Chimpanzees play the ultimatum game. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(23), 1–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23319633 doi: 10.1073/pnas.1220806110

Raff, R. A., Wray, A. & Biology, C. (1989). Heterochrony: Developmental and evolutionary results. *Journal of Evolutionary Biology*, *2*, 409–434.

Reuter, I., Cheremushkin, E., Kel, A. E., Go, E. & Wingender, E. (2003). : a tool for searching transcription factor binding sites in DNA sequences . *Nucleic Acids Research*, *31*(13), 3576–3579. doi: 10.1093/nar/gkg585

Rice, S. H. (2002). Human Evolution through Developmental Change. In N. Minugh-Purvis & K. J. Mcnamara (Eds.), *Human evolution through developmental change* (pp. 155–170). Baltimore: Johns Hopkins University Press.

Rogers, J., Garcia, R., Shelledy, W., Kaplan, J., Arya, A., Johnson, Z., ... Cameron, J. (2005). An initial genetic linkage map of the rhesus macaque (Macaca mulatta) genome using human microsatellite loci. *Genomics*, *87*, 30–38. doi: 10.1016/j.ygeno.2005.10.004

Rolland, A. D., Niederhauser-wiederkehr, C., Chung, S. S. W., Demougin, P., Primig, M., Gattiker, A., ... Je, B. (2007). The conserved transcriptome in human and rodent male gametogenesis . *PNAS*, *104*(20), 8346–8351.

Schultz, N., Hamra, F. K. & Garbers, D. L. (2003, oct). A multitude

of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(21), 12201–6. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=218736{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1073/pnas.1635054100

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., ... Kasprzyk, A. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, *43*(Web Server issue), 589–598. doi: 10.1093/nar/gkv350

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., ... Kaessmann, H. (2013). Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports*, *3*, 2179–2190. doi: 10.1016/j.celrep.2013.05.031

Ting, C.-T., Tsaur, S.-C., Wu, M.-L. & Wu, C.-I. (1998). Chau-Ti Ting, Shun-Chern Tsaur, Mao-Lien Wu, Chung-I Wu*. *Science*, *282*, 1501–1504.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, R., ... Pachter, L. (2013). NIH Public Access. *Nature Protocols*, *7*(3), 562–578. doi: 10.1038/nprot.2012.016.Differential

Trevino, V., Falciani, F. & Barrera-Saldana, H. A. (2007). DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. *Molecular Medicine*, *13*(October), 527–541. doi: 10.2119/2006

Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *Physiology (Bethesda, Md.)*, *4*(3), 446–458. doi: 10.1371/journal.pbio.0040072

Wang, B., Qi, T., Chen, S.-Q., Ye, L., Huang, Z.-S. & Li, H. (2016). RFX1 Maintains Testis Cord Integrity by Regulating the Expression of Itga6 in Male Mouse Embryos. *Molecular Reproduction & Development*, *83*, 606–614. doi: 10.1002/mrd.22660

Wilhelm, D., Palmer, S. & Koopman, P. (2007). Sex Determination and Gonadal Development in Mammals. *Physiological Reviews*, *87*, 1–28. doi: 10.1152/physrev.00009.2006.

Wood, B. (2002). Hominid revelations from Chad. *Nature*, *418*(July), 133–135. doi: 10.1038/418133a

Wood, B. & Collard, M. (1999). The human genus. *Science*, *284*,

65–71. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.284.5411.65{%}5Cnhttp://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed{&}id=10102822{&}retmode=ref{&}cmd=prlinks{%}5Cnpapers3://publication/uuid/10780B33-C8BC-448D-8AFE-2EB2B64BF70D doi: 10.1126/science.284.5411.65

Wu, C.-i., Johnson, N. A. & Palopoli, M. F. (1996). Haldane' s rule and its legacy: why are there so many sterile males? *TREE*, *11*(July), 281–284.

