Multicopy Gene Family Evolution on Primate Y Chromosomes

MULTICOPY GENE FAMILY EVOLUTION ON PRIMATE Y CHROMOSOMES

Ana-Hermina Ghenu, Hons. B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2015) McMaster University

(Biology) Hamilton, Ontario

TITLE: Multicopy gene family evolution on primate Y chromosomes

AUTHOR: Ana-Hermina Ghenu, Hons. B.Sc. (University of Toronto)

SUPERVISOR: Dr. Ben J. Evans

NUMBER OF PAGES: [xvii], 117

ABSTRACT

Unlike the autosomes, the Y chromosome in humans and other primates has few protein coding genes, with only a few dozen single-copy genes and several tandem duplicated gene families, called the “ampliconic” genes. The interaction of many biological and evolutionary factors is responsible for this structural heterogeneity among different parts of the genome.

We sequenced and assayed the copy numbers of Y-linked, single-copy genes and ampliconic genes in a group of closely related macaque monkeys, then fit models of gene family evolution to this data along with whole genome data from human, chimpanzee, and rhesus macaque. Our results (i) recovered evidence for several novel examples of gene conversion in papionin monkeys, (ii) indicate that ampliconic gene families evolve faster than autosomal gene families and than single-copy genes on the Y chromosome, and that (iii) Y-linked singleton and autosomal gene families evolved faster in great apes than they do in other Old World higher Primates.

These findings highlight the evolutionary eccentricity of duplicated genes on the Y chromosome and suggest an important role for natural selection and gene conversion in the evolution of Y-linked gene duplicates.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisor, Ben Evans, for giving me the opportunity to work independently on this project, for his continued enthusiastic encouragement, and for helping me with my sometimes incomprehensible writing. Next, I would like to thank my committee member, Ben Bolker, for his patient mentorship, his enlightening and inspiring conversations, and for teaching me how to use `R`. I would also like to thank my committee members, Brian Golding and JP Xu. Brian – thanks for your wry wit and constant willingness to share your superior intellect with us lowly graduate students. JP – thanks for your encouragement and continued interest in my work. I am much obliged to the Biology Deptartment, the School of Graduate Studies, the Ontario Graduate Scholarship Program, and the National Science and Engineering Research Council for their enduring support. Finally, I'm indebted to Gabriel Marais and Isabel Gordo for taking me under their tutelage during my semester abroad at the Instituto Gulbenkian de Ciência.

My time at McMaster was wonderfully enlivened by my friends and colleagues in the Biology Department. Thanks to current and former Evans' lab members Adam Bewick, Jane Shen, Brian Alcock, Ben Furman, Shireen Bliss, Simone Mendel, Sonya Shri, Zach Hughes, and Graham Colby for your excellent conversations and for putting up with me constantly yelling or muttering at my code and qPCR's. I would be remiss if I didn't mention Yifei Huang for explaining to me many a 'very simple model' and helping me understand directed, bifurcating graphs. I am particularly grateful to everyone who offered me their friendship, laughter, and a shoulder to cry on over the last five years: Tara Sadoway, Lindsay Keegan, Jake Szamosi, Dave Leaman, Chai Molina, Jonathan Dushoff, Carolyn Lenz, Jen Klunk, Nathalie Mouttham, Spencer Hunt, Ranya Amir, Aaron Vogan, John Allison, Chyun Shi, Greg Barltrop, and Connie O'Connor.

Thanks to Señor Bandito (aka Kitten Little) for letting me know that I need to throw him his mousie already because life really isn't so serious all the time. And to my brother, Mike Ghenu, for reminding me of just how lucky I am to be a research scientist -slash-graduate student. A very special thank you goes to my partner, Eric Dewitt, for more often than not staying up late to talk me out of various freak-outs and for always believing in my best possible self. Finally, none of this would have been possible without my parents, Şerban and Nina Ghenu. Thanks Dad for nurturing my curiosity about science (and the world in general) from a young age by casually leaving excellent, carefully selected books and magazines around the house. And thanks Mom for making sure that I didn't starve over the last couple of years and unconditionally supporting me in all of my decisions.

