CONVERGENT EVOLUTION OF PRIMATE TESTIS TRANSCRIPTOMES IN RESPONSE TO MATING STRATEGY DIFFERENCES

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

BY

ETKA YAPAR

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

FEBRUARY 2021

Approval of the thesis:

CONVERGENT EVOLUTION OF PRIMATE TESTIS TRANSCRIPTOMES IN RESPONSE TO MATING STRATEGY DIFFERENCES

submitted by **ETKA YAPAR** in partial fulfillment of the requirements for the degree of **Master of Science in Biological Sciences Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşe Gül Gözen
Head of Department, **Biological Sciences** _____

Assoc. Prof. Dr. Mehmet Somel
Supervisor, **Biological Sciences, METU** _____

Examining Committee Members:

Prof. Dr. Ayşe Elif Erson Bensan
Dept. of Biological Sciences, METU _____

Assoc. Prof. Dr. Mehmet Somel
Dept. of Biological Sciences, METU _____

Assist. Prof. Dr. Alexey Yanchukov
Dept. of Biology, Bülent Ecevit University _____

Date: 01.02.2021

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, Surname: Etka Yapar

Signature :

ABSTRACT

CONVERGENT EVOLUTION OF PRIMATE TESTIS TRANSCRIPTOMES IN RESPONSE TO MATING STRATEGY DIFFERENCES

Yapar, Etka
M.S., Biology
Supervisor: Assoc. Prof. Dr. Mehmet Somel

February 2021, 71 pages

In independent mammalian lineages where females mate with multiple males (multi-male mating strategies), males have evolved larger testicles relative to those lineages where females mate with fewer males (single-male mating strategies). This convergent evolution of relative testis size is attributed to the sexual selection acting as the immense sperm competition between males of multi-male mating lineages. Here I analyze bulk testis transcriptomes of humans, chimpanzees, gorillas, rhesus macaques, mice and rats in attempt to detect traces of convergent evolution of gene expression and find that bulk testis transcriptomes also appear to be convergently evolved. Then, using deconvolution I compare the relative (postmeiotic + meiotic) versus (premeiotic + somatic) cell type ratios (or testis tissue composition) among testes of the species analyzed and show that convergent patterns seen in bulk testis transcriptomes are largely attributable to cell type composition changes among testes of species. However, analyzing cell type-specific gene expression data from spermatocytes and spermatids of humans, rhesus macaques and mice, I show that there is also convergent evolution at the cell-autonomous level, albeit in modest amounts when compared with bulk testis convergence. Finally, I analyze testis development data from mouse

and macaque to show that when the adult bulk testis transcriptomes are compared against testis development of said two species, single-male species like human and gorilla are paedomorphic relative to bulk testis transcriptome profiles of multi-male primates, chimpanzee and macaque. This suggests that shifts in the timing or the rate of testis development could explain convergences in relative testis mass, tissue composition and bulk transcriptomes.

Keywords: transcriptomics, comparative evolutionary biology, bioinformatics

ÖZ

PRİMAT TESTİS TRANSKRİPTOMLARININ ÜREME TİPİ FARKLARINA CEVABEN YAKINSAK EVRİMİ

Yapar, Etka
Yüksek Lisans, Biyoloji
Tez Yöneticisi: Doç. Dr. Mehmet Somel

Şubat 2021 , 71 sayfa

Dişi bireylerin birden çok erkek bireyle çiftleştiği birbirinden bağımsız memeli türlerinde erkek bireyler, dişi bireylerin daha az erkekle çiftleştiği türlere karşılaştırıldığında daha büyük testisler evrimleştirmişlerdir.Vücut büyüklüğüne oranla büyük testislerin geliştiği bu yakınsak evrim daha önce sperm rekabeti olarak isimlendirilen ve çok erkekli üreyen türlerin erkek bireyleri üzerine etkiyen bir çeşit eşeysel seçilim baskısına atfedilmiştir. Bu tezde, testis anatomisi üzerinde etkili ve yukarıda bahsedilen yakınsak evrimin gen ifadesi düzeyinde de gözlemlenebilir olup olmadığını test etmek amacıyla insan, şempanze, goril ve makak tüm testis transkriptom verisi analiz edilmiş olup tüm doku gen ifadesi düzeyinde de bu yakınsak evrim izlerinin saptanabileceği gösterilmiştir. Daha sonra, uygun istatistiki metodlar kullanılarak mayoz sonrası ve mayoz geçiren hücrelerin mayoz öncesi ve somatik hücrelerin göreli miktarları analiz edilen türlerin tüm testis gen ifadesi düzeylerinden çıkarsama yoluyla saptanmış ve tüm testis düzeyinde gen ifadesinin yakınsak evriminin büyük oranda bahsi geçen hücre tiplerinin testis dokusundaki göreli miktarlarının değişimiyle açıklanabileceği gösterilmiştir. Buna karşın, insan, makak ve fare spermatosit ve sper-

matid hücre tiplerine özgü gen ifadesi verisinin analiz edilmesi sonucunda, hücre içi gen ifadesi değişimlerinin de testis doku bileşimi kadar güclü bir etkiye sahip olmasalar da anlamlı etkileri olduğu görümüştür. Son olarak, bahsi geçen tüm testis gen ifadesi verilerinin makak fare testis gelişimi gen ifadesi verisi ile karşılaştırmalı analiz edilmesiyle insan ve goril testislerinin tüm doku gen ifadesi düzeyleri açısından bakıldığında makak ve şempanze testislerine göre pedomorfik durumda olduğu, yani erişkin insan ve goril testislerinin ergen veya çocuk makak veya şempanze testislerine daha benzer oldugu gözlenmiştir. Bu bulgu, testis gelişimi sırasındaki zamanlama değişimlerinin potansiyel olarak yukarıda bahsi geçen anatomik ve gen ifadesi düzeyinde testis yakınsak evrimi örüntülerini açıklayabilecek gelişimsel ana mekanizma olabileğini göstermesi açısından önem arzetmektedir.

Anahtar Kelimeler: transkriptom, karşılaştırmalı evrimsel biyoloji, biyoenformatik

To all my cat babies, Inci, Maya, Bonbon and Shiraz.

ACKNOWLEDGMENTS

This work would not have been possible without the support of many people to whom I would like to express my gratitude. First, I would like to mention my supervisor and my mentor for the past five years during my studies in his research group, Mehmet Somel. I believe I could not go through all this without his guidance, his neverending optimism, and -at the same time- his utterly realistic ways of reassuring us at times of failure.

Of course, not all credit is due to him for the support I received during those hard times. I want to thank all members of the CompEvo Lab for the social atmosphere, for the friendship, for everything they have been to me for all these years. Among them, I have to single out Ekin Sağlıcan, Ezgi Özkurt, and Melike Dönertaş for undeniably being the backbone of the testis evolution project, whose results are represented in this thesis; and Hamit Izgi and Gözde Turan for the careful proofreading. Here I also would like to thank my thesis committee members, Ayşe Elif Erson Bensan and Alexey Yanchukov, for the evaluation of my thesis and their insightful suggestions that made this work stronger.

I should also thank all of our collaborators, within or outside of METU, that were essential for this project to work out: Rori Rohlfs for providing us the EVE software and days worth of discussion on our results; Philipp Khaitovich and his colleagues Song Guo, Haiyang Hu, and Zheng Yan for generating the invaluable biological data and sharing it with us; and Babür Erdem for all the things he has done for the project. Here I should also mention Bluma Lesch, whose data we use for the cell-autonomous convergency tests, for helping me when I was starting to analyze RNA-Seq data for the first time in my life by providing additional information about the lab protocols used for data generation and the discussion regarding my pre-processing results.

I also want to thank my sweetheart, Ece, whom I shared my life with for the past three years, for being there for me on every occasion, her love and endless support,

kindness, and companionship that makes this cruel world a better place to live for me. My family for everything I am today, for bringing me into this world, raising me, and supporting me when I chose the path of being a researcher.

Lastly, I want to thank the Scientific and Technological Research Council of Turkey, the Turkish Academy of Sciences, and METU BAP office for supporting this project with the 2232 Fellowship (no. 114C040), the BAGEP-2014 awards, and project funding with the code BAP-07-02-2015-009 respectively.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

dNTP	Deoxynuclotide triphosphate
DEG	Differentially expressed gene
GTF	Gene transfer file
cDNA	Complementary DNA
RNA-Seq	RNA sequencing
GO	Gene ontology
BP	Biological process
TF	Transcription factor
SMBL	Single-male branch length
MWU	Mann-Whitney U test
PS	Pachytene spermatocyte
RS	Round spermatid
TFBS	Transcription factor binding site
TSS	Transcription start site

CHAPTER 1

INTRODUCTION

1.1 Testis histology

Testis is the male gonad that both produce male gametes, sperms, and the male reproductive hormone, testosterone. It is possible to examine the histology of mammalian testis in two distinct parts. In the interstitial region, the most commonly found cell types are Leydig cells and immune cells such as mast cells and macrophages. Leydig cells are worth mentioning in the context of reproduction in that they produce and secrete testosterone, which helps the development and maintaining of the secondary male characteristics. Seminiferous tubules are where spermatogenesis takes place and host many cell types with particular importance. In addition to their "nurturing " role in maintaining spermatogenesis, Sertoli cells are also the first cell type that emerges during testis differentiation from urogenital ridge during fetal development and assists the testicular differentiation, in turn, would yield other cell types. Spermatogonia are the precursor cells that can either produce copies of themselves or spermatocytes, which will go under meiotic division through spermatogenesis to produce the mature sperm (Fietz et al., 2017).

1.2 Sexual selection

Despite its *ad nauseam* referral to explain sexual selection, male peafowls' ornaments developed in parallel with female individuals' selective mate choice behavior is a textbook case.

In detail, male individuals of this species carry larger tails and ornamental feathers

with colorful eyespots, or ocelli, when compared to the rather dull appearance of the female individuals. Density, but the not number, of ocelli found in an individual male, has been shown to be a reliable estimator for increased mating success before, along with other traits such as train length (Loyau et al., 2005).

This dimorphism may seem not very interesting at the surface, but it paints an absurd picture from a survivalist standpoint. That is because bizarre traits like the peacock's long train and ocelli suggest a reduced survival advantage. However, survival at the cost of not being able to mate would not bring any evolutionary advantage. Although this specific peacock trait has been shown not to affect the metabolic cost of free movement (Thavarajah et al., 2016), the survival versus mating success dilemma still holds. After all, without passing their genes to their offspring living would mean the least for the peacocks.

Although this specific example is often regarded as the definitive example of sexual selection, it only touches upon one specific form of sexual selection, which happens intersexually -mate choice- and overlooks an equally important form of sexual selection that happens intrasexually, competition.

In general, sexual selection may refer to any force of natural selection that directly acts on a species' reproductive success rather than survival (Darwin, 1896). In the case of mate choice, as explained above, selective mating of one sex based on several phenotypes of the opposite sex creates an immense selection pressure on the sex that is being selected for their appearance, which the case of peacocks displays an extreme example. On the other hand, the competition -usually among the males- may result in the evolution of arguably more interesting behavioral and anatomical traits. It is possible to further classify intrasexual selection into two complementary steps as the forces acting on traits related to the points before copulation and those after copulation. For instance, under female promiscuity, post-copulatory intrasexual selection among males might be the dominant form of sexual selection, whereas under male promiscuity -or harems- the effect of pre-copulatory male competition would be more critical in that it will determine the access to a receptive female.

1.3 Sperm competition

Post-copulatory intrasexual selection, which was mentioned in the above section, is historically referred to as sperm competition. It is of particular importance for topics covered in the following sections regarding the convergent evolution of testis anatomy and histology.

1.3.1 Convergent testis size evolution in relation to sexual selection

It has been previously reported that it is possible to explain differences in testis mass among different species by different mating strategies. In lineages with female promiscuity (e.g., polygynandry or polyandry), there is profound sperm competition among individual males. This selection pressure is anticipated to result in ejaculation in higher volumes, higher testis mass, and more frequent copulation (Short, 1979). Almost four decades ago, Harcourt and colleagues showed that differences in their mating behavior could explain the relative testis weight of species by analyzing testis and body mass data from approximately 30 primate species. In detail, species with female promiscuity, or multi-male mating species, generally have higher residual testis mass than species with less female promiscuity or single-male mating species (Harcourt et al., 1981). Thus the mechanism behind the testis weight convergence explained above could be the elevation of this kind of selection pressure in multi-male lineages like chimpanzee and macaque or the relaxation of it in single-male lineages such as human and gorilla. Later studies explored this testis mass and mating strategy relationship with a broader set of mammalian species (Dixson, 2012; Harcourt et al., 1995; Hosken, 1997; Kenagy et al., 1986; Ramm et al., 2005). See **Figure 1.1** for visualization of compiled testis weight / body weight data taken from (Kenagy et al., 1986).

1.3.2 Testis tissue composition and rate of spermatogenesis

The ratio of seminiferous tubules to connective tissue in humans has been shown to be close to 1, whereas this ratio is around 2-2.5 for multi-male primate species such

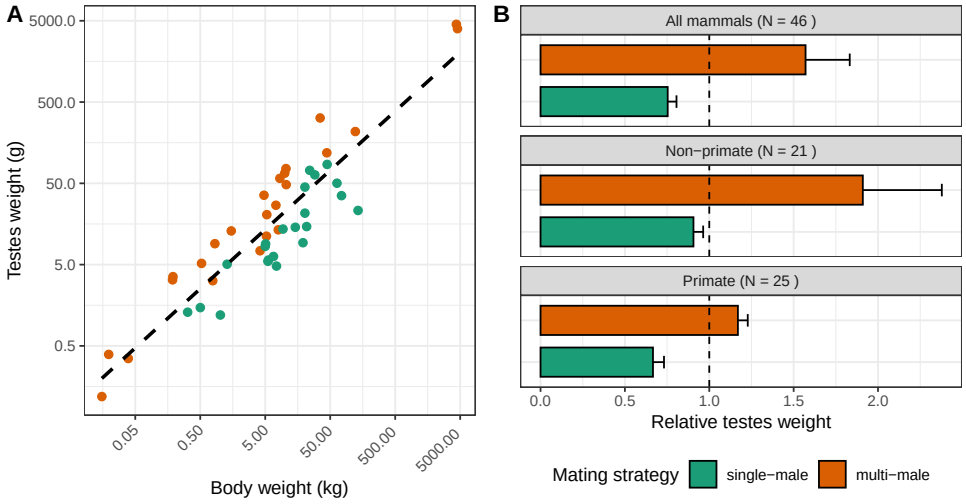

Figure 1.1: **(A)** Relationship between body weight and combined testes weight of 46 mammals. The plot and the linear model was constructed with both axes log transformed. **(B)** Distribution of relative testis size stratified according to different mating strategies. "Relative testes weight" was calculated as the ratio of actual log testes weight and fitted log testes weight according to the linear model used in panel A. Data for both panels and the plot idea for panel B were taken from Kenagy et al., 1986. See Table 1 and Figure 1 of cited work, respectively.

as chimpanzees or macaques (A. H. Schultz, 1938). It is suggested that strong sperm competition also selects for faster spermatogenesis (Ramm et al., 2010). Findings related to convergences of testis mass, histology, and spermatogenesis rates in response to sexual selection altogether highlight that the histological and anatomical evolution of testes is a fast process since it is possible to observe this kind of convergence within a four-species primate phylogeny with closely related species: human, chimpanzee, gorilla and macaque. Among many factors that allow this rapid evolution, plausible ones may include the modularity of the testis development organization (Sekido et al., 2013; Stockley, 2004) and the potent positive selection on reproductive phenotypes (Dixson, 2012; Short, 1979; Stockley, 2004) of species mentioned above.

1.4 Transcriptome evolution

Here in this context, transcriptome evolution may refer to any divergent or convergent changes in gene expression that has occurred between two or more lineages. To study the evolution of gene expression in a transcriptome-wide manner, high-throughput methods are needed.

1.4.1 Methods for high-throughput gene expression measurement

As stated above, evolutionary studies on transcriptomes need gene expression data coming from tens of thousands of genes in parallel. There are two major technologies that allow the generation of data of such nature.

1.4.1.1 Microarray

Since the first emergence of the photolithographic techniques that made synthesizing oligonucleotide probes onto glass slides with high precision in 1991 (Fodor et al., 1991) and the creation of cDNA microarray to be used to hybridize total cellular mRNA pool from *Arabidopsis thaliana* as a novel way of measuring gene expression (Schena et al., 1995), DNA microarray (or DNA chip) technologies allow high-throughput parallel measurement of gene expression.

The actual quantification step relies on emitted light during the hybridization of mRNAs with oligonucleotide probesets engineered specifically for each gene or transcript. Aliquots of the source mRNA pool are distributed to micro spots on the chip, which itself includes a set of ligated oligonucleotide probes specific to one feature. After hybridization, emitted light from the chip is recorded, and it is possible to quantify gene expression with the raw light intensity data and the metadata of the geometric coordinates of used probesets and their target features.

1.4.1.2 RNA Sequencing

In the year 2008, three studies published within two months apart from each other marked the beginning of a new era in transcriptomics (Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008). These were the first examples of using RNA Sequencing to measure gene expression of Arabidopsis, mammals, and yeast, respectively. RNA Sequencing, or RNA-Seq for short, uses the NGS technology on cDNA fragments created from pools of isolated mRNA. After obtaining the cDNA libraries, the process is essentially the same as the DNA sequencing. Although the exact molecular technology that enables sequencing DNA fragments can change between different platforms, most technologies rely on punctuated replication of the isolated fragments and obtain base calls between sequential steps of this controlled synthesis of DNA inside the respective instruments.

In summary, this process requires a pool of fragmented, amplified, size selected, and adapter-ligated DNA fragments, or sequencing libraries, to operate. Ligated sequencing adapters allow the positioning and ligation of DNA fragments to the sequencing machine's flow cell for clonal amplification of DNA strands to populate their microenvironments on the flow cell surface to form what is called clusters. Amplification of this sort allows for more precise detection of emitted light by resulting in a larger surface that will give out the signal during sequencing. Ligated adapters also contain primer seed sequences needed for the sequencing runs to start. The sequencing step involves the parallel and sequential amplification of ligated DNA fragments scattered through the flow cell. In detail, each cycle of sequencing, which refers to a single nucleotide addition for all the template fragments, consisting of several steps:

1. washing of the flow cell with color tagged dNTPs,
2. consequent incorporation of said tagged dNTPs into growing chains,
3. washing out unbound dNTPs followed by imaging the flow cell surface for base calls,
4. washing of fluorophore and the bound groups that are preventing elongation by blocking the 3'-OH group.

This strictly controlled sequencing progression is also called "Cyclic reversible termination," which refers to the added blocking groups and fluorophores attached to the dNTPs (Goodwin et al., 2016).

The obtained short NGS reads are then aligned to a reference genome and/or transcriptome and hits per feature are counted to have a measure of gene expression.

1.4.1.3 Comparison of two technologies

There are differences in resolution and accuracy of the two platforms' measurements of annotated genes' expression. First, RNA-Seq performs better on genes with extremely low or high expression in terms of sensitivity and accuracy compared to microarray (Zhao et al., 2014). Additionally, being a sequencing-based platform, RNA-Seq is not affected by limitations related to hybridization-based experiments such as signal saturation, background light noise, or biases resulting from differences in hybridization efficiencies of different probes (Goodwin et al., 2016).

It is essential to acknowledge that the raw data generated by the two platforms differ in their nature and what they actually represent. While the light intensity data from microarray experiments can be a direct measure of gene expression, the raw data for RNA-Seq experiments is simply a text file with nucleotide sequences of DNA fragments found in sequencing libraries. Thus, by nature, RNA-Seq data can reveal additional biological information that microarray data cannot. For instance, it is possible to identify novel transcripts of a gene with RNA-Seq data, whereas the microarray data is limited to the probesets used in the hybridization of mRNA pools at the time of the experiment. Additionally, although not of direct importance in gene expression studies, it is also possible to identify single nucleotide variants (SNVs) for the coding regions of the genomes using RNA-Seq data (Zhao et al., 2014).