Wyckoff, G. J., Wang, W. & Wu, C. I. (2000). Rapid evolution of male reproductive genes in the descent of man. *Nature*, *403*, 304–309. Retrieved from file://localhost/References/Paperdownloads/sexualselectionmatechoice/Wyckoff2000Nature.pdf doi: 10.1038/35002070

APPENDIX A

IN-HOUSE PYTHON CODE FOR THE PREPARATION OF THE GTFS

```
### preparing new gtfs
# ensembl gene id, transcript id, chromosome name, strand,
   exon start(bp), gene end(bp) for each species were
   downloaded from ensembl83 version, unique results only on
    4may2016.
import itertools
inf2 = file("EXON_START_END_POSITIONS_FOR_EACH_SPECIES")
inf2.readline() # get rid of the header
ensTrDict = dict()
ensTrChrDict = dict() # for chr and strand
counter = 0
for line in inf2:
  counter += 1
#  print counter
  line = line.split()
  lineGene = line[0] # gene id
  lineTrID = line[1] # transcript id
  lineChr = line[2] # chr
  lineStrand = line[3] # strand
  lineStrPos = int(line[4]) # exon start
  lineEndPos = int(line[5]) # exon end
  linePos = [lineStrPos,lineEndPos]
  linePos.sort()
 if (ensTrDict.has_key(lineTrID)): #if the transcript has
    already been listed
    new = ensTrDict[lineTrID] + [linePos]
    new.sort()
    ensTrDict[lineTrID] = new
  else:
    ensTrDict[lineTrID] = [linePos]
```

```
    ensTrChrDict[lineTrID] = [lineGene, lineChr, lineStrand]


inf = file("MAF_OF_THE_SPECIES_USED")
outf = file("NEWLY_CONSTRUCTED_GTF_OF_EACH_SPECIES","w")

# make a dictionary of block start and end positions from
   the MAF
mafResDict = dict()
counter = 0
speciesx = 'speciesxxx' # speciesx changes with the species
   of interest
for line in inf:
  if (speciesx in line):
    counter += 1
 #   print counter
    line = line.split()
    lineID2 = line[1].split("_")[1]
    lineTrID = lineID2.split(".")[0]
    lineStrPos = int(line[2]) + 1 # because there are line
       [2] positions behind the alignment start
    lineEndPos = int(line[2]) + int(line[3])
    linePos = [lineStrPos, lineEndPos]
    linePos.sort() # sort start and end
    if (mafResDict.has_key(lineTrID)):
      new = mafResDict[lineTrID] + [linePos]
      new.sort() # sort the blocks
      mafResDict[lineTrID] = new
    else:
      mafResDict[lineTrID] = [linePos]

# summarize the dictionary by joining consecutive alignment
   blocks

mafResDict2 = dict()
for trID in mafResDict.keys():
  mafResDict2[trID] = []
```

```
  start = mafResDict[trID][0][0] # the start position of the
      first block
  end = mafResDict[trID][0][1] # the end of the first block
  for i in range(1,len(mafResDict[trID]),1): # loop
     initiates from the second block
    if (mafResDict[trID][i][0] == (end + 1)): # if start of
       the first block follows the previous end
      end = mafResDict[trID][i][1] # end is updated
    elif (mafResDict[trID][i][0] != (end + 1)):
      mafResDict2[trID] = mafResDict2[trID] + [[start,end]]
      start, end = mafResDict[trID][i][0:2] # start, end
         updated
  mafResDict2[trID] = mafResDict2[trID] + [[start,end]] #
     finally add the last entries


# check that MAF blocks do not overlap:
# not anymore but they do join one after another

count1 = 0
count2 = 0
for trID in mafResDict2.keys():
  overlap1 = False
  overlap2 = False
  geneID, chr, strand = ensTrChrDict[trID]
  for i in range(0,(len(mafResDict2[trID])-1),1): # loop
     initiates from the second block
    prevend = mafResDict2[trID][i][1] # the end of the first
        block
    nextstart = mafResDict2[trID][i+1][0]
    if (nextstart - prevend == 0):
      overlap1 = True
    elif (nextstart + prevend < 0):
      overlap2 = True
  if (overlap1):
    print "A", trID, strand
    print mafResDict2[trID]
```

```
    count1 += 1
  if (overlap2):
    print "B", trID, strand
    print mafResDict2[trID]
    count2 += 1
# preparing gtf

for trID in mafResDict2.keys():
#  print trID # get the chr and strand info
  geneID, chr, strand = ensTrChrDict[trID]
  if (strand == "1"):
    strand = "+"
  else:
    strand = "-" # start and end of exons as relative
       positions in CDS (starting from 1)
  ensTrDictTemp = []
  exEnd = 0
  for exon in ensTrDict[trID]:
    exStr = exEnd + 1
    exEnd = exon[1] - exon[0] + exStr
    ensTrDictTemp = ensTrDictTemp + [[exStr, exEnd]] # the
       number of exons
  n = len(ensTrDict[trID])  # check if the gene is on the
     negative strand or not
  if (strand == "-"): # reverse the relative positions in
     the maf for that transcript, based on the max position
     in ensTrDictTemp # same size but the positions reversed
     , so that it now follows the positive strand sequence
    mafResDictTemp = []
    for x in mafResDict2[trID]:
      maxN = 1 + max(list(itertools.chain(* ensTrDictTemp)))
          # max + 1
      res = [maxN - x[0], maxN - x[1]]
      res.sort()
      mafResDictTemp = mafResDictTemp + [res] # sort the
         list
    mafResDictTemp.sort()
```

```
else:
  mafResDictTemp = mafResDict2[trID] # now calculate the
     corresponding positions
for mafX in mafResDictTemp: #run across all mafs fro that
   transcript
  mafX0, mafX1 = mafX # relative coordinates of that maf
  for i in range(0,n,1): # run across all exons
    ex0, ex1 = ensTrDictTemp[i] # relative coordinates of
       the exon
    exGen0, exGen1 = ensTrDict[trID][i] # genomic
       coordinates of the exon
    if ((mafX0 >= ex0) & (mafX1 <= ex1)): # if the maf
       segment is within that exon
      mafGen0 = mafX0 - ex0 + exGen0 # genomic coordinates
          of the maf (start)
      mafGen1 = mafX1 - ex0 + exGen0 # genomic coordinates
          of the maf (end)
      outf.write(chr + "\t" + "x" + "\t" + "exon" + "\t" +
          str(mafGen0) + "\t" + str(mafGen1) + "\t" + "."
         + "\t" + strand + "\t" + "." + "\t" + "gene_id␣"
         + geneID + ";" + "␣transcript_id␣" + trID + "\n")
    elif ((mafX0 >= ex0) & (mafX0 <= ex1) & (mafX1 > ex1))
       : # if the maf segment is partly within that exon
      mafGen0 = mafX0 - ex0 +exGen0 # genomic coordinates
         of the maf (start)
      mafGen1 = exGen1 # genomic coordinates of the maf (
         end) for that exon
      outf.write(chr + "\t" + "x" + "\t" + "exon" + "\t" +
          str(mafGen0) + "\t" + str(mafGen1) + "\t" + "."
         + "\t" + strand + "\t" + "." + "\t" + "gene_id␣"
         + geneID + ";" + "␣transcript_id␣" + trID + "\n")
      stop = False
      while stop != True:
        for j in range((i+1),n,1): # run across all the
           remaining exons
          ex0, ex1 = ensTrDictTemp[j] # relative
             coordinates of the exon
```

```
            exGen0, exGen1 = ensTrDict[trID][j] #genomic
               coordinates of the exon
            if ((mafX1 >= ex0) & (mafX1 <= ex1)): # if the
               maf segment remainder is within that exon and
                finishes there
              mafGen0 = exGen0 # genomic coordinates of the
                 maf (start)
              mafGen1 = mafX1 - ex0 + exGen0 # genomic
                 coordinates of the maf (end) for that exon
              outf.write(chr + "\t" + "x" + "\t" + "exon" +
                 "\t" + str(mafGen0) + "\t" + str(mafGen1) +
                  "\t" + "." + "\t" + strand + "\t" + "." +
                 "\t" + "gene_id␣" + geneID + ";" + "␣
                 transcript_id␣" + trID + "\n")
              stop = True
            elif((mafX1 >= ex0) & (mafX1 >= ex1)): # if the
               maf segment remainder is embedded in that
               exon
              mafGen0 = exGen0 # genomic coordinates of the
                 maf (start) for that exon
              mafGen1 = exGen1 # genomic coordinates of the
                 maf (end) for that exon
              outf.write(chr + "\t" + "x" + "\t" + "exon" +
                 "\t" + str(mafGen0) + "\t" + str(mafGen1) +
                  "\t" + "." + "\t" + strand + "\t" + "." +
                 "\t" + "gene_id␣" + geneID + ";" + "␣
                 transcript_id␣" + trID + "\n")



outf.close()
```