DECLARATION OF ACADEMIC ACHIEVEMENT

Chapter 2 of this thesis has been prepared as a manuscript and is currently in review at *BMC Genomics*. For this chapter, laboratory work, analysis, programming, and manuscript preparation was primarily an individual effort, with contributions from Ben Evans and Ben Bolker.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Symbols and Abbreviations

$B = ID$ a model of gene family evolution where the duplication and deletion rates differ and the innovation rate is equal to the duplication rate. 26, 28, 30, 31, 44, 75

BD a model of gene family evolution where the duplication and deletion rates can differ (no innovation). 26, 28, 44, 60, 75

C_q quantification cycle. 58, 59, 72, 73

L a model of gene family evolution where the duplication rate is equal to the deletion rate (no innovation). 26, 28, 60, 75

$L = I$ a model indicating the special case where the duplication, deletion, and innovation rates are all equal. 26, 28, 30, 31, 39, 44, 75

LI a model of gene family evolution where the duplication and deletion rates are equal and there is an independent innovation rate. 26, 28, 44

N_e effective population size. 4, 5, 8, 12, 19, 32, 98, 101

λ indicates the special case when the per copy duplication rate is equal to the per copy deletion rate. 26, 29, 31, 52, 75

b the per copy duplication, or ‘birth,’ rate. 26, 52, 75

d the per copy deletion rate. 26, 52, 75

dN rate of substitutions at non-synonymous sites. 13

dS rate of substitutions at synonymous sites. 13

i the innovation rate per gene family. 26, 29, 52, 75

AG ampliconic gene. 11–14, 20–22, 24–42, 49–51, 56, 57, 61, 75, 95, 97–102

BAC bacterial artificial chromosome. 41, 54, 62

BIC Bayesian information criterion. 29–31, 52, 56, 61, 75

bp nucleotide base pairs. 42, 48, 55, 56, 70, 100

CI confidence interval. 75

CV coefficient of variation. 24, 25, 30, 59, 74

df degrees of freedom. 75

DNA deoxyribonucleic acid. 14, 19, 21, 27, 31, 41, 54, 55, 100

gDNA genomic DNA. 41, 43, 55, 57, 58, 60, 63, 72

IUCAC Institutional Animal Care and Use Committee. 41

LOD limit of detection. 58, 72

logLik log likelihood. 75

M-value geNorm mean stability value; see [1]. 43, 59, 74

MAP maximum *a posteriori*. 35

MLE maximum likelihood estimate. 52, 60

MRCA most recent common ancestor. 3, 7, 11, 29, 45, 56, 60, 95

msrY male specific region of the Y chromosome. 5–8, 11–14, 18–34, 36–42, 45–47, 49, 51, 52, 55, 56, 59, 61, 67, 78, 83, 95, 97–101

My million years. 11, 34, 42, 97

NAHR non-allelic homologous recombination. 7

NRQ normalized relative quantity. 59, 74

nt nucleotides. 47, 70

NTC no template control. 59, 73

PCR polymerase chain reaction. 23, 41, 42, 54, 55, 57, 58, 64, 67

qPCR quantitative PCR. 21–27, 31, 38–40, 43, 44, 49, 50, 57–61, 64, 67, 70–72, 74, 84–92

RDML Real-time PCR Data Markup Language. 42

w(BIC) Bayesian information criterion weights. 75

CHAPTER 1

INTRODUCTION

In the broadest sense, my thesis investigates two related questions: How are Y chromosomes different from autosomes?; and, Why are two Y chromosomes from different species often so dissimilar from one another, even when their autosomes are very similar? These questions have plagued geneticists and evolutionary biologists for a long time. The first was posed – and partially answered – by Hermann Muller in 1918 [2]. The second became of interest almost 50 years ago [3–6], when the development of detailed karyotyping techniques allowed geneticists to compare the Y chromosomes of different mammalian species.

The general answer to these questions is that the genomic and evolutionary processes operating in Eukaryotes act heterogeneously among different evolutionary lineages as well as between different parts of the genome, like the Y chromosome and the autosomes. Admittedly, this answer seems rather tautological. But, molecular evolution is about the four population genetic processes handed down to us from the modern evolutionary synthesis [7]. So, perhaps, reducing a question about observations to a question about processes is apt.