Both microarray and RNA-Seq platforms have their limitations regarding raw data pre-processing that stem from the respective platforms' method of operation. It is faster to computationally pre-process the raw data to obtain a gene expression matrix for microarray data. In contrast, for RNA-Seq data, this process usually takes hours with legacy software (Bray et al., 2016). This difference stems from the sequence

alignment step usually required for pre-processing the raw RNA-Seq data. It is also worth mentioning that the RNASeq platform produces raw data that is substantially larger in file size than microarrays (Zhao et al., 2014).

1.4.2 Previous studies on testis transcriptome evolution

There are several research articles that are worth mentioning in terms of -sometimes indirect- findings regarding primate testis transcriptome evolution.

Using microarray technology, Khaitovich and colleagues showed that there is substantially higher gene expression divergence between humans and chimpanzees and low within-species divergence in bulk testis transcriptomes relative to brain and liver transcriptomes (Khaitovich et al., 2005, 2006b). In addition to this observation, it is also reported that genes expressed in testis display faster evolution of protein-coding sequences than other genes (Khaitovich et al., 2006a, 2005). Brawand and others corroborated this unique feature of testis among other organs in that it shows the highest pace of evolution among six tissues at the transcriptome level, utilizing bulk RNA-Seq technology (Brawand et al., 2011). Even though that convergence is not quantified, it was also noted in the same study that the human bulk testis transcriptome appears more similar to the gorilla than to the chimpanzee.

Unfortunately, comparative aspects of molecular mechanisms of primate testis development have yet received meager attention. One exception is a recent study by Cardoso-Moreira and colleagues, in which gene expression from numerous organs was studied across different species' development (Cardoso-Moreira et al., 2019). The authors noted that testis is among the tissues with the highest number of genes with signs of positive selection and also has the highest number of trajectory changes, that is having the highest number of newly acquired gene expression patterns across different mammals throughout organ development. They continue with underlining that even though the testis is one of the fastest evolving organs among mammals, humans have slower evolving testes than more promiscuous species like mice, rats, or rabbits. Though not common at the whole organ level, the authors also reported the existence of heterochronic patterns, that is, changes in the rate of development, in different organs with sub-groups of genes, particularly in testis regarding the onset of

meiosis.

Another example is a new study, Wang et al., 2020, which presents essential findings about evolutionary mechanisms by which gene expression is regulated at the transcriptome and translatome (sample universe of mRNAs that are actively being translated) layers by using RNA-Seq data from the brain, liver, and testis across six species. They also mention some unique features of testis gene expression. First, they report that the translatome layer generally shows similar or greater gene expression variation than the transcriptome layer for the brain and liver. This trend is reversed for testis transcriptome and translatome layers of four out of six species included in the study. The authors attempt to explain this reverse trend in the testis by attributing it to an inverse relationship between transcript abundance and respective translational efficiencies of said transcripts, or anticorrelations. This is strengthened by the observation that four species with reduced translatome expression variation are the same four species that show the strongest anticorrelations. They further hypothesize that these anticorrelations that are unique to testis could stem from strong translational repression of genes that are mainly expressed in meiotic and post-meiotic cell types, which mature testis include in great abundance relative to other cell types. When the respective rates of evolution are compared between translatome and transcriptome layers, the number of genes with faster transcriptome evolution are consistently higher for all three organs. However, when the relative numbers of genes with faster translatome versus transcriptome evolution compared across organs, testis has the highest ratio. Moreover, when the distributions of relative paces of translatome versus transcriptome evolution among organs are compared, testis has the biggest contrast.

1.5 Motivation of this study

A previous master's thesis from our research group was presented three years ago (Sağlıcan, 2018). That work already visited some of the research questions discussed here. In summary, Ekin aimed to show that the anatomical testis convergent evolution is also paralleled by human, chimpanzee, gorilla, and macaque bulk testis transcriptomes and also showed possible developmental and histological mechanisms by which the said bulk testis gene expression convergence could be explained. The main

findings of the said work can be summarized in four major points:

1. In the scope of genes separating human and chimpanzee bulk testis transcriptomes, testes of multi-male mammalian species such as macaques, mice and rat show significantly higher correlation to chimpanzee than human, and the single-male gorilla shows an opposite trend.

2. When the gene expression profiles of individual cell types from mouse testis are used and summarized as premeiotic/somatic (PRE) and postmeiotic/meiotic (POST) cells, it has been seen that single-male species such as human and gorilla display significantly lower POST:PRE ratios when compared to multi-male species included in the analyses.

3. When the adult bulk testis samples of human, chimpanzee, gorilla, and macaque are compared to mouse and macaque testis developmental time points at the transcriptome level, single-male primate species display the most similarity to earlier time points of macaque and mouse testis development than multi-male species do. In other words, human and gorilla have paedomorphic testicles when compared to chimpanzee and macaque.

4. Clustering genes common for macaque and mouse testis development and adult bulk testis samples of four primate species; It is possible to obtain two clusters such that genes belong to one cluster display overall increase throughout testis development and show a multi-male overexpression for the adult bulk testis samples, whereas the other cluster having the opposite pattern. Moreover, the former cluster also shows enrichment in Gene Ontology Biological Process categories related to reproduction.

However, it is also worth mentioning what Ekin suggested as future work and limitations in her thesis to improve that comprehensive analysis:

1. The gene expression convergence in bulk testis transcriptomes of primate species was not methodologically tested for each gene. Instead, differential transcriptome-wide correlations to human and chimpanzee were used as a proxy for measuring the overall convergence that can be detected using gene expression data.

2. The analyses listed were only performed in testis, without a tissue control, which would show the patterns discovered are unique to the testis.

3. Although there is apparent convergence of POST/PRE ratios of single-male and multi-male species testes, the relative contribution of cell-autonomous effects to the convergence seen in bulk testis transcriptomes remains to be answered.

1.5.1 Novel contributions of this study

Here, I analyzed a superset of the data that was used in Ekin's thesis, aiming to address several limitations listed above by re-analyzing the data with improved methods and including additional data from adult primate brain transcriptome with the same four species, a human testis development dataset and a cell type-specific gene expression data of spermatocytes and spermatids from human, macaque, and mouse. Below I explain these novel contributions.

First, I used a probabilistic framework, the EVE model (Rohlfs et al., 2015) (see **Section 2.3.2** for detail), to model convergent gene expression changes using the human, chimpanzee, gorilla and macaque bulk testis data. This new method allows me to test individual genes for convergent evolution and infer the strength of transcriptome-wide convergence using the proportion of genes that display convergent evolution. Second, for the transcriptome-wide convergence and developmental analyses I utilized a prefrontal cortex RNA-Seq data as a tissue control. Third, I used the above-mentioned spermatocyte and spermatid data from human, macaque and mouse to detect possible cell-autonomous gene expression convergence and later compared this effect with the cell type ratio differences. Lastly, I improved the methodology for gene clustering, functional annotation and transcription factor binding site analyses by including a human testis development dataset and using more recent and robust databases and statistical methods for these analyses. With the improved version of these analyses, we submitted the findings discussed here as a research article to the journal *Evolution*.

CHAPTER 2

MATERIAL AND METHOD

2.1 Raw data used and preprocessing

2.1.1 Creating custom transcriptome annotation

Measuring gene expression from RNA-Seq data involves aligning NGS reads to a reference genome and/or transcriptome and then counting the hits per gene and/or transcript. This process relies on normalization according to the total number of reads per library (read depth) and lengths of corresponding genes. Thus, it is highly sensitive to differences between lengths of annotated genes or coding sequences. Taking this problem into account becomes even more crucial when one is dealing with cross-species data since the completeness of genome annotation can differ substantially among species. To overcome this problem, different studies employed similar approaches where a custom transcriptome annotation was created to only include perfectly alignable, 1:1 orthologous exon sets among the all the species involved in a particular analysis (Brawand et al., 2011; Sağlıcan, 2018). In particular, I used the same approach taken in (Sağlıcan, 2018) whenever I processed a raw RNA-Seq data to be included in this study, and used raw RNA-Seq data preprocessed into FKPM gene expression matrices created by the author, where I was using data from that particular study.

The specific preprocessing approach as it is implemented in (Sağlıcan, 2018) is summarized below.

Two sets of custom GTF files were created to include only 1:1 orthologous and alignable exons:

- GTF A: Content of this transcriptome annotation is limited to 1:1 orthologous and alignable exons of only the primate species analyzed
- GTF B: Content of this transcriptome annotation is limited to 1:1 orthologous and alignable exons of all the mammalian species analyzed.

In any case depicted above, the process of creating the final annotation is the same except the scope of species the annotation includes. In detail, the 1:1 orthologous gene set information was downloaded from Ensembl version 83 (Cunningham et al., 2015; Yates et al., 2020) using the Biomart tool of the website. Then, for each orthologous gene set, exonic sequences for each species downloaded again through Ensembl. The longest coding sequence was selected and used for genes with multiple coding transcripts. TBA software (Blanchette et al., 2004) was used to align sequences of species inside one gene set using multiple sequence alignment approach. Finally, a custom Python script is used to filter and cut the Ensembl GTFs for species utilizing the MSA information so that only exonic sequences that were fully alignable for all species were included in the final output. The exact script can be accessed from the Appendix of the related thesis (Sağlıcan, 2018). Moreover, gaps encountered in one or more species were removed from the output. There was a final filtering step based on the resulting length of coding sequences so that only genes with total length >100bp were retained.

2.1.2 Common pre-processing steps for all RNA-Seq data

RNA-Seq reads were mapped to Ensembl version 83 genomes and transcriptomes using `TopHat2` software (Kim et al., 2013). For the reported hits, maximum of 1 mismatch between read and the reference and 2 multi-hits total in the genome/-transcriptome were allowed, alignment search was limited to known splice juctions (`-no-novel-juncs`), and maximum and minimum intron length parameters were changed to 1Mb and 40bp respectively. Using the optional fields which are specific for the aligner, resulting alignments were filtered to contain unique mappers only. Retained alignments were quantified as FPKM gene expression per gene using `Cufflinks` software (Trapnell et al., 2010), using the appropriate custom GTF file (see the previous section). Finally, resulting FPKM values were transformed as

$\log_2(FPKM + 1)$ and then quantile normalized with `preprocessCore` R package (Bolstad, 2019; Gautier et al., 2004).

In **Tables 2.1** and **2.2**, I listed the accession codes for data used here and the organization of combined datasets I created from those data sources, respectively.

Table 2.1: Accession numbers

Description	Accession code
(Brawand et al., 2011)	GSE30352
(Khaitovich et al., 2005)	E-AFMX-11
(Chalmel et al., 2007)	E-TABM-130
(Namekawa et al., 2006)	GSE4193
(Lesch et al., 2016)	GSE68507
(Cardoso-Moreira et al., 2019)	E-MTAB-6814

2.2 Creation of combined datasets

2.2.1 Primate / Mammal bulk testis / brain

For both the adult primate and mammal bulk testis combined datasets, I used a revised normalization approach which leverages the fact that both data sources included human and chimpanzee samples when merging the main two data sources (Brawand et al., 2011; Khaitovich et al., 2005), adopted from (Sağlıcan, 2018). The process in detail explained below.

1. First I merged two data based on common genes between them and recorded the average gene expression per gene per data source but only using the same number of human and chimpanzee samples (two humans and two chimpanzees for both data sources). This value was later used in the last step of the normalization and created as a way to preserve the relationship among genes before normalization.

2. Then, I calculated the average and standard deviation of gene expression of

human and chimpanzee samples per gene per data source. These value I denote $\mu_{h,c}$ and $\sigma_{h,c}$. I maintained the balance of human/chimpanzee sample sizes within each dataset by discarding a random human individual from (Khaitovich et al., 2005) data.

3. After obtaining the $\mu_{h,c}$ and $\sigma_{h,c}$ values per data source, I scaled two data sources separately by subtracting each data source's $\mu_{h,c}$ from gene expression values and dividing this result by respective $\sigma_{h,c}$ values.

4. As a last step, I added the value calculated in step 1 to the scaled matrix. This way I preserve the information regarding relative expression among genes.

2.2.2 Mouse cell types

This combined dataset was created in a similar way explained above for combined bulk testis datasets since both mouse cell type data sources included samples from spermatid and spermatogonium cells.

2.2.3 Human, macaque and mouse testis development

This combined dataset was used for input in k-means clustering for profiling gene expression among three species' testis development (see **Section 2.3.7**). To create the combined testis development dataset, I first scaled each gene's expression values to mean = 0 and sd=1 for three matrices separately. Then I simply joined all data sources together using common genes.

2.3 Statistical analyses

All basic statistical analyses and data visualization were done in R v3.6.3 (R Core Team, 2020). All species icon illustrations were downloaded as royalty free vector images from various websites such as freepik.com, vectorstock.com, and vecteezy.com. Mann-Whitney U (MWU) test was used to assess significance of group differences and in particular to identify differentially expressed genes (DEGs). For all applica-

Table 2.2: Summary of the data used to create combined datasets.

Created Dataset	Content	Used Data
Primate bulk testis	8 Human, 7 Chimpanzee, 1 Gorilla, 2 Macaque	(Brawand et al., 2011; Khaitovich et al., 2005)
Mammal bulk testis	Primate bulk testis + 2 Mouse + 2 Rat	(Brawand et al., 2011; Chalmel et al., 2007; Khaitovich et al., 2005)
Primate bulk prefrontal cortex	10 Human, 11 chimpanzee, 1 Gorilla, 2 Macaque	(Brawand et al., 2011; Khaitovich et al., 2005)
Mouse cell type profiles	2 Sertoli, 6 Spermatogonia, 4 Spermatocyte, 4 Spermatid	(Chalmel et al., 2007; Namekawa et al., 2006)
Human, Macaque, Mouse PS & RS	3 Human, 2 Macaque, 2 Mouse for both cell types	(Lesch et al., 2016)
Human, Mouse, Macaque testis dev.	13 Human [7mo-55y], 15 Mouse [1d-60d], 12 Macaque [16d-26y]	(Cardoso-Moreira et al., 2019; N. Schultz et al., 2003), Novel
Mouse neocortex dev.	8 Mouse, [1 to 122 days of age]	Novel

tions of multiple testing correction, Benjamini-Hochberg (BH) method was used to adjust the nominal p-values into q-values and 0.1 was used as a threshold.

2.3.1 Transcriptome-wide correlations of adult bulk testis data

For checking the relative overall correlations of bulk testis transcriptomes of the gorilla, macaque, mouse and rat to human versus chimpanzee, I first identified differentially expressed genes between human and chimpanzee samples in the combined adult primate bulk testis dataset using the MWU test. This yielded n=4,295 DEGs. Then using only those genes, I calculated the Spearman correlation between pairs of (i) aforementioned four species and (ii) human versus (iii) chimpanzee, using the average gene expression across samples when there is more than one sample for the species in the set (i). Then, I checked the difference between human versus chimpanzee correlation coefficients for each species separately to see if any species show significantly more similarity to human or chimpanzee using two approaches:

1. directly using the two-sided MWU test p-value.
2. with a permutation test approach by permuting the human/chimpanzee labels for $N = 10^5$ iterations to obtain a null distribution and then calculating an empirical p-value comparing the obtained U value from the MWU test against the null distribution of U values created by permutation.

2.3.2 Using EVE for tests of convergent evolution across species

Here my goal was to test genes in the adult primate bulk testis data for convergent evolution among the four primate species according to their different mating systems. Using this phylogeny, which consists of closely related four-species such that the neighboring pairs do not share the same mating system (**Figure 2.1**), allows for a high-resolution setting in which the convergence tests can be conducted.

To test for convergent gene expression patterns, I used the Expression Variance and Evolution (EVE) model (Rohlfs et al., 2015), which is also implemented as an R package (Gillard et al., 2020). EVE relies on the Ornstein-Uhlenbeck (OU) process to

model the phenotypic evolution, while accounting for within species variance in gene expression. In summary, EVE tests two alternative probabilistic models per gene, representing (i) the null model of stabilizing selection towards a common optimum for gene expression level among all species, and (ii) the alternative model of convergent evolution towards two different optima separating species that are determined *a priori* for testing to be in the same group. Here in this context, the optima to which the gene expression is pulled towards is also called theta (θ). To sum up, I test each gene and obtain a likelihood ratio test statistic (LRT) value of two competing models:

- $\mathrm{M}_0 : \theta_{multi-male} = \theta_{single-male}$
- $\mathrm{M}_A : \theta_{multi-male} \neq \theta_{single-male}$

LRT values approximate chi-square distribution with one degree of freedom (χ^2_1). This, therefore, means that it is possible to obtain a p-value to assess the significance of each gene's convergent expression. EVE analysis described above for the bulk testis convergence is also employed with different datasets in this study targeting different hypotheses. A summary list of all EVE tests employed is given below.

1. Bulk testis data for convergent evolution under different mating strategies across the 4 primates. Testing separation of human and gorilla from chimpanzee and macaque. [(human-gorilla) vs (chimpanzee-macaque)]
2. Bulk testis data for control of mating strategy tests. [(human-macaque) vs (chimpanzee-gorilla)]
3. Spermatocyte or spermatid pooled cell transcriptome data for convergent evolution under different mating strategies [(macaque-mouse) vs (human)]
4. Spermatocyte or spermatid pooled cell transcriptome data for control of mating strategy tests [(human-mouse) vs (macaque)]
5. Primate brain transcriptome data for convergent evolution under different mating strategies across the 4 primates [see item 1 for species]
6. Primate brain transcriptome data for the control of mating strategy tests. [see item 2 for species]

7. Cell type proportion ratios using adult primate bulk testis [see item 1 for species]

8. Cell type proportion ratios using adult mammal bulk testis [(human-gorilla) vs (chimp-macaque-mouse-rat)]

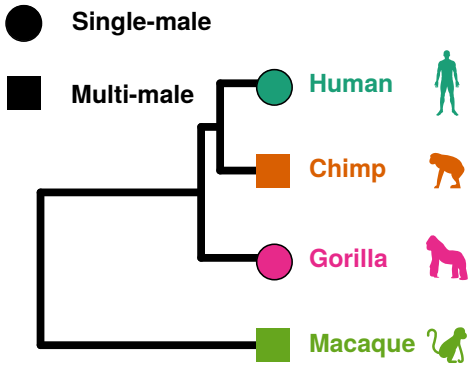

Figure 2.1: The adult primate bulk testis phylogeny.

2.3.3 Prediction of cell type proportions

To predict relative contributions of meiotic/postmeiotic (POST) versus somatic/premeiotic (PRE) cells to the adult bulk testis transcriptomes, I used a deconvolution like approach (Gong et al., 2011). To achieve this, I first created POST and PRE cell type profiles by simply calculating mean expression per gene across all PRE or all POST cells in the combined mouse cell type dataset (n=7,315 genes in total), which will be denoted by E_{PRE} and E_{POST}, respectively. Out of these genes, 4,724 were common with the adult mammal bulk testis data and therefore available for following the analysis. Then, with the gene expression values across bulk testis samples, denoted E_{BT}, I built a linear regression model:

$$E_{BT} = a + b_{PRE} * E_{PRE} + b_{POST} * E_{POST} + \epsilon \tag{2.1}$$

where b_{PRE} and b_{POST} represent regression coefficients, a is the intercept, and ϵ is the error. Using these resulting coefficients, I then calculated $log_2(b_{POST}/b_{PRE})$ ratios

per sample to express the relative POST versus PRE cell type contribution to bulk testis gene expression.

2.3.4 Assessing the role of cell type composition shifts in convergence of bulk testis transcriptomes

To test how well cell type ratio shifts explain convergence seen in bulk testis level, I incorporated adult primate bulk testis and combined mouse cell type datasets. Using common genes between two datasets, I calculated two effect sizes to compare, using Cohen's D with pooled variance for each gene.