APPENDIX B

IN-HOUSE PYTHON CODE FOR THE MODIFICATION OF THE CDFS

```
cdf=open("CDF_OF_THE_CHIP_DOWNLOADED_FROM_MICROARRAY_LAB","r
    ")
newcdf=open("MODIFIED_CDF_ONLY_TO_CONTAIN_COMMON_PROBES","w"
    )
for h in range(12):
        line=cdf.readline()
        newcdf.write(line)

unitcounter=1
line=cdf.readline()
if line.startswith("["+"Unit"+str(unitcounter)+"]"):
        unit=line
        name=cdf.readline()
        direction=cdf.readline()
        numatoms=cdf.readline()
        numcells=cdf.readline()
        unitnumber=cdf.readline()
        #if unitnumber.startswith("UnitNumber=1"):
                #pass
        #else:
                #print("unitnumber is ",unitnumber)
        unittype=cdf.readline()
        numberblocks=cdf.readline()
        gap=cdf.readline()
        unitblock=cdf.readline()
        geneid=cdf.readline()
        blocknumber=cdf.readline()
        blocknumatoms=cdf.readline()
        blocknumcells=cdf.readline()
        startposition=cdf.readline()
```

```
stopposition=cdf.readline()
cellheader=cdf.readline()
dik=dict()
newnumcells=1
for j in range(int(blocknumatoms.split("=")[1].split
    ("\r")[0])):
        cellline=cdf.readline()
        coord=cellline.split("=")[1].split("\t")[0]+
            ";"+cellline.split("=")[1].split("\t")[1]
        if coord in xy: #xy is the x and y
            coordinates of the common probes of the
            chip that are aligned to both humans and
            chimpanzees
                newcellline="Cell"+str(newnumcells)+
                    cellline[5:]
                dik["cell"+str(newnumcells)]=
                    newcellline
                newnumcells+=1
newcdf.write(unit)
newcdf.write(name)
newcdf.write(direction)
newcdf.write(numatoms[:9]+str(newnumcells)+numatoms
    [-2:])
newcdf.write(numcells[:9]+str(newnumcells)+numcells
    [-2:])
newcdf.write(unitnumber)
newcdf.write(unittype)
newcdf.write(numberblocks)
newcdf.write(gap)
newcdf.write(unitblock)
newcdf.write(geneid)
newcdf.write(blocknumber)
newcdf.write(blocknumatoms[:9]+str(newnumcells)+
    blocknumatoms[-2:])
newcdf.write(blocknumcells[:9]+str(newnumcells)+
    blocknumcells[-2:])
newcdf.write(startposition)
```

```
newcdf.write(stopposition)
newcdf.write(cellheader)
for d in range(len(dik)):
        cl=dik["cell"+str(d+1)]
        newcdf.write(cl)
dik=dict()
newnumcells=1
unitcounter+=1
line=cdf.readline()
newcdf.write(line)
line=cdf.readline()
newcdf.write(line)
line=cdf.readline
```