The remainder of this chapter explains how each of the four population genetic processes (mutation, selection, drift, and linkage) can act heterogeneously and what this means for gene families on primate Y chromosomes. First, I discuss each of the four processes with regard to heterogeneity between the Y chromosome and autosomes. Then, I will turn to discussing lineage heterogeneity in a similar manner. Next, as my thesis is about duplicate genes, I provide some background on methods for making sense of gene family evolution. Finally, as my thesis is also about the ampliconic genes on the Y chromosome, I will provide some background on these gene families and their evolution. Throughout the introduction, examples from Old World primates and mammals will be used when available.

Why heterogeneity between autosomes & the Y?

Why heterogeneity between autosomes & the Y: mutation

The possibility of an increased mutation rate in males compared to females was noticed first by Haldane [8] and later confirmed in mammals [9–11]. Since the Y chromosome spends all of its time in males while autosomes spend about half their time in males, a higher mutation rate in males can result in faster divergence of the Y as compared to the autosomes. Faster male mutation is a result of the larger number of mitotic events that germ cells undergo during spermatogenesis in the testis as compared to the few mitotic events during oogenesis [10]. Estimates of the ratio of mutations in males compared to females range from two to effectively infinite in mammals [10]. More recent estimates from primates suggest a ratio of about three and confirm that the extent of male mutational bias is roughly equal to – or may in fact surpass – the excess of mitotic replications in the male germ-line as compared to the female [12]. Furthermore, differences in the types of mutational events common in males as compared to females [13, 14] can lead to differences in the types of genetic variants common on the Y chromosome as compared to the autosomes.

Why heterogeneity between autosomes & the Y: selection

Contrary to previous theories, the Y chromosome is an important determinant of male phenotype [15–17]. The Y chromosome has been found to be associated with male fertility in humans [18] and even less obvious traits like lifespan [19] and locomotive activity [20] in *Drosophila* species. Therefore, any Y-linked genetic variation that underpins phenotypic variants relevant for fitness is a potential target for natural selection.

Genes on the Y chromosome are only ever inherited and expressed in males, therefore, there can be male-specific selection on these genes which cannot occur on the male and female inherited autosomes [15]. For example, the Y chromosomes of human and cattle both have a high proportion of genes involved in sperm development [21], with almost half of the gene families on the bovine Y chromosome having been independently acquired since the most recent common ancestor (MRCA) of primates and ungulates [22]. Another example is domesticated chickens, which have undergone artificial selection for female-specific processes like egg-laying and, as a result, have experienced evolutionary changes in expression in genes on the female-specific W chromosome as compared to their non-domesticated ancestors [23].

Furthermore, genes which move from the autosomes to the Y chromosome, especially early on in the history of the Y, have functions which are beneficial for males but detrimental for females [24]. The process of male-specific adaptation in gene function and expression can occur very rapidly after the expansion of a new region of the Y chromosome [25], underscoring the fact that selection can be strong on the Y chromosome.

Why heterogeneity between autosomes & the Y: drift

The term "drift" describes the random sampling of alleles from one generation to the next in a population of finite size [26]. Alleles that are selectively neutral, or have a sufficiently small selection coefficient, can be driven to fixation or extinction through this stochastic process. The magnitude of drift as compared to the magnitude of selection is expressed by the effective population size (N_e): the smaller the N_e, the more that drift will dominate over selection in determining the fate of a new allele, and vice versa.

The N_e is different between the autosomes and the Y chromosome because there is only one Y chromosome for every four autosomes in the population, if we assume an equal ratio of males and females [27]. This smaller N_e of the Y chromosome means that it is more influenced by drift, resulting in weak purifying selection and faster divergence than autosomes. In addition, when the sex ratio is skewed towards females, the N_e of the Y chromosome may become even smaller. Further, if males have higher variance in reproductive success than females, then the N_e of males is smaller than that of females, and the Y chromosome will have an even smaller N_e than the $1/4$ expectation. This means that a species' mating system will influence the disparity in N_e between the Y chromosome and the autosomes.