- PRE versus POST effect size calculated on combined mouse cell type dataset, denoted $E_{PRE-POST}$
- Single-male versus multi-male effect size on adult primate bulk testis dataset, denoted E_{SM-MM}.

Spearman correlation was used to check the relationship between these two measures. Additionally, I compared the magnitude of cell type effect size (absolute value of $E_{PRE-POST}$) for genes that show convergent evolution according to EVE and for those who do not.

2.3.5 Bulk testis versus cell-autonomous convergent patterns

To compare the relative importance of bulk testis versus cell-autonomous effects of convergent evolution, I used two alternative approaches, which are explained in the following sections.

2.3.5.1 Studied using multiple regression

In this analysis, I used multiple regression to quantify relative importances of two effects. In particular, I built three different linear regression models with the same set of explanatory variables and three different estimates: (a) human-macaque effect

size in bulk testis data [$E_{hsa-mml(BT)}$], (b) human-chimpanzee effect size in bulk testis data [$E_{hsa-ptr(BT)}$], and (c) gorilla-chimpanzee effect size in bulk testis data [$E_{ggo-ptr(BT)}$]. For each model, the same three explanatory variables were used to estimate the effect sizes in the bulk testis data: (i) human-macaque effect size in spermatocytes [$b_{hsa-mml(PS)}$], (ii) human-macaque effect size in spermatids [$b_{hsa-mml(PS)}$], and (c) PRE-POST effect size in combined mouse cell type dataset [$b_{PRE-POST}$]. Three models listed in the following:

$$E_{hsa-mml(BT)} = a + b_{hsa-mml(PS)} * E_{hsa-mml(PS)} + b_{hsa-mml(RS)} * E_{hsa-mml(RS)} + b_{PRE-POST} * E_{PRE-POST} + \epsilon \quad (2.2)$$

$$E_{hsa-ptr(BT)} = a + b_{hsa-mml(PS)} * E_{hsa-mml(PS)} + b_{hsa-mml(RS)} * E_{hsa-mml(RS)} + b_{PRE-POST} * E_{PRE-POST} + \epsilon \quad (2.3)$$

$$E_{ggo-ptr(BT)} = a + b_{hsa-mml(PS)} * E_{hsa-mml(PS)} + b_{hsa-mml(RS)} * E_{hsa-mml(RS)} + b_{PRE-POST} * E_{PRE-POST} + \epsilon \quad (2.4)$$

2.3.5.2 Studied using branch-length correlations

The second comparative way in which I tested the relative importance of bulk testis vs cell-autonomous convergence was through correlations of single-male primate branch-lengths between two datasets. The reasoning behind this analysis was to test if the cell-autonomous effects observed in human, macaque, and mouse spermatid and spermatocyte cell types goes beyond this specific phylogeny and is able to explain the convergence in other possible three-species phylogenies with the same topology [((single-male primate, multi-male primate), multi-male rodent);] which can be formed with the available species in the bulk testis dataset. I tested this by forming three different such trees in the bulk testis dataset, namely:

1. Human-macaque-mouse
2. Human-chimpanzee-mouse
3. Gorilla-chimpanzee-rat.

Gene expression of these possible phylogenies from the bulk testis data was compared to the only available one for the spermatocyte and spermatid datasets, the phylogeny #1 shown above. Through these available three-species trees for bulk testis and germline dataset, a neighbor-joining tree was constructed for each common gene between them, and the length of the single-male primate branch (SMBL) for both trees were recorded. Then these set of SMBL values per gene were compared using Spearman correlation. This was repeated for all three pairwise comparisons. Additionally, for each comparison, a bootstrap confidence interval for the Spearman's ρ values were calculated by repeated sampling of available genes for $N = 10^5$ iterations. Phylogenies used through these pairwise comparisons are summarized by referring to the phylogeny numbers from the above list, are summarized in **Table 2.3** and by schematic representations in **Figure 3.9**

Table 2.3: Summary of the tree sets used for the branch length based comparative analysis of bulk testis versus cell-autonomous convergence

Comparison set	Tree used in Bulk testis	Tree used in Spermatocyte/Spermatid
Set1	Tree #1	Tree #1
Set2	Tree #2	Tree #1
Set3	Tree #3	Tree #1

2.3.6 Developmental comparisons

I utilized mainly the mouse and macaque testis development datasets for this part. To see the time points of mouse or macaque testis development to which adult primate bulk testis samples showed maximum correlations, I first determined the genes which show significant Spearman correlation with age for mouse macaque testis development datasets separately. Then using only genes that change with age, loess

regression models with $\log_2(t)$ were constructed per gene to interpolate gene expression using equally separated time points (N=30) through testis development where t is the age of mouse/macaque. Finally, correlations between each interpolated time point and the adult gene expression from bulk testis data were calculated and the time point to which each bulk testis sample showed the maximum correlation (will be referred to as peaks) were recorded. Differences between the peaks for single- and multi-male primates and as well as human and chimpanzee were calculated using in a similar permutation type approach as detailed in **Section 2.3.1**. This peak-time comparison was done separately each time using mouse and macaque testis development as the reference point. As a tissue control, the same analysis was repeated for the adult primate brain and mouse brain development datasets.

2.3.7 Gene clustering & GO enrichment

To group the genes according to their expression profiles through human, macaque, and mouse testis development, I used k-means the algorithm. Since the k-means algorithm is a heuristic method and is affected by the random starting positions of centroids, I increased both the starting number of centroids and the maximum number of iterations allowed until the algorithm converges to 500. Inspecting the average silhouette scores of trials with $k = \{4, \ldots, 14\}$, I decided on k=6 as a compromise between the size of individual clusters and the within clusters variance.

I run GO BP enrichment on selected clusters (see **Section 3.8**) by using the convergent genes in the particular cluster as the foreground and non-convergent genes from out of that particular cluster as the background set of genes. I used R package `topGO` (Alexa et al., 2019) with the algorithm "parentChild", so as to avoid the redundancy caused by the hierarchical structure of gene ontology annotation (Grossmann et al., 2007).

2.3.8 Transcription factor binding site enrichment

To search for putative transcriptional regulators of the selected clusters (see **Section 3.8**), I first retrieved 2000 bp upstream sequences from the TSSs of the human

genes that are in that particular cluster, using the `biomaRt` (Durinck et al., 2009). Then I retrieved positional weight matrices for human TFs using the `MotifDB` R package (Shannon et al., 2019). These two pieces of information used as inputs for the TRAP software (Roider et al., 2007) that calculated binding affinity of given TF matrices for gene sequences. Finally, to obtain the enrichment information per TF, PASTAA software (Roider et al., 2009) was used. The foreground and background set of genes used for enrichment with PASTAA were the same as in GO enrichment part detailed above. `MotifDB` is a comprehensive database for TF matrices and therefore contains information coming from different individual databases (e.g. TRANSFAC, jaspar, HOCOMOCO, *etc.*). As result of this, there are duplicate TF matrix entries coming from different databases. To de-duplicate the results of PASTAA, I iterated over all matrices available for a single TF and only retained the one with the most significant nominal p-value for enrichment. Following the de-duplication, resulting nominal p-values were corrected for multiple testing using the "Benjamini-Hochberg" method (Benjamini et al., 1995).

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Evaluation of normalization approach

To evaluate the efficiency of normalization approaches used when creating combined datasets, I inspected how respective samples cluster using Principal Components Analysis and hierarchical clustering using the UPGMA method. Both the combined primate dataset and the combined mouse cell type dataset show no signs of batch effect related to data source as samples cluster according to their biological properties (species phylogeny/cell type origin) but not data source (**Figures 3.1 and 3.2**).

Moreover, I tested each gene present in the combined adult primate bulk testis dataset for data source versus species effects using two-way ANOVA. Across 7,305 available genes, 4,081 (55.87%) showed significant species effect, whereas only 5 genes (0.07%) showed significant data source effect (**Figure 3.3**). These results together indicate that data normalization approach taken here is effective for removing major sources of confounding factors related to data source and retains biologically relevant information.

3.2 Bulk testis transcriptomes reflect mating strategies

Looking at the phylogeny of the primate species included in this thesis (**Figure 2.1**), one might expect that single- and multi- male mating strategies have evolved convergently since the adjacent species in the phylogeny do not share the same mating strategy. If this effect is not traceable in the bulk testis transcriptomes, it would be expected for the gorilla and the macaque to show no difference in similarity to human

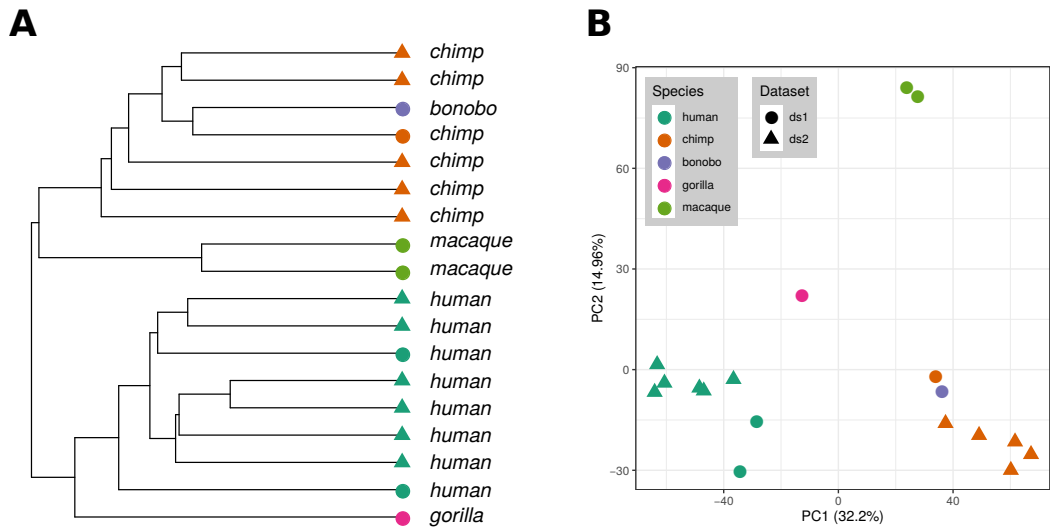

Figure 3.1: Quality control of the combined adult primate bulk testis dataset. **(A)** Dendrogram of UPGMA hiearchical clustering of all individuals. **(B)** PCA, plot of first two components. Percent variation explained by each component shown in respective axis labels. Labels 'ds1' and 'ds2' refer two data sources (Brawand et al., 2011) and (Khaitovich et al., 2005), respectively.

or chimpanzee. If there is a mating type effect on the bulk testis transcriptomes however, gorilla, a single-male species, should show a higher correlation to human when compared to chimpanzee. The opposite should be true for macaque: it should show a higher correlation to chimpanzee as opposed to human.

To test the existence of such a relationship, I first identified differentially expressed genes between human and chimpanzee, treating them as representatives for single- and multi-male mating respectively. Since they are the two closest related species in this phylogeny, it would be expected that the genes which are differentially expressed between the two species to be related to their different mating strategies. This yielded n=4295 DEGs between human and chimpanzee. Then, I used the DEGs in calculation of Spearman correlation to compare the gorilla and the macaque bulk testis transcriptomes to those of human (n=8) and chimpanzee (n=7) individuals. The macaque showed a significantly higher correlation to the chimpanzee than to human, and the reverse was true for the gorilla (permutation test $p < 0.004$, **Figure 3.4**).

I then further added the two rodents included in the analyses, mouse and rat, to this

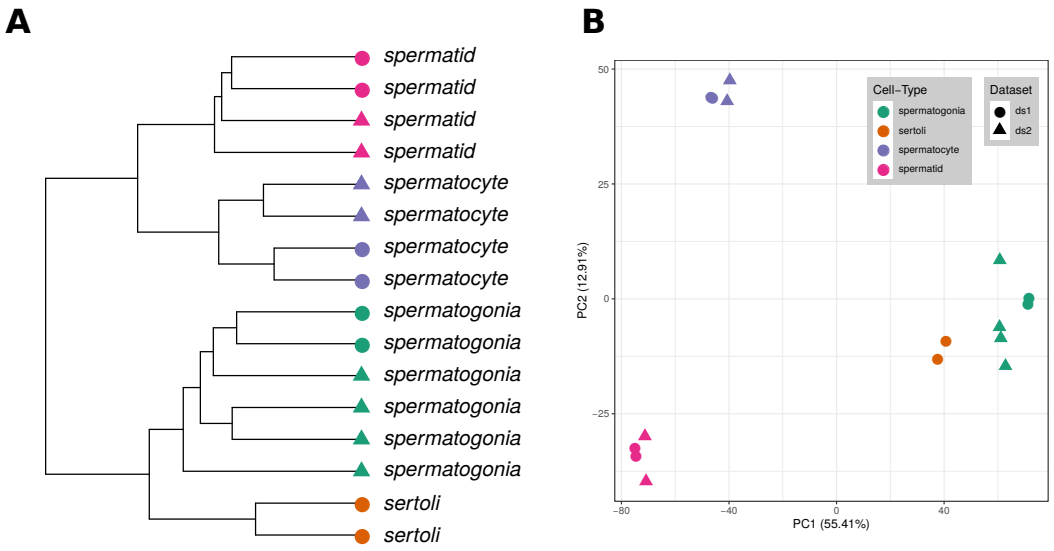

Figure 3.2: Quality control of the combined mouse cell type dataset. **(A)** Dendrogram of UPGMA hiearchical clustering of all cell types. **(B)** PCA, plot of first two components. Percent variation explained by each component shown in respective axis labels. Labels 'ds1' and 'ds2' refer two data sources (Chalmel et al., 2007) and (Namekawa et al., 2006), respectively.

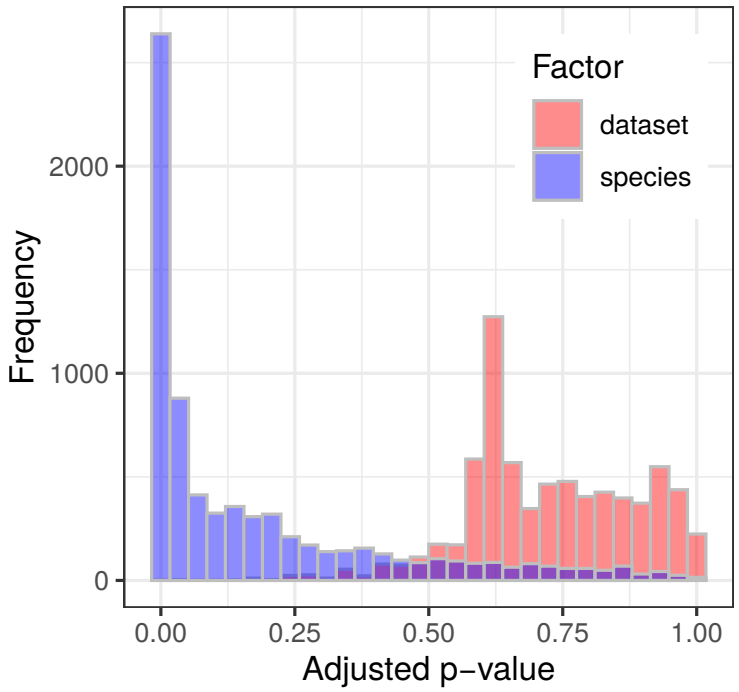

Figure 3.3: Distribution of adjusted p-values related to batch effect two-way ANOVA runs.

comparative correlation analysis. Since the data for those two species were created using a genome annotation covering orthologous genes for all the mammals (primates + rodents), it contains fewer genes because there are less 1:1 orthologous genes between all the species than there are for only primates. Instead of using the same, limited version of the primate samples, I carried out the above-mentioned comparisons with genome annotation A, then simply added the rodents which are only quantified using the GTF file B to the comparisons later. Since there are fewer genes that are common for both primates and rodents, the comparisons including rodents were limited to only n=2847 DEGs. Both mouse and rat showed significantly higher correlations to the chimpanzee than human (permutation test $p < 0.001$, **Figure 3.4**), as expected from their multi-male mating strategy.

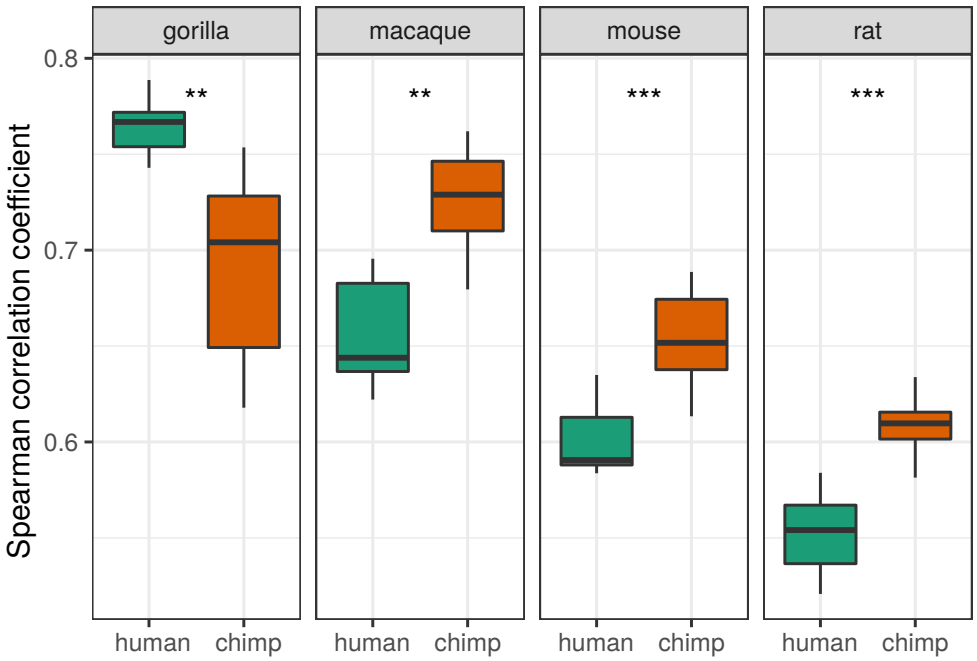

Figure 3.4: Transcriptome wide correlations of gorilla, macaque, mouse and rat to human versus chimpanzee testis transcriptome profiles. Correlations were calculated across genes that are differentially expressed between human and chimpanzee. [**(**):** $p < 0.005$, **(***):** $p < 0.001$].

3.3 Gene expression convergence in bulk testis transcriptomes

While seeing the general trend following species' mating strategies was indicative of convergence at the bulk testis level, I wanted to test individual genes for signs of convergence. For this, I opted for a formal statistical framework, EVE (see **Section 2.3.2**).

To use EVE for testing the convergent evolution of bulk testis transcriptomes according to different mating strategies of the species, I selected human and gorilla to be tested as the branches for which the additional optimum is present. Out of n=7305 available genes in the combined primate adult testis dataset, 2143 (29.3%) showed significant convergence according to mating strategy (grouping species with

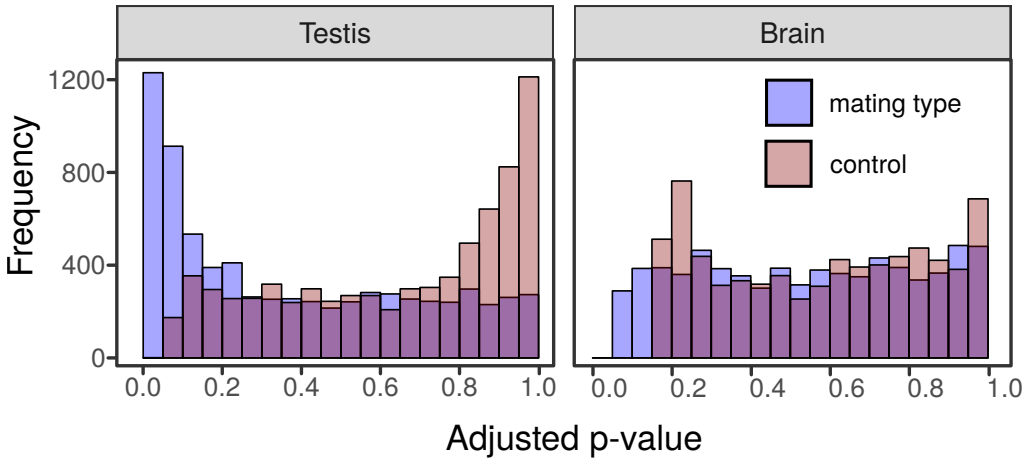

Figure 3.5: Adjusted p-value distributions of convergent evolution tests with EVE for the testis and brain. Blue shaded histograms show the p-value distribution for the tests of mating strategy-related convergent evolution [(human-gorilla) vs. (chimp-macaque)], whereas the red shaded histograms show those of controls [(human-macaque) vs. (chimp-gorilla)].

same mating strategies together). As a control, I also tested the human-macaque (and therefore chimp-gorilla) convergence, and therefore testing the existence of gene expression patterns that group species with different mating strategies together. This yielded only 175 significant genes out of 7305 genes tested.