Why heterogeneity between autosomes & the Y: linkage

Recombination is the 'shuffling' of the alleles of genes that are on the same chromosome; this breaks up the genetic linkage between mutations that are either new or segregating in the population. When this process occurs in germ-line cells, as in the case of meiotic recombination, it produces offspring that are more genetically variable from each other than those produced in the absence of recombination. Natural selection acts more efficiently in a genetically variable population: deleterious mutations are purged from the population because individuals containing many deleterious mutations tend to produce fewer offspring than those with few deleterious mutations, or even multiple advantageous mutations. Therefore, recombination allows selection to operate efficiently, effectively increasing the N_e of the genomic region in which it occurs.

The most obvious difference between the Y and the autosomes is the inhibition of recombination in what is known as the male specific region of the Y chromosome (msrY) (i.e. in contrast to the pseudoautosomal region of the Y chromosome where recombination occurs freely with the pseudoautosomal region of the X chromosome). This nearly complete linkage among genes on the msrY is probably primarily responsible for the accumulation of deleterious mutations [28], the inability to efficiently fix advantageous mutations, rapid divergence [29], degradation of functional genes [30, 31], accumulation of repetitive elements [32], and fusions with autosomes [33] that are often observed among the Y chromosomes of many species.

The processes that lead to the accumulation of deleterious mutations and the "degeneration" of the msrY [34] are different examples of Hill-Robertson effects, where directional selection at one locus interferes with selection at a linked locus, leading to a local decrease in the N_e of the linked genomic region [35–37]. According to the simulation study of [38], the relative contributions of different flavours of Hill-Robertson effects change as the msrY matures.

Early on, just after recombination has ceased, Muller's ratchet [39, 40] dominates. This term describes the decrease in N_e in a finite population as a result of segregating, deleterious mutations with strong effect being removed from the population. Muller's ratchet results in the accumulation of deleterious mutations with moderate effect and an overall decline in the mean fitness of the population [41]. Empirical studies of young Y and neo-Y chromosomes have found that Muller's ratchet can be responsible for either rapid loss of functional genes [42] or rapid gene silencing [43], respectively.

Then, once enough time has passed that most segregating variation has been eliminated, background selection [44, 45] and the "ruby in the rubbish" [46] mechanisms dominate. These terms describe that a new mutation of moderate effect, either deleterious or advantageous, respectively, can persist and rise to fixation only if it occurs on the genetic background with the fewest number of deleterious mutations. The net effect of background selection is a paucity of genetic variation on the msrY and a reduction in the number of Y-linked genes such that the msrY has remarkably low segregating genetic variation (as empirically observed in [11, 47, 48]) and is a smaller target for deleterious mutations [38].

Finally, genetic hitchhiking dominates in relatively old and gene depauperate msrYs [38]. As suggested by its name, this process is the 'hitchhiking' of deleterious alleles that are linked to an advantageous mutation which is sweeping to fixation under positive selection [49]. Genetic hitchhiking can lead to msrY degeneration in circumstances when other processes, like Muller's ratchet, cannot operate efficiently [38, 50]. I do not know of any empirical studies that unequivocally illustrate the impact of genetic hitchhiking, but it is possible for positive selection to act on msrYs [51] and the extreme paucity of genetic variation on the human msrY is not explained by models that consider only deleterious mutations [52].

Essentially, the result is that regions completely lacking recombination will follow a relatively well characterized path towards degeneration and, therefore, will differ substantially from autosomes. In spite of this degeneration process, old msrYs can persist for longer periods of time than previously thought [31, 38, 53] through the persistence of strong purifying selection on the few remaining Y-linked genes [54–56].

Finally, note that the cessation of recombination between the X and Y chromosomes happens gradually as an ongoing process, usually in the form of one or more inversion events [56–58]. This means that within the msrY there are regions which differ in the age since recombination ceased [59]. When a cessation of recombination event occurs on an existent msrY this accelerates the degeneration process along the entire msrY because the new non-recombining region increases the number of genes and the amount of genetic variation on the msrY [38]. Overall, the degeneration of msrYs is a highly dynamic process susceptible to many factors, including the stochastic occurrence of inversions, the strength of selection, and the distribution of fitness effects of new mutations.

Why heterogeneity between lineages?