To be able to say that these convergent patterns of bulk-tissue expression are specific to testis, I also utilized adult primate prefrontal cortex bulk tissue transcriptomes in the same type of EVE runs. In contrast to the widespread convergence between human-gorilla (and chimpanzee-macaque) bulk testis transcriptomes, there were only 289 positive genes for the brain, out of 7212 tested. The strength of convergent evolution seen for bulk testis and the lack thereof for the brain together show that convergence seen in adult bulk testis transcriptomes of primates is related to their different mating strategies.

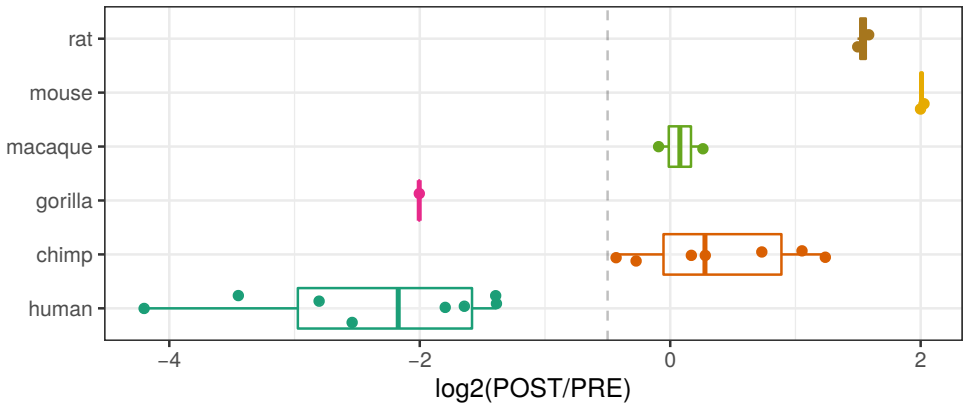

Figure 3.6: Results of cell type deconvolution analyses. Boxplots show the distrubution of log2(POST/PRE) values, or relative contributions of POST versus PRE cell type profiles to bulk testis transcriptomes of species shown in the plot.

3.4 Effect of cell type composition change on bulk testis transcriptomes

It has been previously shown that the major factor of the bulk testis convergence that is observed between human-gorilla and chimpanzee-macaque pairs is in fact the differences in relative cell type composition of single- and multi-male species‘ testicles (Saglican, 2018). This was tested with deconvolution analysis using linear regression and mouse cell type data. To reproduce the observed effect of cell type composition, I used the same two mouse cell type datasets (Chalmel et al., 2007; Namekawa et al., 2006) which together comprise 4 spermatid, 4 spermatocyte, 6 spermatogonia and 2 sertoli cells (8 PRE + 8 POST). Then employing a deconvolution analysis using linear regression, I predicted the relative POST vs PRE cell type abundances for each individual sample present in the adult bulk testis mammal data across N=4724 genes shared between two datasets. Multi-male species showed consistently higher POST/PRE ratios when compared to single-male species, human and gorilla (**Figure 3.6**). The differences of POST/PRE ratios between single- and multi-male species were significant (MWU $p = 2.1 \times 10^{-6}$). The distribution of POST/PRE ratios across species show significant convergence according mating strategies (EVE $p = 0.00061$ when tested with only primates, $p = 0.00079$ when tested with primates + rodents)

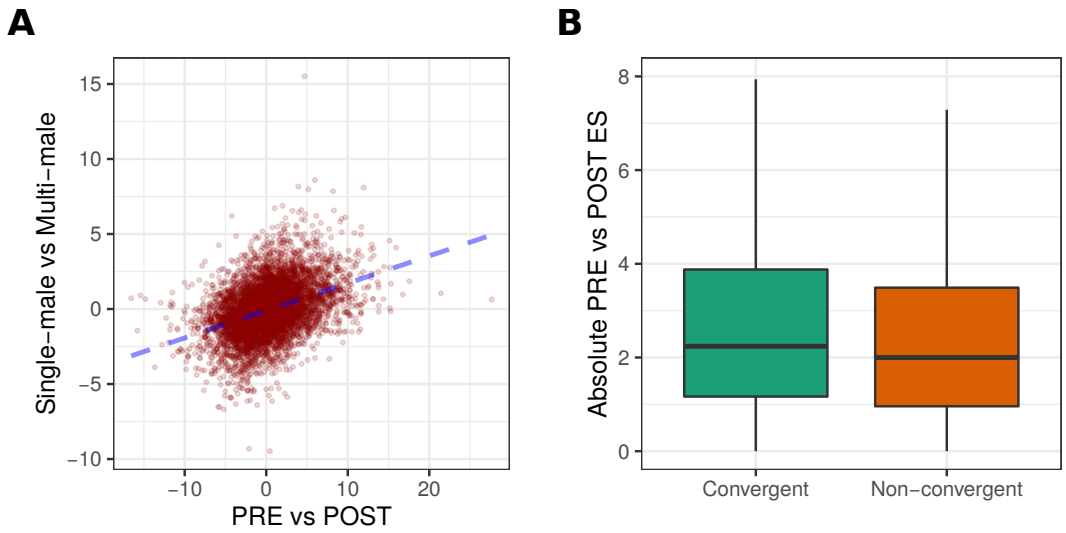

Figure 3.7: Contribution of cell type composition changes to bulk testis convergence. **(A)** Relationship between mating strategy effect size and cell type effect size. **(B)** Distribution of absolute cell type effect sizes accross convergent and non-convergent genes. N=6738 genes for both analyses.

To assess the importance of cell type composition changes on the mating strategy effects seen on the bulk testis transcriptomes, I then checked the relationship between the POST vs PRE effect sizes in the combined mouse cell type dataset and mating strategy effect sizes in the combined adult primate bulk testis dataset. Two values were correlated (Spearman's Rank Correlation, $\rho = 0.40, p < 10^{-15}$, **Figure 3.7A**). Following this logic, I also compared the absolute POST vs PRE effect sizes between genes that show convergence and those who do not. Convergent genes have significantly higher POST vs PRE effect sizes when compared to non-convergent genes. (MWU $p < 10^{-7}$, **Figure 3.7B**)

These findings show that cell type ratio shifts contribute immensely to the convergent evolution of gene expression patterns seen in primate bulk testis transcriptomes that reflect mating strategy convergences. Cell-autonomous changes, which are evolutionary gene expression shifts between respective species' cell types, could be hypothesized as another contributing factor. The following sections include analyses and discussions regarding the importance of cell-autonomous effects on the bulk testis gene expression convergence.

Table 3.1: Summary of the results of EVE runs

EVE run	# of sig. genes	# of tested genes	% sig. genes
Bulk testis	2143	7305	29.3
Testis control	175	7305	2.4
Brain	289	7212	4.0
Brain control	0	7212	0.0
Spermatocyte	358	7135	5.0
Spermatocyte control	248	7135	3.5
Spermatid	598	7061	8.5
Spermatid control	354	7061	5.0

3.5 Gene expression convergence in Spermatocytes and Spermatids

Since the information coming from the bulk testis dataset is the sum of all the mRNA present in the whole tissue, it does not provide information regarding whether the individual cell type transcriptomes of different species have undergone convergent changes, i.e., cell-autonomous changes. To identify and measure possible cell-autonomous effects contributing to the bulk testis convergence, I analyzed a cell-specific gene expression data (Lesch et al., 2016) of pachytene spermatocyte and round spermatid (both are POST) cell types coming from 3 humans, 2 macaques and 2 mice. Although this data includes only a subset of what is available in the bulk testis dataset, it is still a promising phylogeny in that it includes two primates (one single-male, human; one multi-male, macaque) and one rodent (mouse) therefore allowing for tests of convergence. Changes that make the mouse appear closer to the macaque can potentially include mating strategy-related convergence, since phylogenetically mouse is equally distant to human and macaque.

To identify convergent evolution signals in cell-autonomous fashion, I again used EVE for both the spermatocyte and the spermatocyte data, separetely. This time, mating strategy tests were run to check macaque-mouse vs human grouping in the phylogeny, and the control tests were run to human-mouse vs macaque grouping. Out

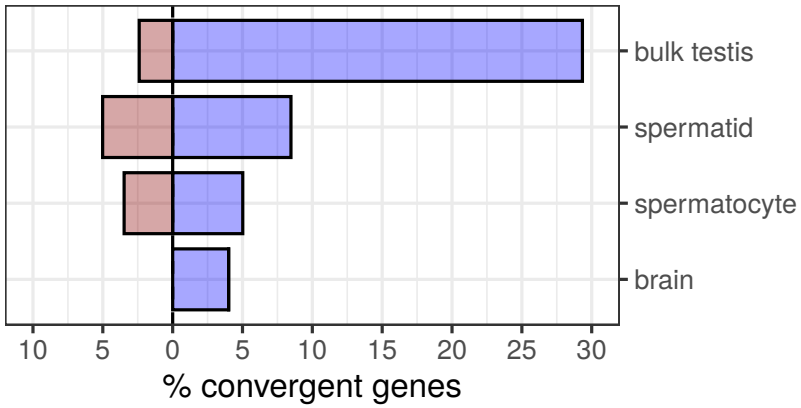

Figure 3.8: Percentage of convergent genes among the bulk testis, spermatid,spermatocyte and the brain datasets. Blue and red bars represent results for mating strategy and control tests, respectively. (As in **Figure 3.5**)

of 7,135 genes tested for spermatocyte, 358 (5%) showed significant convergence for mating strategy, whereas there were 248 (3.5%) positive genes for the control test. Out of 7,061 genes tested for spermatid, 598 (8.5%) showed significant convergence for mating strategy, whereas this number was 354 (5%) for the control test. The percentage of convergent genes across the bulk testis, spermatocyte, spermatid, and the brain datasets can be seen together in **Figure 3.8 and Table 3.1** for the comparison of strength of convergent effects.

It is clear that the percentage of convergent geens seen in spermatocyte and spermatid transcriptomes is low when compared to bulk testis. However it is possible that this difference stems partly from small sample size in species and individuals for the cell type specific datasets. In fact, as expected from this speculation, I find only 14.9% of convergent genes in the bulk testis when I use EVE on the human-macaque-mouse phylogeny using the (Brawand et al., 2011) subset of the mammal bulk testis dataset. This, then, suggest that the small size of the cell type specific dataset might indeed cause a low-power setting in which it is harder to detect convergent signals when compared to the primate bulk testis testing scenario. It also highlights the possible importance of cell-autonomous effects on the primate bulk testis gene expression convergence, at least for the spermatids.

3.6 Bulk testis versus cell-autonomous convergence

Following the observation of the existence of cell-autonomous convergence, although in modest amounts when compared to bulk testis convergence, I wanted to use several comparative donwstream analyses to assess the relative importance of cell-autonomous versus bulk testis convergence. To achieve this I used two alternative approaches, explained in the following two sections.

3.6.1 Multiple regression

To assess the significance of modest convergence seen in cell-autonomous level, I analyzed the adult bulk testis mammal dataset and the cell type-specific dataset (spermatocyte/spermatid data) in a comparative manner using multiple regression. I built three different models in which I estimated three different effect sizes representing single-male and multi-male bulk testis gene expression divergence: (a) human - macaque, (b) human - chimpanzee and (c) gorilla - chimpanzee effect sizes. I used the same three explanatory factors for all three models: (1) human - macaque effect size in spermatocyte, (2) human - macaque effect size in spermatid and (3) POST vs PRE effect size. POST vs PRE effect size (factor 3) was the only reliable estimator in that it was the only factor that had significant p-value consistently across the three models tested (**Tables 3.2- 3.4**). Results presented in aforementioned tables suggest that the human-macaque divergence in spermatocytes and spermatids is not able to estimate single-male vs multi-male divergence in the bulk testis gene expression other than human vs macaque divergence. In other words, human vs macaque divergence in spermatocytes and spermatids does not go beyond this specific phylogeny.

3.6.2 Branch length correlations

In this part, I designed a branch length-based comparative approach leveraging the fact that both the bulk testis and the cell type-specific data can yield a three species phylogeny in which there are one single-male primate, one multi-male primate and a multi-male rodent. Overall, I used three different species combinations (1) human

- macaque - mouse, (2) human - chimpanzee - mouse, (3) gorilla - chimpanzee - rat. Clearly, only the tree #1 was available for the spermatocyte/spermatid data, whereas all three trees were used for the bulk-testis data. If there is mating strategy-related convergence between the macaque and the mouse in spermatocyte and spermatid that is beyond the species identities, it would be expected that effects seen in cell type-specific data would still be comparable to effects seen in bulk testis data regardless of the species combination used in the latter. Therefore, I would need to compare the convergence in the cell type-specific data to those of bulk testis data for the three species combinations available for the bulk testis data.

Table 3.2: Results of the multiple regression analysis, Model a

Model a	Estimate	Std. Error	t value	p-value	corr. coefficient
(Intercept)	-0.002	0.014	-0.137	0.89	
hsa-mml_ps	0.075	0.014	5.338	9.86E-08 ***	0.178
hsa-mml_rs	0.153	0.014	10.895	2.63E-27 ***	0.432
prepost	0.274	0.014	19.530	1.16E-81 ***	0.283

Table 3.3: Results of the multiple regression analysis, Model b

Model b	Estimate	Std. Error	t value	p-value	corr. coefficient
(Intercept)	0.000	0.014	0.003	0.99	
hsa-mml_ps	-0.003	0.014	-0.209	0.83	-0.124
hsa-mml_rs	0.038	0.014	2.705	0.01 *	0.086
prepost	0.335	0.014	24.013	4.01E-120 ***	0.354

Table 3.4: Results of the multiple regression analysis, Model c

Model c	Estimate	Std. Error	t value	p-value	corr. coefficient
(Intercept)	0.000	0.014	-0.028	0.98	
hsa-mml_ps	0.001	0.014	0.061	0.95	-0.070
hsa-mml_rs	0.016	0.014	1.102	0.27	0.097
prepost	0.301	0.014	21.275	4.01E-120 ***	0.324

To achieve this, for each pairwise comparison (1 vs 1, 1 vs 2, 1 vs 3), I built unrooted neighbor-joining trees (Methods) for each gene that is common between two datasets and recorded the length of the single male branch (SMBL). Then, I compared the SMBL values of the spermatocyte/spermatid and the bulk testis data using rank correlation, for each pairwise comparison between two datasets (**Figure 3.9**). Moreover, I constructed 95% confidence intervals for the calculated correlations using bootstrapping of the genes used in the comparison (N = 10,000). This was implemented separately for spermatocyte and spermatid data. Although the correlation of convergence considerably drops when the species combinations are not identical between bulk testis and the spermatocyte/spermatid data in general, all three pairwise comparisons for the spermatid yielded significant positive correlations with the bulk testis convergence (95% CI > 0)

In summary, in accordance with the EVE results, correlations between bulk testis and cell-specific data are stronger in spermatid gene expression than in spermatocyte gene expression. It is worth noting that linear model-based analyses depicted above also point in this direction. This phenomenon might be partly explained by the relative abundance of spermatid cells versus spermatocyte cells within the seminiferous tubules being high, at least in humans (Skakkebaek et al., 1973). Meanwhile, cell-autonomous convergent evolution appears considerably weaker than convergent changes that could be related to cell type ratio shifts.

3.7 Paedomorphism in single-male bulk testis transcriptomes

In a lineage with multi-male mating sytem, evolution of larger testicles can be achieved by acceleration or extended duration of progenitor germ cell divisions during the adolescence period which would result in increased sperm production and number/proportion of germline -thus increased POST:PRE ratio- cells (Montoto et al., 2012). Following this suggested mechanism, delaying or decelerating germ cell proliferation could be the mechanism to develop a single-male testis in terms of histology and gene expression.

Hypothesizing that shifts in rate or onset of development could also be seen in the

transcriptome level, I used a published mouse testis development dataset to test if adult primate bulk testis transcriptomes show differential affinity to mouse testis developmental stages in accordance with the differences in their mating system. Since the reference point here is a multi-male species (mouse) which is equally distant to human, chimpanzee, gorilla or macaque, I would expect single-male species human and gorilla to show the highest correlation to earlier stages of mouse testis development when compared to the chimpanzee or the macaque, which show a multi-male behavior. I also included the PRE and POST cell profiles in this analysis to see the differences between them in terms of the point of maximum correlation among mouse testis development.

For each of the aforementioned four species and the two cell types, I simply deter-

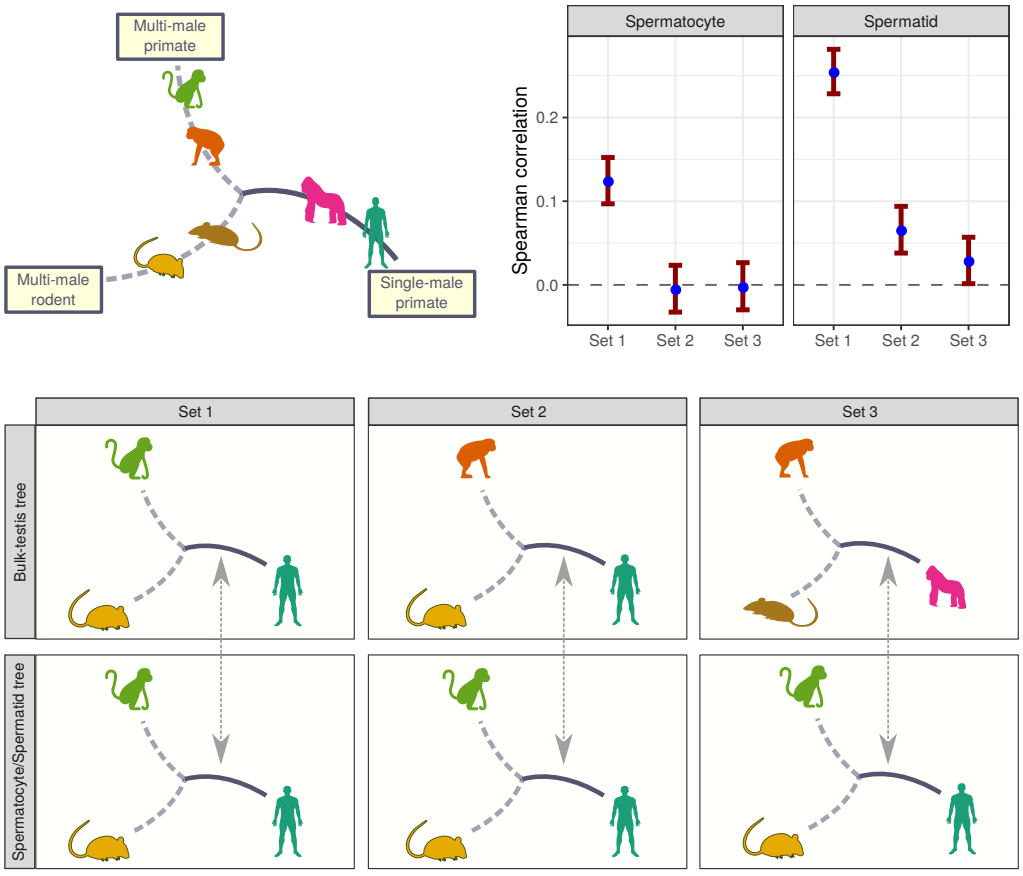

Figure 3.9: Schematic representation (bottom) and the results (top-right) of branch length based analysis of bulk testis vs. cell-autonomous convergent changes.

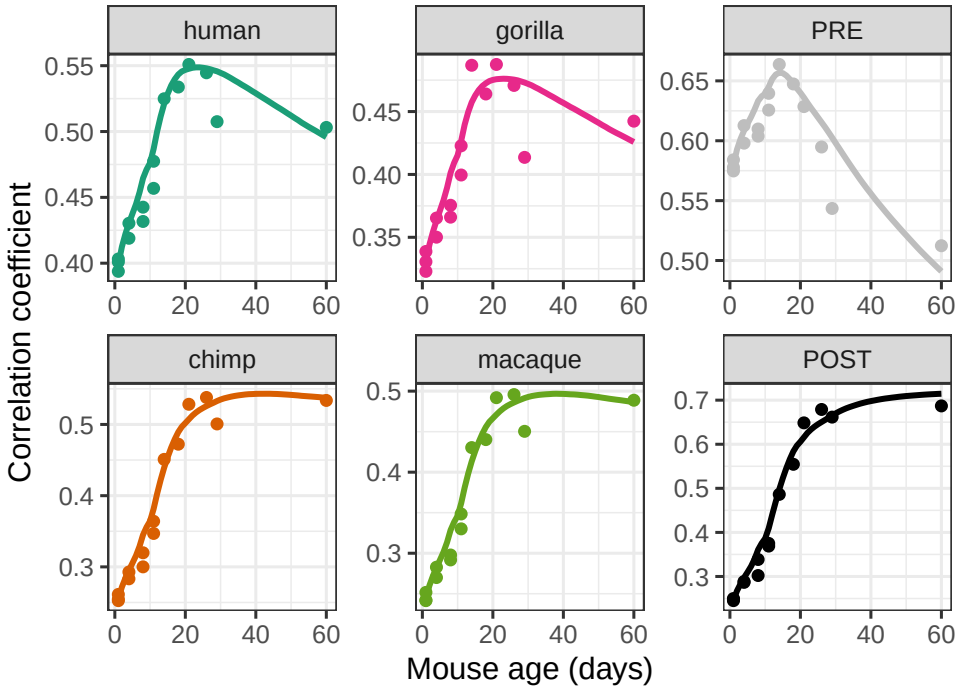

Figure 3.10: Results of developmental comprarisons using mouse testis development as the reference. Comparisons were limited to 2295 genes showed significant Spearman correlation with age in mouse testis development.

mined the age to which they show the highest correlation thoughout mouse testis development. Consistent with their nature, PRE and POST cell type profiles showed highest affinity to pre-adolescent and adult mice, respectively (**Figure 3.10**). Moreover, human and gorilla adult testes showed highest similarity to sub-adult mice (median=23 days) whereas chimpanzee and macaque showed the highest similarity to adult mice (median=41 days). Differences between the peak time points for human and chimpanzee, as well as the differences between single- and multi-male species were significant (permutation test, $p < 10^{-5}$ for both comparisons).

I also utilized a macaque testis development data to investigate this phenomenon furhter. Implementing the same analysis as with the mouse testis development part, I found a trend in the same direction ($p < 10^{-5}$, **Figure 3.11**). In summary both the mouse and the macaque testis development datasets supports the fact that adult bulk testis transcriptome profiles of single-male species, human and gorilla, are paedomorphic when compared to those of multi-male species, chimpanzee and macaque.

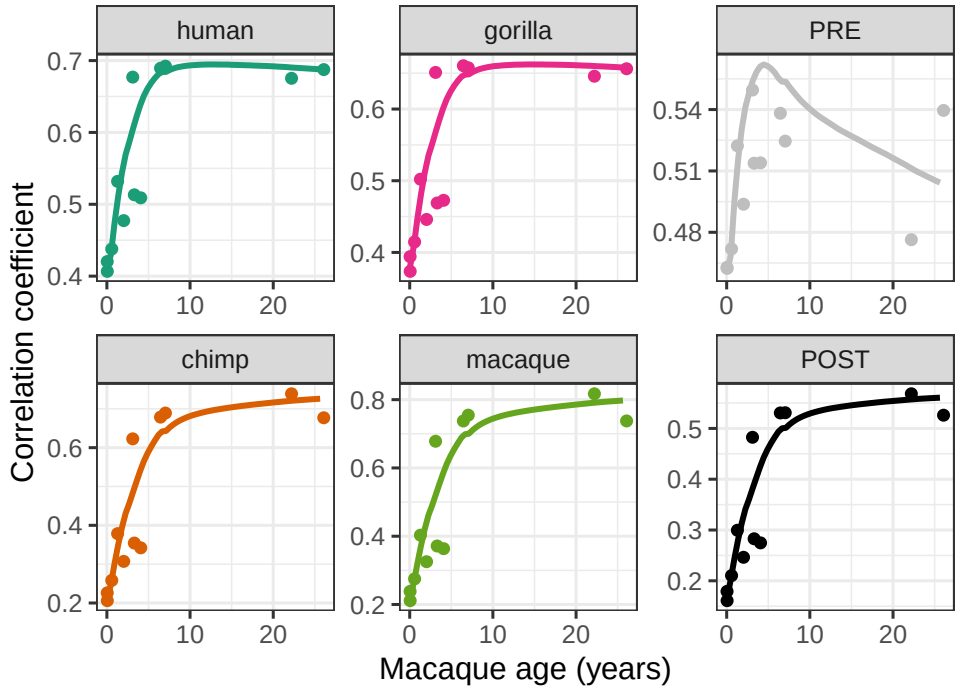

Figure 3.11: Results of developmental comprarisons using macaque testis development as the reference. Comparisons were limited to 2321 genes showed significant Spearman correlation with age in macaque testis development, respectively.

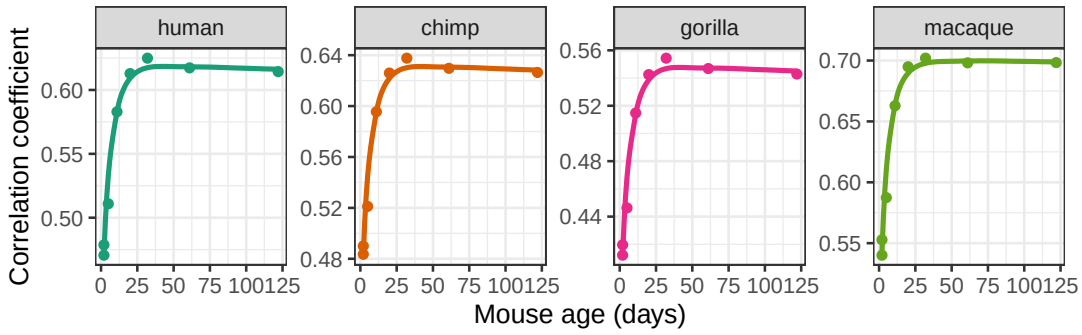

Figure 3.12: Results of mouse brain development analysis. Across mouse brain development, both the human and the gorilla (median=57 days) and the chimpanzee and macaque (median=43 days) show the most correlations to adult mice.

It has been previously reported that humans show neotenic gene expresssion relative to chimpanzee and macaque for gene sets functionally related to neuronal proceesses (Liu et al., 2012; Somel et al., 2009). Therefore it is tempting to ask whether this observed transcriptome-wide paedomorphism for human and gorilla is also present

in the brain or it is unique to testis and thus related to convergent patterns we see in bulk testis. To test this hypothesis, I utilized an additional dataset from mouse neocortex development bulk tissue, comprising 8 mice with ages newborn to 122 days. I used this dataset in the same way as described above for mouse and macaque testis development datasets, and perform the same analysis this time comparing adult human, chimpanzee, gorilla, and macaque prefrontal cortex bulk tissue gene expression to this mouse prefrontal cortex development time series. For the brain development, all adult primates showed the highest correlation to young adult or adult mice (median=57 and 43 days of age, for single-male and multi-male species, respectively; **Figure 3.12**). Contrary to testis development results, neither the difference between human and chimpanzee peak time points, nor the single- versus multi-male difference was significant. This corroborates previously reported findings showing that neotenic gene expression in the cerebral cortex is not detectable transcriptome-wide, but rather confined to specific functional processes. Therefore, the paedomorphism of human and gorilla transcriptomes is limited to testis gene expression and not seen in brain.

3.8 Searching for putative regulators of convergent expression patterns

The convergence in testis development and cell type ratio changes shown here could result from a limited number of changes in central developmental regulators, thus, changing gene expression in their targets. To identify such regulators or gene groups, I clustered the genes shared between human, macaque, and mouse testis development gene expression data into six groups using k-means. (**Figure 3.13**). Out of six total, two yielded significant enrichment in genes showing convergent evolution according to EVE analysis (Fisher's exact test, $q < 0.005$). One of the clusters that showed enrichment, was noteworthy in that genes in this cluster were enriched in multiple GO BP categories related to spermatogenesis (**Figure 3.15**), and the overall trend of gene expression for all three species was a consistent increase during testis development. Notably, genes in this cluster (N=960) also show higher expression for the chimpanzee and macaque relative to the human and gorilla in adult bulk testis primate dataset (**Figure 3.14**).

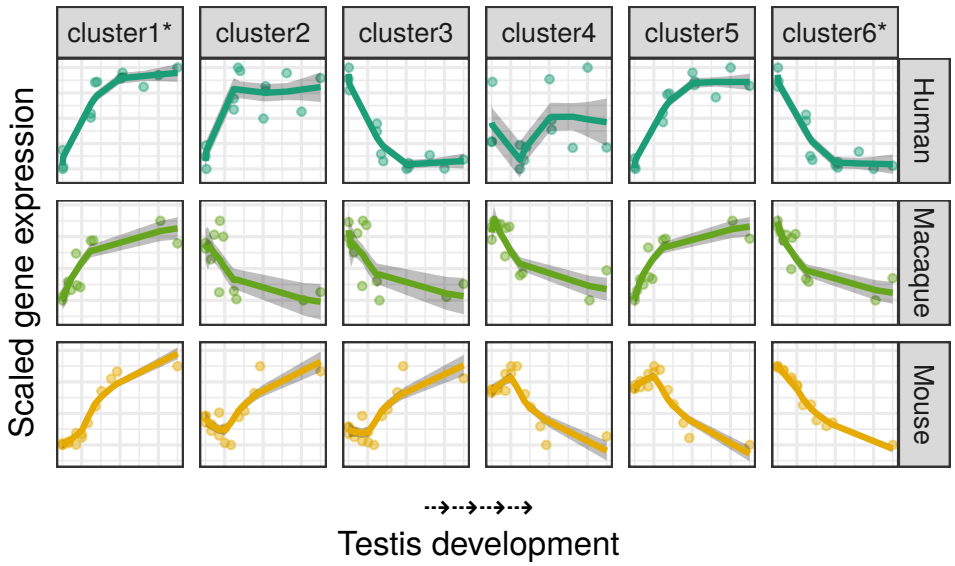

Figure 3.13: Patterns of gene expression throughout human, macaque and mouse testis development across six k-means clusters. **(*):** Clusters enriched for convergent genes.

To identify any putative transcription factors that regulate the expression of genes in cluster #1, I performed TFBS analysis using candidate promoter regions from the human sequences of the respective genes. Out of 475 transcription factors tested for binding site enrichment, 145 displayed enrichment in this cluster (BH corrected Fisher's exact test, at $q < 0.10$). Out of these, 32 (22.9%) showed convergent evolution according to EVE analyses (**Figure 3.16A**). These 32 transcription factors include three that previously reported to take role in controlling organ growth: *TEAD1, MAX*, and *MXI1*. *TEAD1* is the ultimate target of the Hippo pathway, which has a role in the fine tuning of the proliferation/differentiation balance and thus controlling organ growth (Watt et al., 2017; Yu et al., 2013). TEAD1 also reported to involve in a complex with MAX such that each protein acts as the others co-activator (Lin et al., 2017). Moreover, MAX forms a complex with MYC to promote growth and proliferation (Nair et al., 2003). MXI also binds to MAX so that there are fewer MAX/MYC complexes to induce proliferation (Schreiber-Agus et al., 1998; Zervos et al., 1993). Expression patterns of *TEAD1* and *MAX* (**Figure 3.16A**) might seem contradictory to the functional roles depicted above at the first glance. However, this seemingly opposite pattern of gene expression becomes more understanble when taking complex

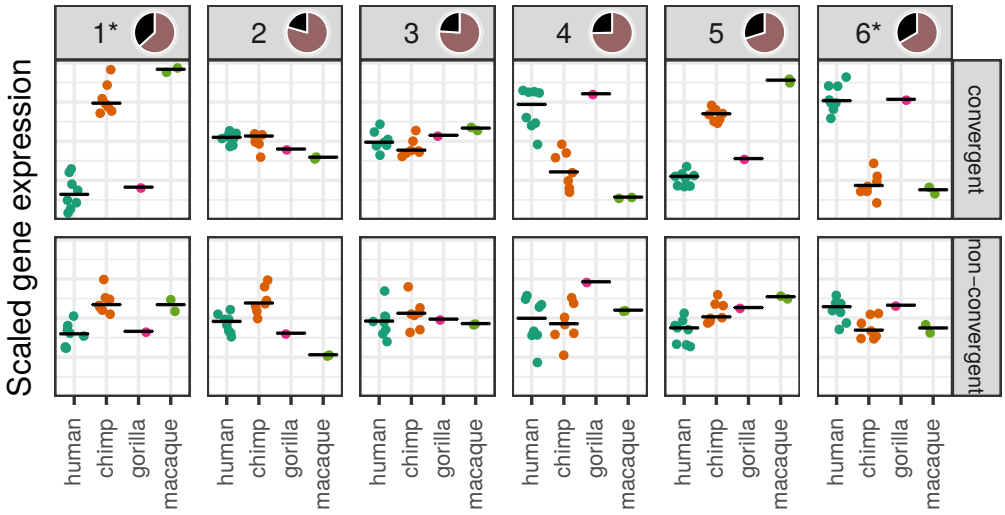

Figure 3.14: Patterns of average gene expression throughout adult primate bulk testis transcriptomes across six k-means clusters. Top row shows gene expression for genes that show convergent evolution in that particular cluster whereas bottom row show that of non-convergent genes. Proportion of convergent genes in a given cluster is given as the dark portion of the pie charts located in the pane labels. **(*):** Clusters enriched for convergent genes.

relationships with their targets into account. Both factors are reported to inhibit their targets' expression if they are overexpressed without their cofactors (YAP1/TAZ and MYC, respectively) for the aforementioned pathways or contexts (Gu et al., 1993; Watt et al., 2017). Therefore it is not easy to unravel a possible cellular mechanism that might have resulted in the apparent convergent evolution of testis sizes between human-gorilla and chimpanzee-macaque pairs with the information and tools available for this study. Interestingly, it has been previously reported that the regulation of the Hippo pathway directly regulates testis growth in atlantic salmon (Kjærner-Semb et al., 2018). Overall, the promising preliminary findings of this thesis together with the aforementioned observation in salmon testis, clearly points to the Hippo pathway as a candidate and a possible area for further research to improve this study.

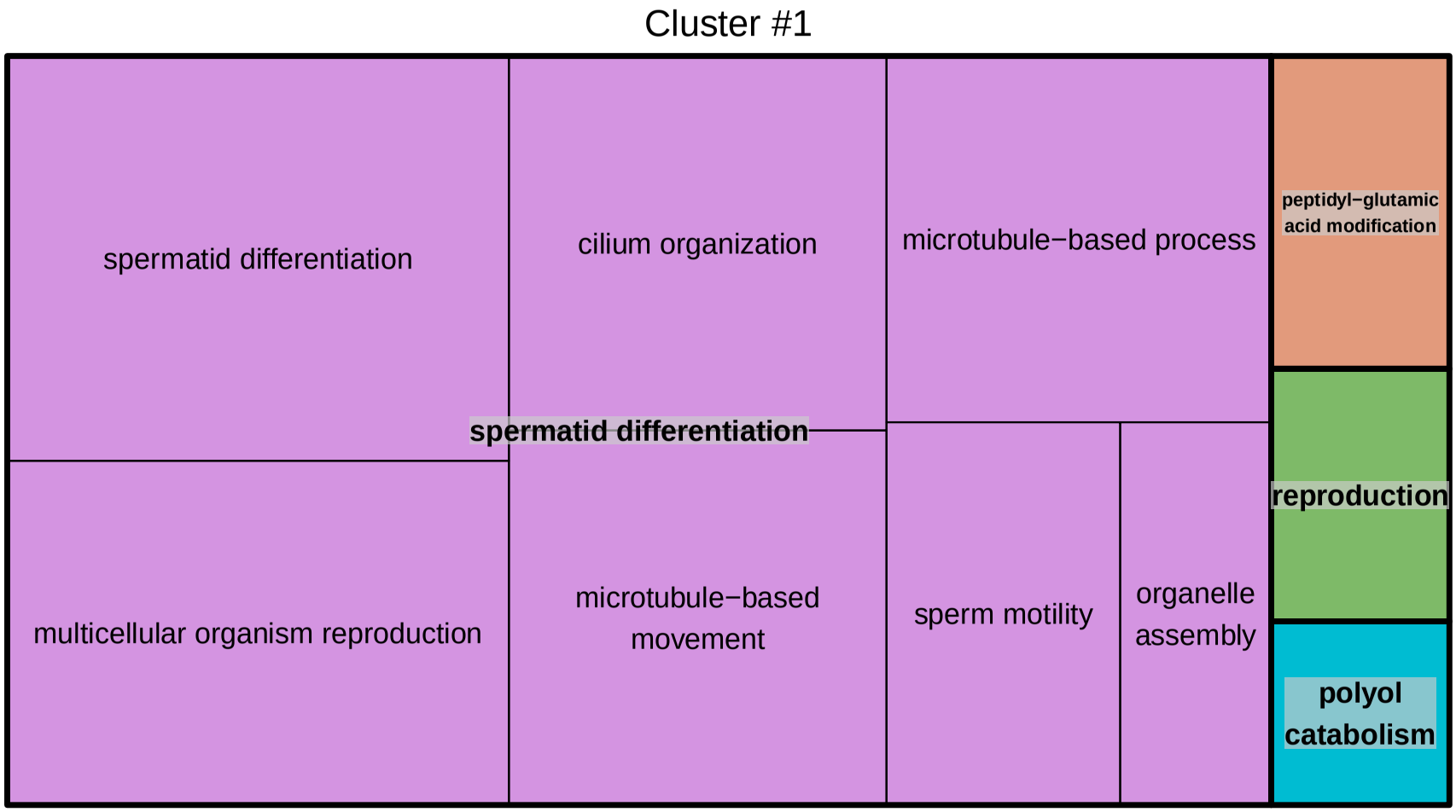

Figure 3.15: Enriched GO terms in cluster #1 summarized by REVIGO

Among the TFs that show enrichment for possible binding sites among the candidate human promoter sequences for the genes in cluster #1 but do not necessarily show convergent gene expression among the four primates, there were also several interesting factors that are worth discussing: *RFX1, RFX2*, and *DLX5*. These factors are particulary interesting because of their reported roles in regulation of spermatogenesis and steroid hormone synthesis in response to luteinizing hormone (LH). RFX2, is one of the major regulators of spermatogenesis and forms a heterodimer with RFX1 (Kistler et al., 2015; Wu et al., 2016), whereas DLX5 regulates testicular steroidogenesis together with GATA-4, which then binds to the *STAR* gene promoter to induce steroid synthesis (Nishida et al., 2008).