Why heterogeneity between lineages: mutation

Repair mechanisms can differ between species, leading to differences in which types of mutations occur in different species. For example, a higher content of inverted repeats and repetitive sequences may lead to increased chromosome fragility and increased opportunities for duplication or deletion through non-allelic homologous recombination (NAHR), or "ectopic recombination," and microhomology-mediated events [60, 61]. This means that heterogeneity between lineages can be mutationally driven.

The rate of segmental duplications is known to vary among higher primate species. It seems that the rate of segmental duplications began to accelerate in the ancestor of great apes [62, 63], culminating in a burst of segmental duplications in the MRCA of humans, chimpanzees, and gorillas [64–66]. This has resulted in a prevalence of NAHR-generated intrachromosomal structural variation in extant apes [67, 68]. In contrast, Old World monkeys like the rhesus macaque are less likely to undergo NAHR and segmental duplication than humans or chimpanzees [68, 69]. These differences in mutational mechanisms can explain the heterogeneity of duplication and deletion rates between lineages.

Why heterogeneity between lineages: drift

Lineages differ in the overall size of their populations, leading to differences in their N_e and the role of drift. For example, among human, chimpanzee, and rhesus macaque, macaques have a much larger extant population size and a larger N_e than the other two [63]. Chimpanzees, despite their smaller extant population size, have a much larger N_e than humans, who underwent a large bottleneck in their evolutionary history.

Why heterogeneity between lineages: linkage

Although there is lineage heterogeneity in the recombination rate within autosomes [63], I will focus on lineage heterogeneity in msrY linkage. As explained above, the size of the msrY as a target of selection and mutation influences the strength of msrY degeneration. This means that species which happened to have many genes on their ancestral msrY or have recently acquired more genes on their msrY are susceptible to faster degeneration.

Another source of heterogeneity among the msrYs of different species is the occurrence of infrequent recombination with the X [70, 71]. This type of recombination is probably mediated by high sequence identity [72] between X and Y paralogs ("gametologs" [73]). Consistent with this mechanism, all of the X-to-Y gene conversion events that have been identified thus far are found on regions of the msrY that recently ceased recombining with the X chromosome [70, 71, 74, 75]. Overall, this means that species which are closely related and/or happen to have high sequence identity between gametologs will tend to be more likely to engage in this type of gene conversion.

A primer on investigating gene family evolution

Evolutionary biologists usually use models of nucleotide substitution to estimate the rate of gene family duplication or deletion (for example, [76–78]). This method can be used when substitutions accumulate separately between independently evolving gene copies, acting as a molecular clock to mark the time since the duplication event that created the copies. Unfortunately, ampliconic gene copies on Old World primate Y chromosomes do not evolve independently since gene conversion is rampant (see below). Therefore, we had to use a different, sequence-free method for estimating the rate of ampliconic gene family evolution.

We used a comparative phylogenetic approach to model gene copy numbers as discrete traits that evolve along a given phylogeny. This type of model has been used before to estimate the rates of gene family evolution from copy number data [79–82]. The gene copy number is treated as a discrete, ordered trait with evolution progressing sequentially (e.g. $0 \rightarrow 1 \rightarrow 2$ or $2 \rightarrow 1 \rightarrow 0$). These types of models are continuous-time Markov models with a time-homogeneous birth-death process where the probability of an event is state dependent. Figure 1.1 illustrates what is meant by these technical descriptors.

Felsenstein's pruning algorithm is used to calculate the likelihood of the observed data along the entire phylogeny [83]. Although the model is a time-homogenous process along any particular branch, different branches can take on different rates. In this way, rate heterogeneity among lineages can be incorporated into the model.

One difficulty of these types of models is that there is no stationary distribution: if the Markov process is run for an infinite period of time, it will either go to zero or to infinity [84]. The pruning algorithm requires known prior probabilities so that the states at the root of the phylogeny can be marginalized out by the "pulley principle;" in nucleotide substitution models, the stationary distribution is used for the prior probabilities [83]. Without a stationary distribution, the maximum likelihood estimator can take on biologically unreasonable values.