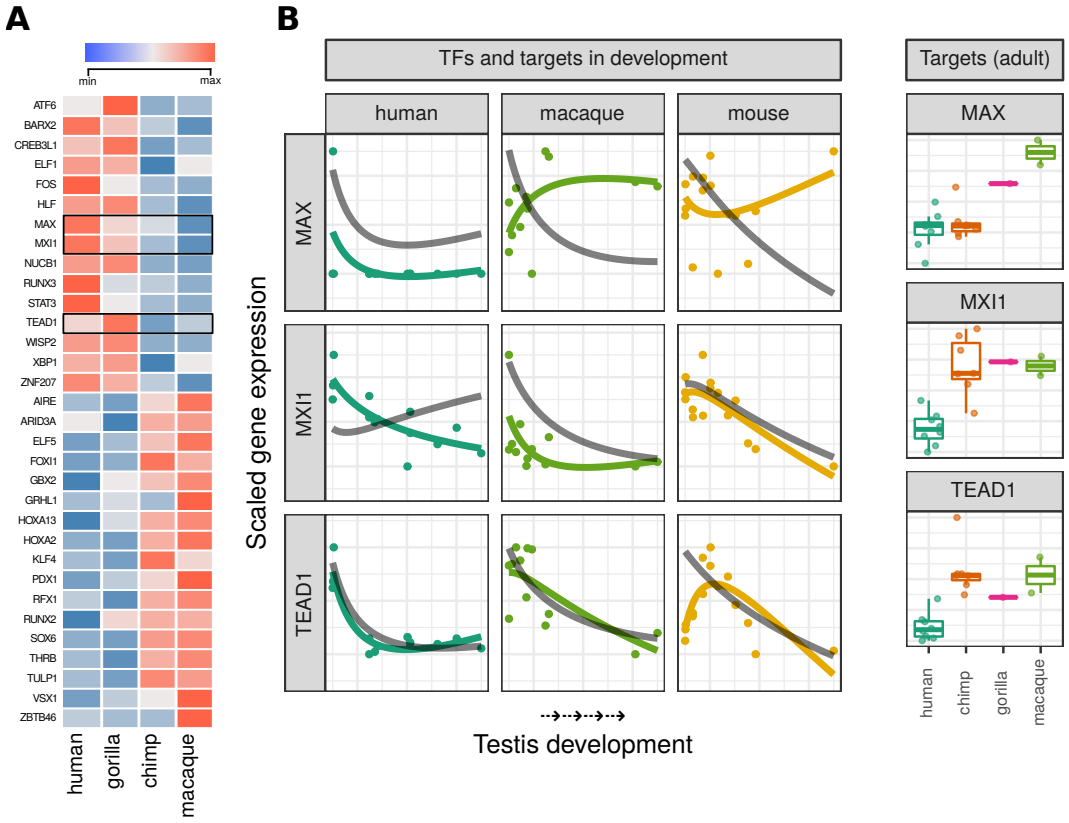

Figure 3.16: Results of transcription factor binding site enrichment analysis. **(A)** Bulk testis gene expression of enriched TFs in cluster #1 and showing convergent gene expression across human, chimpanzee, gorilla and macaque. **(B)** Gene expression profiles of selected TFs (marked with rectangles on pane A) and their target genes, throughout testis development (left) and adult bulk testis transcriptomes (right).

CHAPTER 4

CONCLUSION

Here in this study, I corroborate that the convergent evolution of testis anatomy, which could be explained by mating strategy differences among mammalian species, is also detectable in bulk testis transcriptomes of mainly primates, by implementing novel analyses regarding the possible histological or molecular effectors of the said convergence. Overall, the findings -of both the repeated and novel analyses presented here- can be summarized in three main conclusive points.

1. Cell-autonomous convergent changes in gene expression of spermatogenic cell types can be detected, but only weakly.
2. Instead, convergent evolution patterns observed transcriptome-wide for bulk testis are explained to a large extent by convergent cell type ratio changes among the species analyzed. In other words, convergent changes in relative abundances of already specialized cell types (which we group as PRE and POST here) in response to changing levels of selection pressure still remains to be the main factor on the convergent evolution bulk testis transcriptomes, among others that we are yet able to measure with future transcriptome data.
3. We observed that human and gorilla bulk testis transcriptomes appear paedomorphic relative to those of chimpanzee and macaque (or hypermorphic in the latter pair relative to the former). These paedomorphic patterns also reflect the anatomical states of the testes size of respective species. Thus, although not measured or analyzed directly here, heterochrony, or change in rate or timing of development, can be the underlying developmental mechanism behind the observed shifts in cell type ratios and gene expression convergence in bulk testis.

Despite the promising findings reported here, this study's technical limitations should be addressed in a clear manner. The limitations mainly stem from the fact that we do not have first-hand access to biological data but instead use published data, and also that postmortem samples from chimpanzees and gorillas are hard to obtain to begin with. These limitations and future prospects are summarized below in four main points:

1. There are only two single-male species analyzed here, with the gorilla represented by only one biological sample. To assess the veracity of the patterns we report, additional species should be incorporated into the analyses. This would require high-quality testis samples collected from species with unambiguously documented mating strategies.

2. We use data obtained from independent biological samples for the analyses comparing cell-type composition versus cell-autonomous convergent effects. Ideally this type of analysis should be done with bulk testis and pooled cell RNA-Seq data obtained from the same set of individuals.

3. All the analyses here are confined to 1:1 orthologous genes among the species. This prevents us from analyzing expression changes in lineage-specific genes. These limitations may also have resulted in an overestimation of the proportion of genes with convergent patterns mirroring mating strategy differences.

4. Results of functional enrichment analyses and the identified candidate transcriptional regulators reported here should be validated using state of the art molecular techniques. A particularly interesting finding to test would be the differential regulation of the Hippo pathway between single- and multi-male species' testis using primary cell cultures.

REFERENCES

Alexa, A. and J. Rahnenfuhrer (2019). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.38.1.

Benjamini, Y. and Y. Hochberg (Jan. 1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. Smith, K. M. Roskin, et al. (2004). "Aligning multiple genomic sequences with the threaded blockset aligner". In: *Genome Research* 14.4, pp. 708–715. DOI: 10.1101/gr.1933104.

Bolstad, B. (2019). *preprocessCore: A collection of pre-processing functions*. R package version 1.48.0.

Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, et al. (Oct. 2011). "The evolution of gene expression levels in mammalian organs". In: *Nature* 478.7369, pp. 343–348. DOI: 10.1038/nature10532.

Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter (May 2016). "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5, pp. 525–527. DOI: 10.1038/nbt.3519.

Cardoso-Moreira, M., J. Halbert, D. Valloton, B. Velten, C. Chen, Y. Shao, et al. (July 2019). "Gene expression across mammalian organ development". In: *Nature* 571.7766, pp. 505–509. DOI: 10.1038/s41586-019-1338-5.

Chalmel, F., A. D. Rolland, C. Niederhauser-Wiederkehr, S. S. Chung, P. Demougin, A. Gattiker, et al. (May 2007). "The conserved transcriptome in human and rodent male gametogenesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.20, pp. 8346–8351. DOI: 10.1073/pnas.0701883104.

Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, et al. (Jan. 2015). "Ensembl 2015". In: *Nucleic Acids Research* 43.D1, pp. D662–D669. DOI: 10.1093/nar/gku1010.

Darwin, C. (1896). "Principles of sexual selection". In: *The descent of man and selection in relation to sex*. Vol. 1. D. Appleton, pp. 253–320.

Dixson, A. F. (2012). *Primate sexuality : comparative studies of the prosimians, monkeys, apes, and humans*. Oxford University Press, p. 785.

Durinck, S., P. T. Spellman, E. Birney, and W. Huber (2009). "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt". In: *Nature Protocols* 4, pp. 1184–1191.

Fietz, D. and M. Bergmann (2017). "Functional Anatomy and Histology of the Testis". In: *Endocrinology of the Testis and Male Reproduction*. Ed. by M. Simoni and I. T. Huhtaniemi. Cham: Springer International Publishing, pp. 313–341. DOI: 10.1007/978-3-319-44441-3_9.

Fodor, S. P., J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas (Feb. 1991). "Light-directed, spatially addressable parallel chemical synthesis". In: *Science* 251.4995, pp. 767–773. DOI: 10.1126/science.1990438.

Gautier, L., L. Cope, B. M. Bolstad, and R. A. Irizarry (Feb. 2004). "affy–analysis of Affymetrix GeneChip data at the probe level". In: *Bioinformatics* 20.3, pp. 307–315. DOI: 10.1093/bioinformatics/btg405.

Gillard, G. B., L. Grønvold, M. Mekrog Holen, Ø. Monsen, B. F. Koop, E. B. Rondeau, et al. (2020). "Comparative regulomics reveals pervasive selection on gene dosage following whole genome duplication". In: DOI: 10.1101/2020.07.20.212316.

Gong, T., N. Hartmann, I. S. Kohane, V. Brinkmann, F. Staedtler, M. Letzkus, et al. (Nov. 2011). "Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples". In: *PLoS ONE* 6.11. Ed. by M. Rattray, e27156. DOI: 10.1371/journal.pone.0027156.

Goodwin, S., J. D. McPherson, and W. R. McCombie (June 2016). *Coming of age: Ten years of next-generation sequencing technologies*. DOI: 10.1038/nrg.2016.49.

Grossmann, S., S. Bauer, P. N. Robinson, and M. Vingron (Nov. 2007). "Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis". In: *Bioinformatics* 23.22, pp. 3024–3031. DOI: 10.1093/bioinformatics/btm440.

Gu, W., K. Cechova, V. Tassi, and R. Dalla-Favera (1993). "Opposite regulation of gene transcription and cell proliferation by c-Myc and Max". In: *Proceedings of the National Academy of Sciences of the United States of America* 90.7, pp. 2935–2939. DOI: 10.1073/pnas.90.7.2935.

Harcourt, A. H., P. H. Harvey, S. G. Larson, and R. V. Short (1981). "Testis weight, body weight and breeding system in primates". In: *Nature* 293.5827, pp. 55–57. DOI: 10.1038/293055a0.

Harcourt, A. H., A. Purvis, and L. Liles (June 1995). "Sperm Competition: Mating System, Not Breeding Season, Affects Testes Size of Primates". In: *Functional Ecology* 9.3, p. 468. DOI: 10.2307/2390011.

Hosken, D. J. (1997). "Sperm competition in bats". In: *Proceedings of the Royal Society B: Biological Sciences* 264.1380, pp. 385–392. DOI: 10.1098/rspb.1997.0055.

Kenagy, G. J. and S. C. Trombulak (Feb. 1986). "Size and Function of Mammalian Testes in Relation to Body Size". In: *Journal of Mammalogy* 67.1, pp. 1–22. DOI: 10.2307/1380997.

Khaitovich, P., W. Enard, M. Lachmann, and S. Pääbo (Sept. 2006a). "Evolution of primate gene expression". In: *Nature Reviews Genetics* 7.9, pp. 693–702. DOI: 10.1038/nrg1940.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, et al. (Sept. 2005). "Evolution: Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees". In: *Science* 309.5742, pp. 1850–1854. DOI: 10.1126/science.1108296.

Khaitovich, P., J. Kelso, H. Franz, J. Visagie, T. Giger, S. Joerchel, et al. (2006b). "Functionality of Intergenic Transcription: An Evolutionary Comparison". In: *PLoS Genetics* 2.10, e171. DOI: 10.1371/journal.pgen.0020171.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg (Apr. 2013). "TopHat2: Accurate alignment of transcriptomes in the presence of in-

sertions, deletions and gene fusions". In: *Genome Biology* 14.4, R36. DOI: 10.1186/gb-2013-14-4-r36.

Kistler, W. S., D. Baas, S. Lemeille, M. Paschaki, Q. Seguin-Estevez, E. Barras, et al. (July 2015). "RFX2 Is a Major Transcriptional Regulator of Spermiogenesis". In: *PLOS Genetics* 11.7. Ed. by P. E. Cohen, e1005368. DOI: 10.1371/journal.pgen.1005368.

Kjærner-Semb, E., F. Ayllon, L. Kleppe, E. Sørhus, K. Skaftnesmo, T. Furmanek, et al. (2018). "Vgll3 and the Hippo pathway are regulated in Sertoli cells upon entry and during puberty in Atlantic salmon testis". In: *Scientific Reports* 8.1, pp. 1–11. DOI: 10.1038/s41598-018-20308-1.

Lesch, B. J., S. J. Silber, J. R. McCarrey, and D. C. Page (Aug. 2016). "Parallel evolution of male germline epigenetic poising and somatic development in animals". In: *Nature Genetics* 48.8, pp. 888–894. DOI: 10.1038/ng.3591.

Lin, K. C., H. W. Park, and K. L. Guan (Nov. 2017). "Regulation of the Hippo Pathway Transcription Factor TEAD". In: *Trends in Biochemical Sciences* 42.11, pp. 862–872. DOI: 10.1016/j.tibs.2017.09.003.

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, et al. (May 2008). "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis". In: *Cell* 133.3, pp. 523–536. DOI: 10.1016/j.cell.2008.03.029.

Liu, X., M. Somel, L. Tang, Z. Yan, X. Jiang, S. Guo, et al. (Apr. 2012). "Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques". In: *Genome Research* 22.4, pp. 611–622. DOI: 10.1101/gr.127324.111.

Loyau, A., M. S. Jalme, and G. Sorci (2005). "Intra- and intersexual selection for multiple traits in the peacock (Pavo cristatus)". In: *Ethology* 111.9, pp. 810–820. DOI: 10.1111/j.1439-0310.2005.01091.x.

Montoto, L. G., L. Arregui, N. M. Sánchez, M. Gomendio, and E. R. Roldan (Mar. 2012). "Postnatal testicular development in mouse species with different levels of sperm competition". In: *Reproduction* 143.3, pp. 333–346. DOI: 10.1530/REP-11-0245.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (July 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. DOI: 10.1038/nmeth.1226.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, et al. (June 2008). "The transcriptional landscape of the yeast genome defined by RNA sequencing". In: *Science* 320.5881, pp. 1344–1349. DOI: 10.1126/science.1158441.

Nair, S. K. and S. K. Burley (Jan. 2003). "X-ray structures of Myc-Max and Mad-Max recognizing DNA: Molecular bases of regulation by proto-oncogenic transcription factors". In: *Cell* 112.2, pp. 193–205. DOI: 10.1016/S0092-8674(02)01284-9.

Namekawa, S. H., P. J. Park, L. F. Zhang, J. E. Shima, J. R. McCarrey, M. D. Griswold, et al. (Apr. 2006). "Postmeiotic Sex Chromatin in the Male Germline of Mice". In: *Current Biology* 16.7, pp. 660–667. DOI: 10.1016/j.cub.2006.01.066.

Nishida, H., S. Miyagawa, M. Vieux-Rochas, M. Morini, Y. Ogino, K. Suzuki, et al. (2008). "Positive regulation of steroidogenic acute regulatory protein gene expression through the interaction between Dlx and GATA-4 for testicular steroidogenesis". In: *Endocrinology* 149.5, pp. 2090–2097. DOI: 10.1210/en.2007-1265.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramm, S. A., G. A. Parker, and P. Stockley (May 2005). "Sperm competition and the evolution of male reproductive anatomy in rodents". In: *Proceedings of the Royal Society B: Biological Sciences* 272.1566, pp. 949–955. DOI: 10.1098/rspb.2004.3048.

Ramm, S. A. and P. Stockley (Apr. 2010). "Sperm competition and sperm length influence the rate of mammalian spermatogenesis". In: *Biology Letters* 6.2, pp. 219–221. DOI: 10.1098/rsbl.2009.0635.

Rohlfs, R. V. and R. Nielsen (Sept. 2015). "Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution". In: *Systematic Biology* 64.5, pp. 695–708. DOI: 10.1093/sysbio/syv042.

Roider, H. G., A. Kanhere, T. Manke, and M. Vingron (Jan. 2007). "Predicting transcription factor affinities to DNA from a biophysical model". In: *Bioinformatics* 23.2, pp. 134–141. DOI: 10.1093/bioinformatics/btl565.

Roider, H. G., T. Manke, S. O'keeffe, M. Vingron, and S. A. Haas (Feb. 2009). "PASTAA: Identifying transcription factors associated with sets of co-regulated genes". In: *Bioinformatics* 25.4, pp. 435–442. DOI: 10.1093/bioinformatics/btn627.

Sağlıcan, E. (2018). "Transcription Evolution Among Hominids". MA thesis. Middle East Technical University.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (Oct. 1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". In: *Science* 270.5235, pp. 467–470. DOI: 10.1126/science.270.5235.467.

Schreiber-Agus, N., Y. Meng, T. Hoang, H. Hou, K. Ghen, R. Greenberg, et al. (June 1998). "Role of Mxi1 in ageing organ systems and the regulation of normal and neoplastic growth". In: *Nature* 393.6684, pp. 483–487. DOI: 10.1038/31008.

Schultz, A. H. (Nov. 1938). "The relative weight of the testes in primates". In: *The Anatomical Record* 72.3, pp. 387–394. DOI: 10.1002/ar.1090720310.

Schultz, N., F. K. Hamra, and D. L. Garbers (Oct. 2003). "A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets". In: *Proceedings of the National Academy of Sciences of the United States of America* 100.21, pp. 12201–12206. DOI: 10.1073/pnas.1635054100.

Sekido, R. and R. Lovell-Badge (2013). "Genetic Control of Testis Development". In: *Sexual Development* 7.1-3, pp. 21–32. DOI: 10.1159/000342221.

Shannon, P. and M. Richards (2019). *MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs*. R package version 1.28.0.

Short, R. V. (Jan. 1979). "Sexual Selection and Its Component Parts, Somatic and Genital Selection, as Illustrated by Man and the Great Apes". In: *Advances in the Study of Behavior* 9.C, pp. 131–158. DOI: 10.1016/S0065-3454(08)60035-2.

Skakkebaek, N. E. and C. G. Heller (Mar. 1973). "Quantification of human seminiferous epithelium. I. Histological studies in twenty-one fertile men with nor-

mal chromosome complements." In: *Journal of Reproduction and Fertility* 32.3, pp. 379–389. DOI: 10.1530/jrf.0.0320379.

Somel, M., H. Franz, Z. Yan, A. Lorenc, S. Guo, T. Giger, et al. (Apr. 2009). "Transcriptional neoteny in the human brain". In: *Proceedings of the National Academy of Sciences of the United States of America* 106.14, pp. 5743–5748. DOI: 10.1073/pnas.0900544106.

Stockley, P. (2004). "Sperm competition in mammals". In: *Human Fertility* 7.2, pp. 91–97. DOI: 10.1080/14647270410001699054.

Thavarajah, N. K., P. G. Tickle, R. L. Nudds, and J. R. Codd (Nov. 2016). "The peacock train does not handicap cursorial locomotor performance". In: *Scientific Reports* 6.1, pp. 1–6. DOI: 10.1038/srep36512.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, et al. (May 2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature Biotechnology* 28.5, pp. 511–515. DOI: 10.1038/nbt.1621.

Wang, Z. Y., E. Leushkin, A. Liechti, S. Ovchinnikova, K. Mößinger, T. Brüning, et al. (2020). "Transcriptome and translatome co-evolution in mammals". In: *Nature* November 2020. DOI: 10.1038/s41586-020-2899-z.

Watt, K. I., K. F. Harvey, and P. Gregorevic (Nov. 2017). "Regulation of tissue growth by the mammalian Hippo signaling pathway". In: *Frontiers in Physiology* 8.NOV, p. 942. DOI: 10.3389/fphys.2017.00942.

Wu, Y., X. Hu, Z. Li, M. Wang, S. Li, X. Wang, et al. (Feb. 2016). "Transcription Factor RFX2 Is a Key Regulator of Mouse Spermiogenesis". In: *Scientific Reports* 6. DOI: 10.1038/srep20435.

Yates, A. D., P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, et al. (Jan. 2020). "Ensembl 2020". In: *Nucleic Acids Research* 48.D1, pp. D682–D688. DOI: 10.1093/nar/gkz966.

Yu, F. X. and K. L. Guan (Feb. 2013). "The Hippo pathway: Regulators and regulations". In: *Genes and Development* 27.4, pp. 355–371. DOI: 10.1101/gad.210773.112.

Zervos, A. S., J. Gyuris, and R. Brent (Jan. 1993). "Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites". In: *Cell* 72.2, pp. 223–232. DOI: 10.1016/0092-8674(93)90662-A.

Zhao, S., W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu (Jan. 2014). "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells". In: *PLoS ONE* 9.1. Ed. by S.-D. Zhang, e78644. DOI: 10.1371/journal.pone.0078644.

APPENDIX A

ENRICHED GO TERMS FOR CLUSTERS #1 AND #6

Table A.1: Enriched GO Biological Processes in Cluster #1

GO.ID	Term	qval
GO:0048515	spermatid differentiation	1.07E-07
GO:0032504	multicellular organism reproduction	1.17E-06
GO:0007018	microtubule-based movement	1.39E-05
GO:0044782	cilium organization	1.39E-05
GO:0007017	microtubule-based process	1.45E-05
GO:0007281	germ cell development	8.54E-05
GO:0097722	sperm motility	0.0008536
GO:0022412	cellular process involved in reproductio...	0.00144045
GO:0044703	multi-organism reproductive process	0.003366977777778
GO:0060271	cilium assembly	0.010314333333333
GO:0035082	axoneme assembly	0.010314333333333
GO:0070925	organelle assembly	0.010314333333333
GO:0018200	peptidyl-glutamic acid modification	0.012194285714286
GO:0001539	cilium or flagellum-dependent cell motil...	0.012194285714286
GO:0019953	sexual reproduction	0.017072
GO:0000003	reproduction	0.028453333333333
GO:0022414	reproductive process	0.028453333333333
GO:0007286	spermatid development	0.028453333333333
GO:0048609	multicellular organismal reproductive pr...	0.04268
GO:0007276	gamete generation	0.061886
GO:0046174	polyol catabolic process	0.075198095238095

Table A.2: Enriched GO Biological Processes in Cluster #6

GO.ID	Term	qval
GO:0040011	locomotion	0.0001357425
GO:0043062	extracellular structure organization	0.0001357425
GO:0051270	regulation of cellular component movemen...	0.00090495
GO:0040012	regulation of locomotion	0.00090495
GO:0051674	localization of cell	0.001170402
GO:0006928	movement of cell or subcellular componen...	0.00231265
GO:0009611	response to wounding	0.0103422857142857
GO:0050793	regulation of developmental process	0.0127363333333333
GO:0032502	developmental process	0.0127363333333333
GO:0051239	regulation of multicellular organismal p...	0.0285196363636364
GO:0030030	cell projection organization	0.0285196363636364
GO:0009719	response to endogenous stimulus	0.03167325
GO:0010647	positive regulation of cell communicatio...	0.0445513846153846
GO:0001568	blood vessel development	0.06033
GO:0023056	positive regulation of signaling	0.076418
GO:0008219	cell death	0.0770883333333333
GO:0007565	female pregnancy	0.0770883333333333
GO:0023051	regulation of signaling	0.0770883333333333
GO:0048666	neuron development	0.0793815789473684
GO:0010941	regulation of cell death	0.0874785
GO:0042445	hormone metabolic process	0.0877527272727273
GO:0006022	aminoglycan metabolic process	0.0877527272727273
GO:0045595	regulation of cell differentiation	0.0928153846153846
GO:0051174	regulation of phosphorus metabolic proce...	0.0928153846153846
GO:0019220	regulation of phosphate metabolic proces...	0.0928153846153846
GO:0048584	positive regulation of response to stimu...	0.0928153846153846
GO:0022610	biological adhesion	0.0960811111111111
GO:0050808	synapse organization	0.0969589285714286

APPENDIX B

CLUSTERING RESULTS WITH DIFFERENT K VALUES

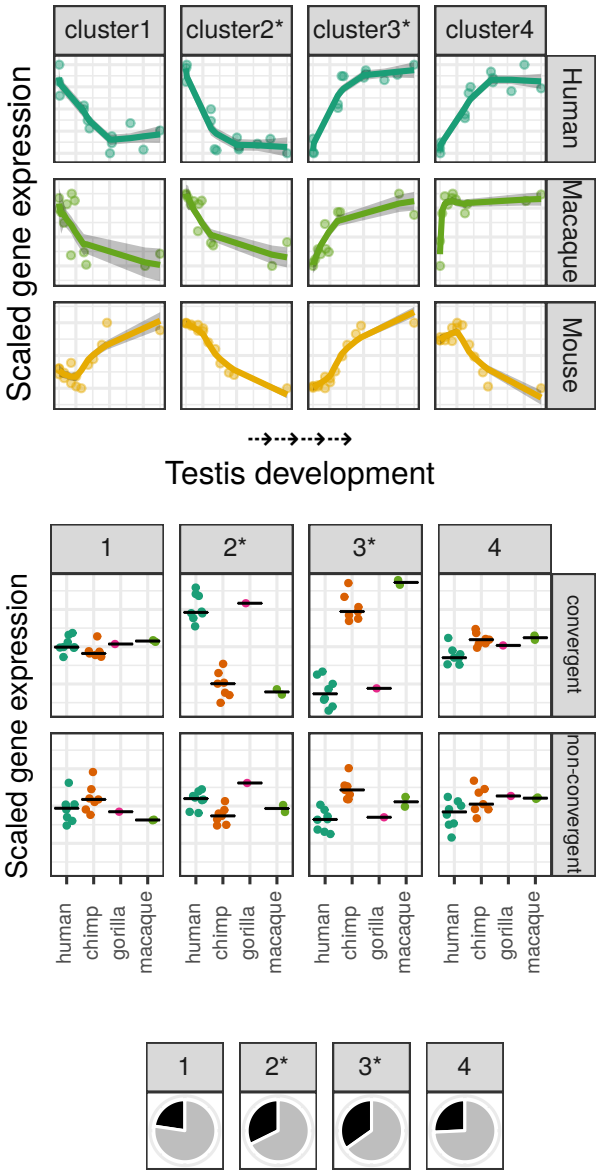

Figure B.1: Reproduction of Figures 3.13 and 3.14 with k=4

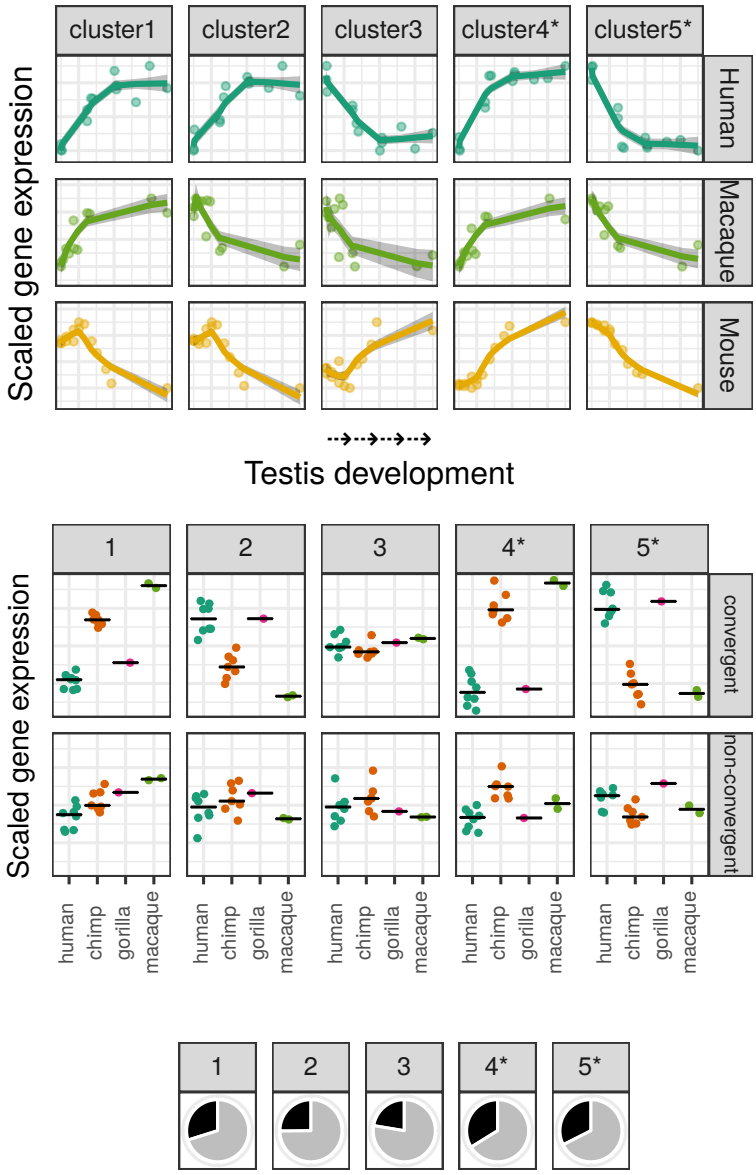

Figure B.2: Reproduction of Figures 3.13 and 3.14 with k=5

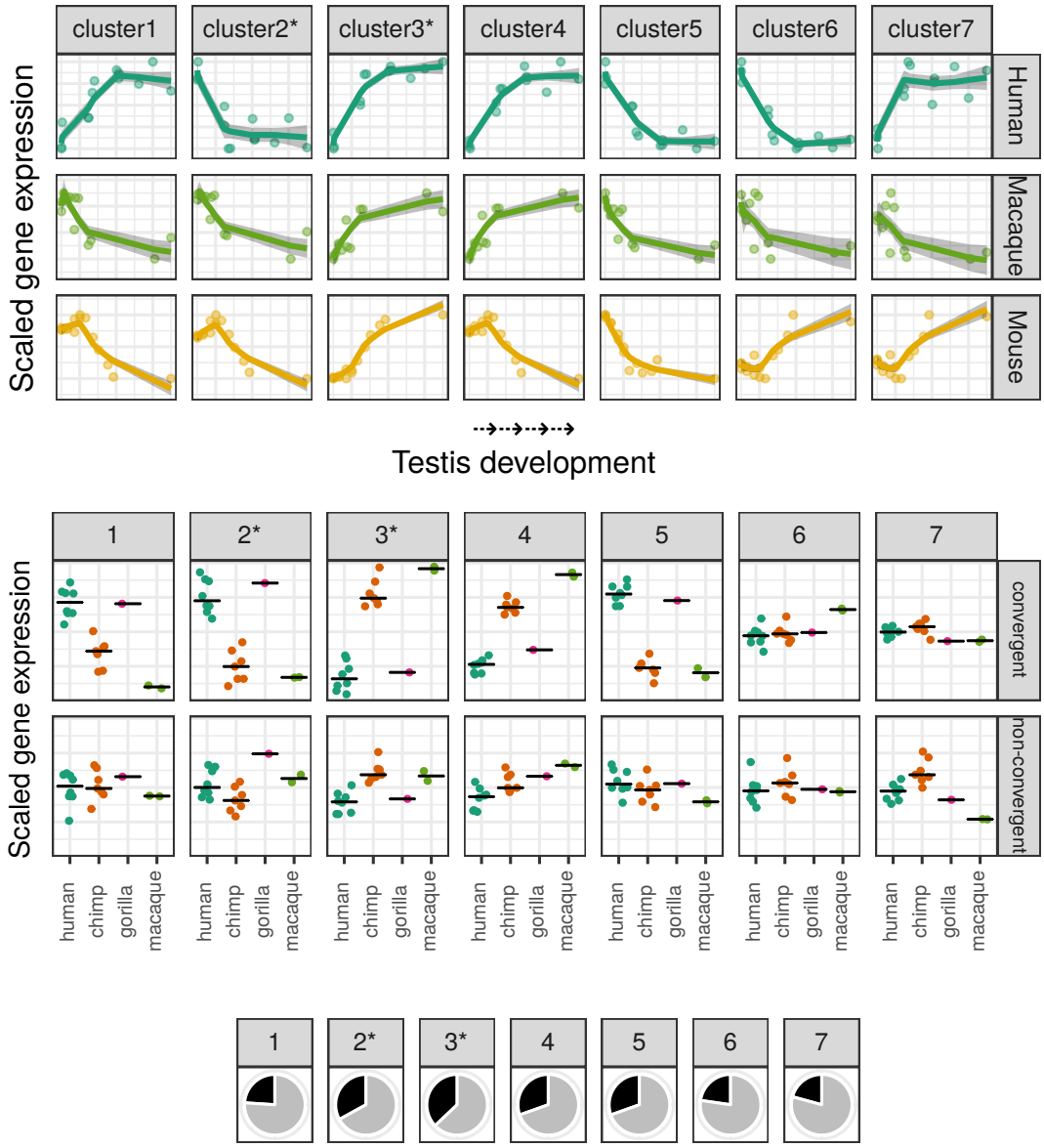

Figure B.3: Reproduction of Figures 3.13 and 3.14 with k=7

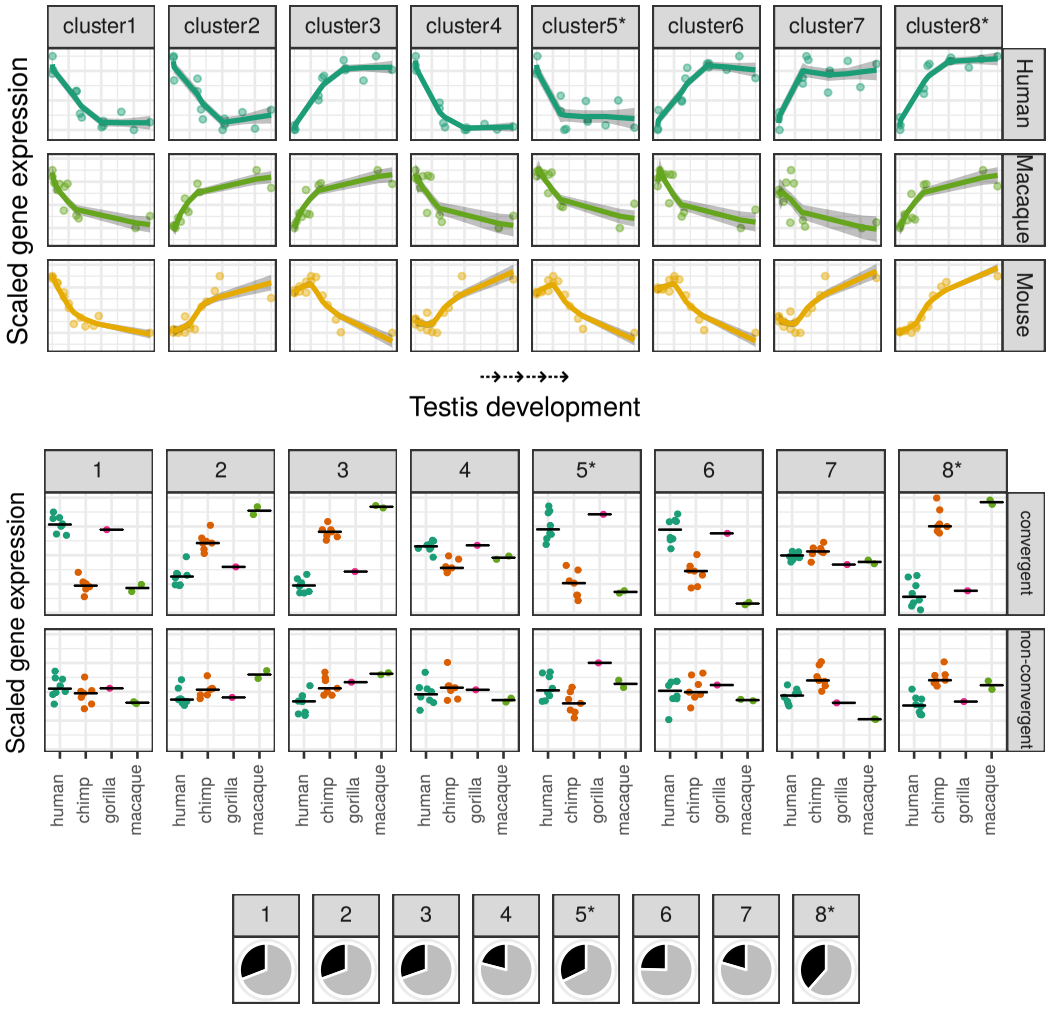

Figure B.4: Reproduction of Figures 3.13 and 3.14 with k=8

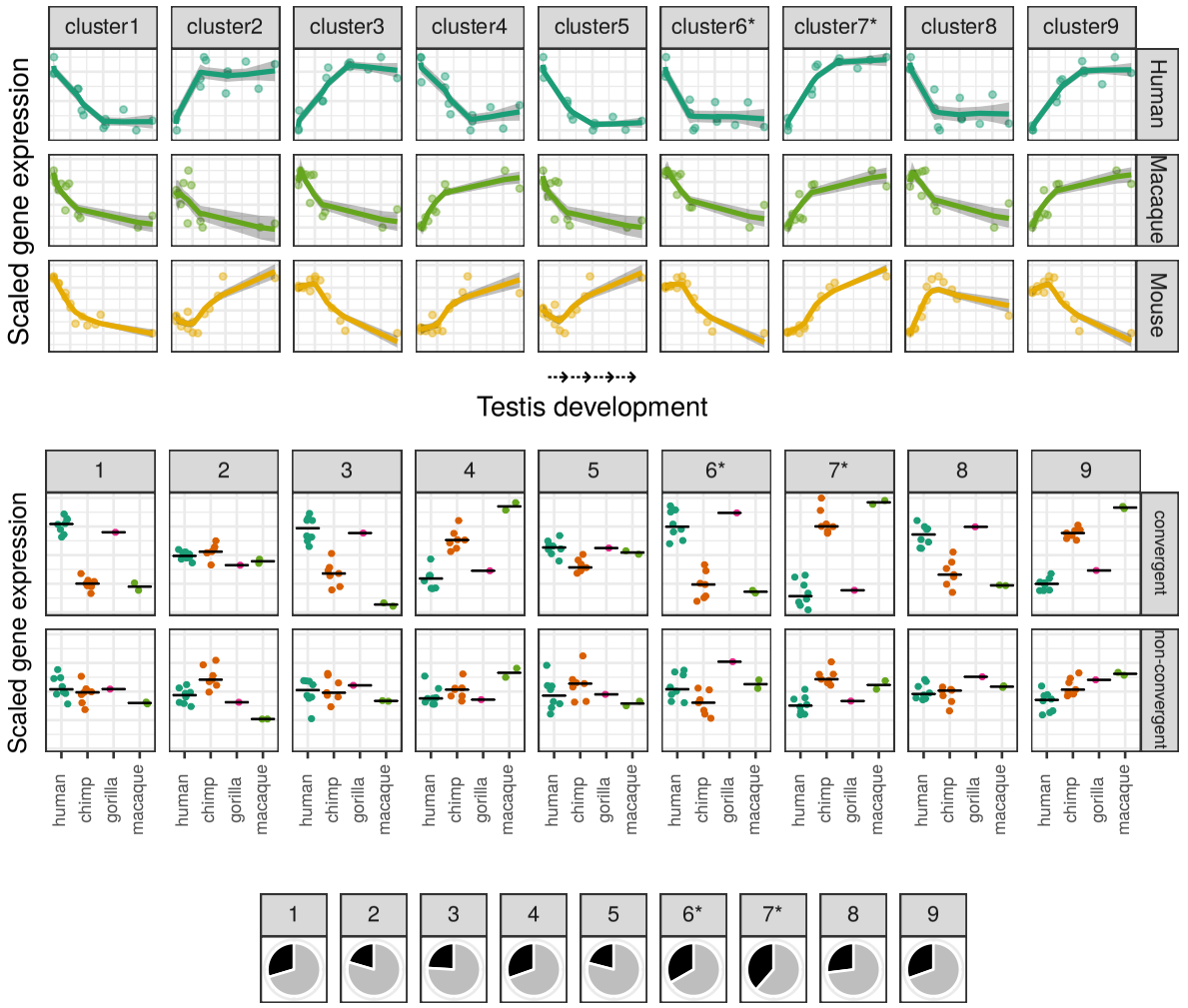

Figure B.5: Reproduction of Figures 3.13 and 3.14 with k=9

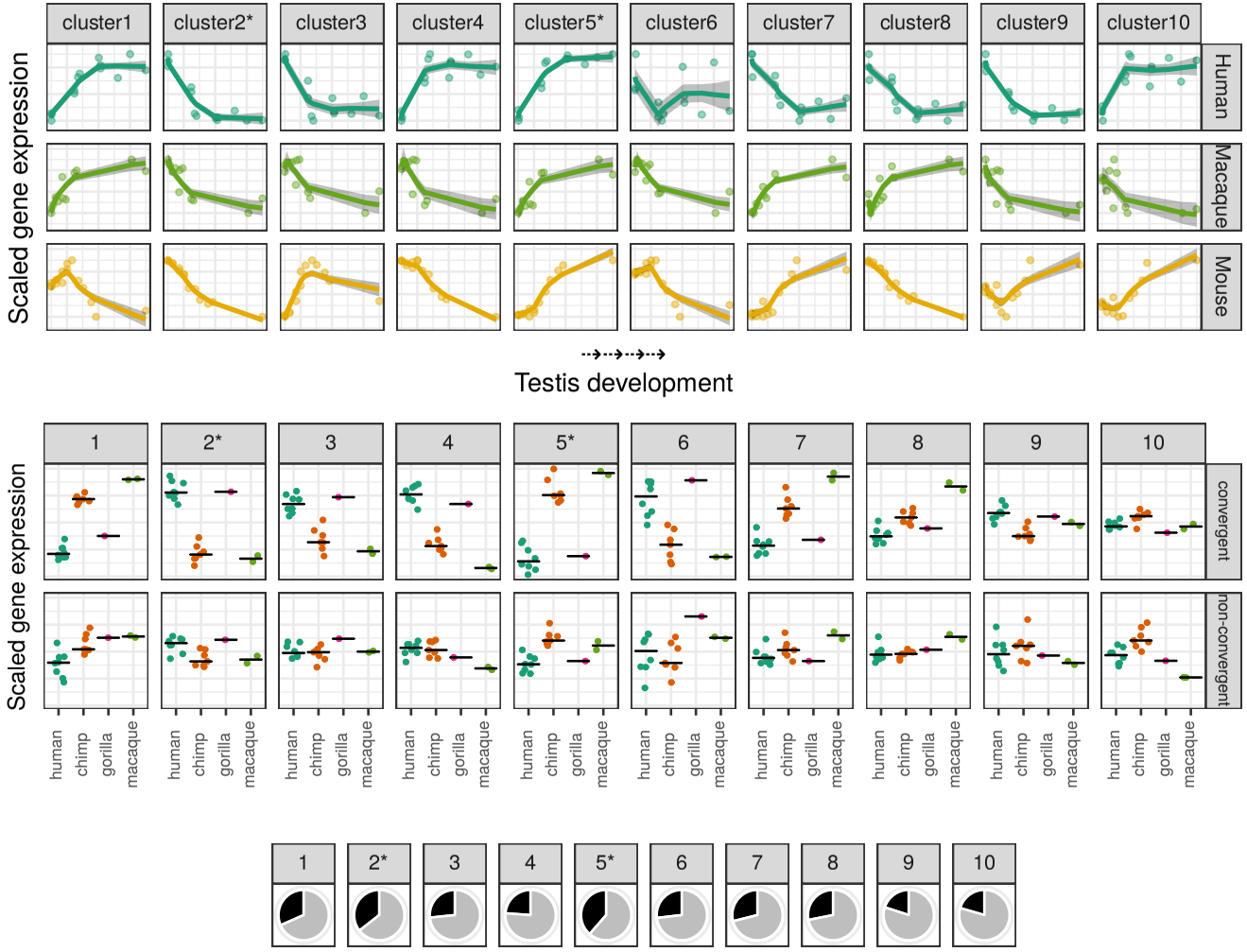

Figure B.6: Reproduction of Figures 3.13 and 3.14 with k=10

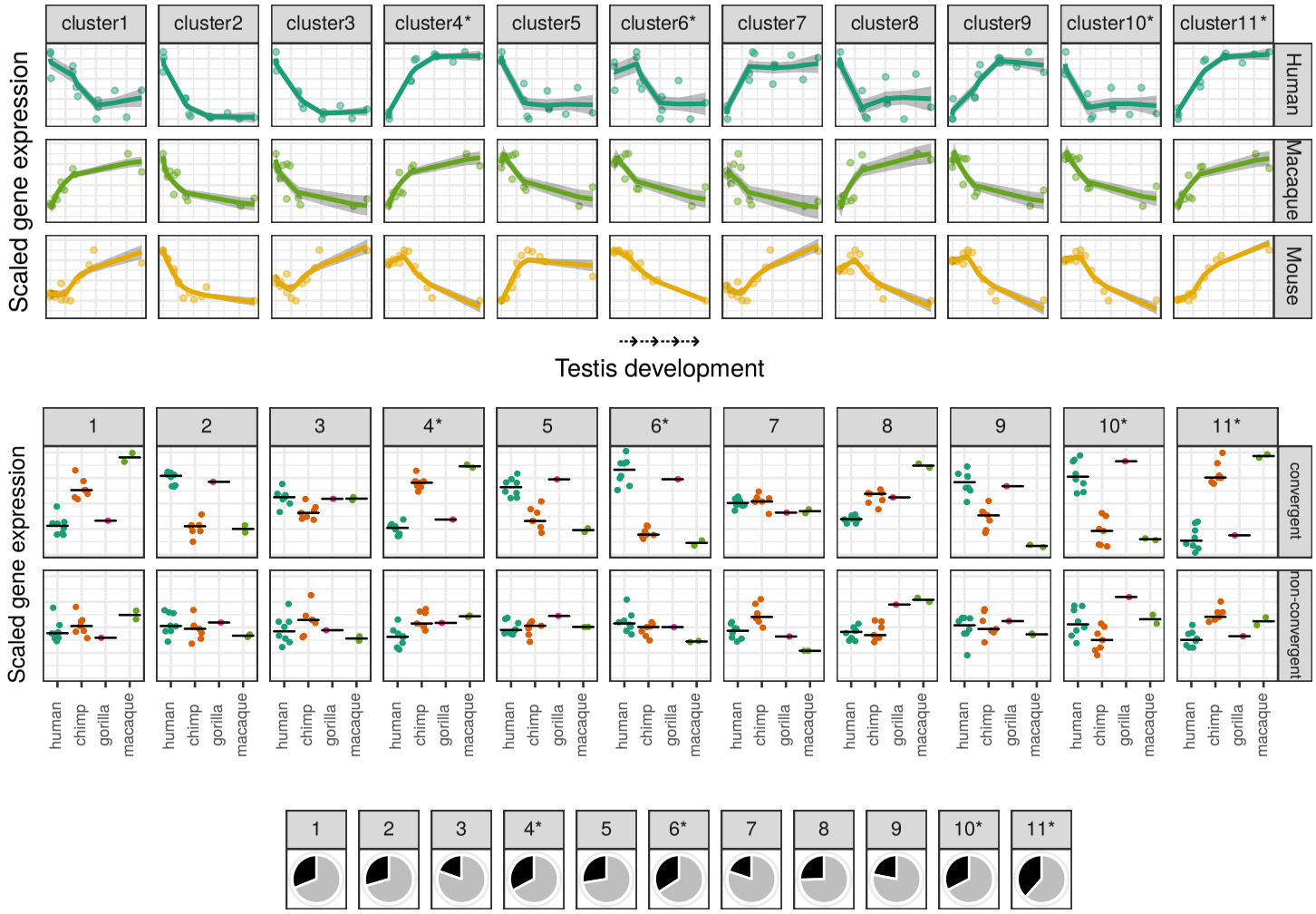

Figure B.7: Reproduction of Figures 3.13 and 3.14 with k=11

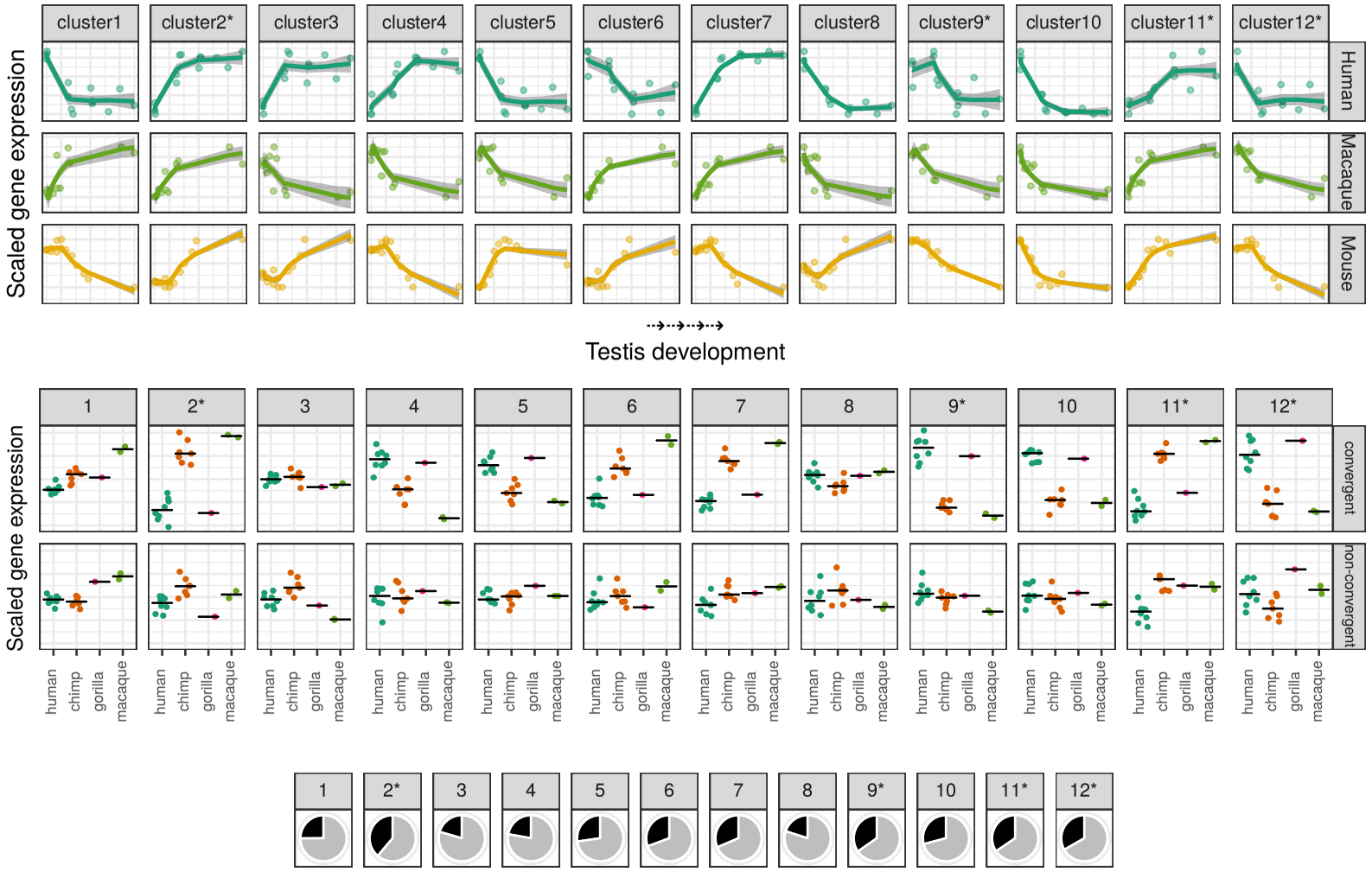

Figure B.8: Reproduction of Figures 3.13 and 3.14 with k=12

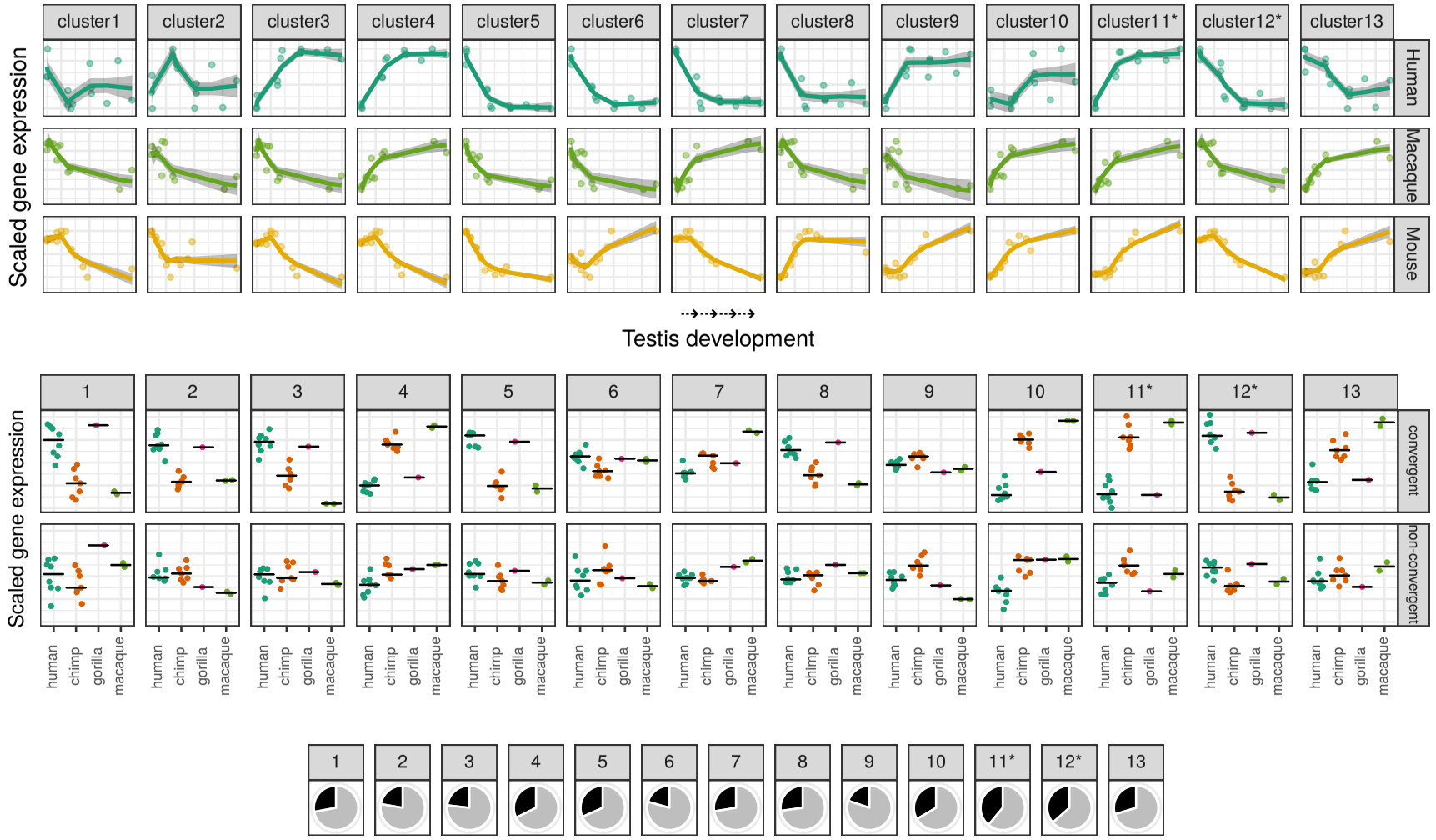

Figure B.9: Reproduction of Figures 3.13 and 3.14 with k=13

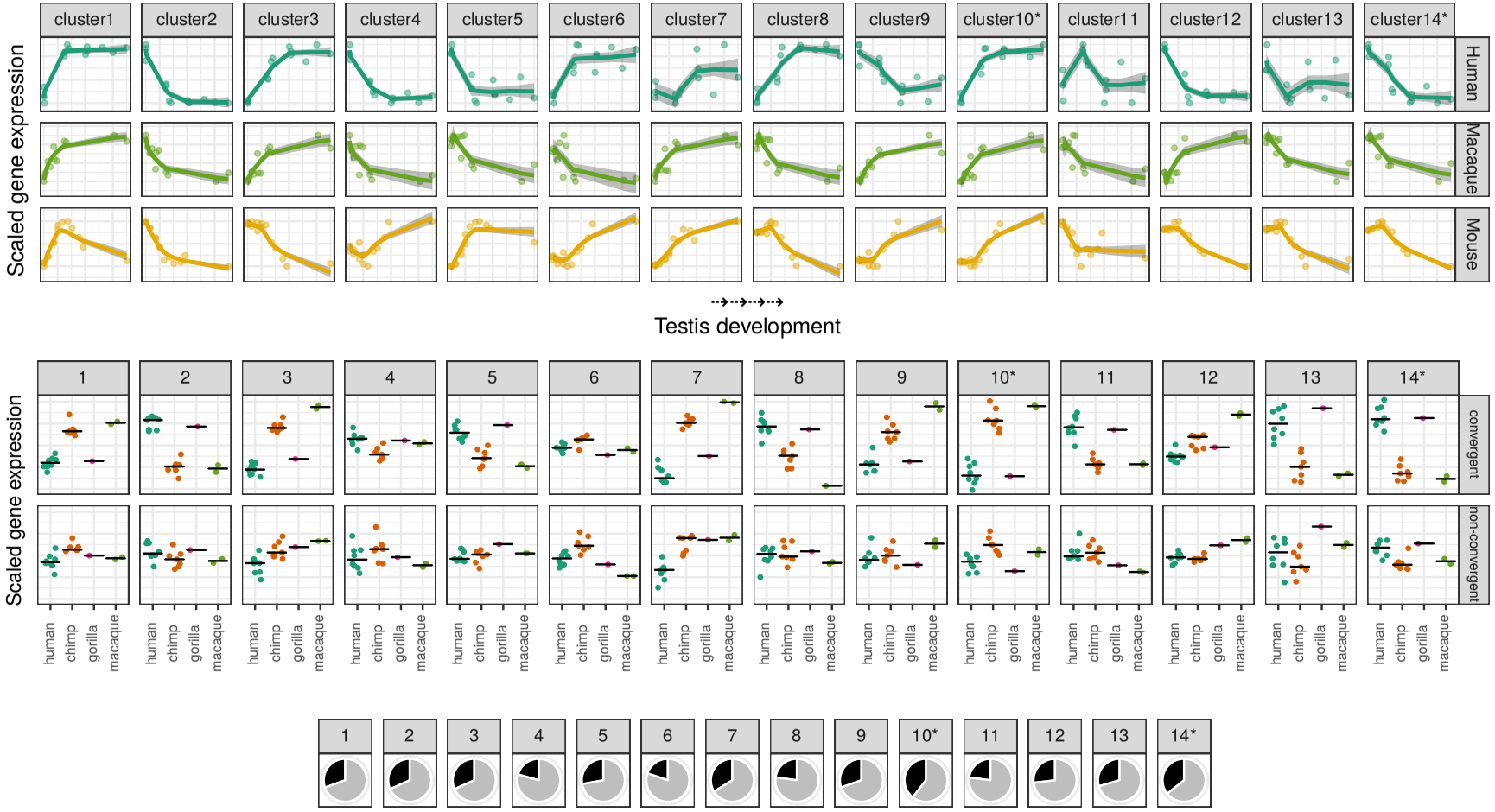

Figure B.10: Reproduction of Figures 3.13 and 3.14 with k=14