Of the previous programs that have been built to deal with these models, the most prominent are CAFE [79, 85, 86] and BadiRate [82]. Both of these programs were built specifically to identify gene families whose copy numbers are evolving unusually (i.e. either too fast or too slow) compared to other gene families in the given genomes. Both of these programs solve the stationary distribution problem by fitting a standard distribution (often Poisson) to the observed copy number data and using it as the prior distribution. The main difference between the two programs is that CAFE sets an arbitrary bound on the maximum gene family size, while BadiRate does not and instead assumes $0 < duplication \leq deletion$ [87]. A second difference between the two programs is that CAFE allows only duplication and deletion to operate, while BadiRate, which is based on a model of gene content in Archaea [87], allows duplication, deletion, and innovation to operate. A final difference between the two programs is that CAFE uses an approximation for calculating the event probabilities [84] which is numerically unstable [87] for biologically relevant cases, while BadiRate uses a Galton-Watson process whose calculation of event probabilities is numerically stable [87].

Unfortunately, because these two programs use different assumptions, it was difficult for us to figure out which one is ‘right.’ So we built a program that allows us to compare the fit of different models of copy number evolution for our data (see Chapter 2).

A primer on mammal Y chromosomes and their ampliconic genes

As methods for sequencing male-specific regions [88] and repetitive regions have improved, investigating sex chromosome evolution has become ‘sexier’ than ever. We humans are personally acquainted with a relatively unique Y chromosome. The human msrY is relatively old, small, and has a peculiar structure with many inverted segmental duplications. Our msrY was born in the therian mammal MRCA, just prior to the divergence of placental (eutherian) mammals and marsupial (metatherian) mammals ≈ 180 million years (My) ago [89]. The birth of our msrY was caused by two events in quick succession on the therian MRCA chromosome homologous to platypus chromosome six [90]. First, allelic recombination ceased between the proto-Y and proto-X [91]. Then, a gain of function mutation neofunctionalized the *SOX3* copy on the proto-Y into the trigger of male sexual development, *SRY* [92]. Since its birth, our msrY has increased its genetic content by undergoing five cessation of recombination events with the X chromosome [57, 59] – the newest just before the divergence of Old World primates [54] – as well as a recent translocation event from the X just after our divergence from chimpanzees [93].

The picture that has emerged from complete and partial msrY sequences of eutherian mammals [54, 94–99] is of two discrete classes of protein coding genes. The X-degenerate genes are single copy, broadly expressed gene families that began to diverge from their gametologs on the X chromosome when recombination between the X and msrY was inhibited; while the ampliconic genes (AGs) are multi-copy, testis-expressed gene families of variable origin that increased in copy number after their origination on the msrY [100]. Some AG families arose from X-linked genes, others were transposed from the autosomes to the msrY, and a few may be of *de novo* origin [100]. Within a species, functional AG paralogs tend to have high sequence identity ($> 99.9\%$) as a result of frequent ectopic recombination ("gene conversion") between copies found in tandem arrays, segmental duplications, or inverted segmental duplications. As expected from models of msrY degeneration, the X-degenerate genes are under strong purifying selection [55, 101]. AGs, on the other hand, exhibit much higher inter-specific divergence and have a longer average lifespan on the msrY than X-degenerate genes [98, 99]. AG families have evolved multiple times independently in different mammalian lineages, with different gene families increasing in copy number [89].

Since the existence of the AGs was unanticipated by traditional models of msrY degeneration, several hypotheses have been postulated to explain the evolution and persistence of this gene class. The neutral evolution hypothesis is that gene duplication is just another consequence of degeneration, albeit an elaborate one. Under this hypothesis, duplicate copies were fixed by drift in a step-wise fashion [102] and gene conversion occurs between these copies for as long as they have high sequence identity [72, 103]. An inadvertent result of this sequence homogenization is the removal of new pseudogenizing mutations, which leads to a decreased death rate of duplicates [104]. The faster substitution rate of AGs may result from an effectively increased neutral mutation rate due to the increased target size of these multi-copy genes [104, 105].

By adding selection on different aspects of this system, we can arrive at other hypotheses. For example, if there is a selective advantage to having increased copy numbers of a particular gene family, then fixation of a new AG copy will be favoured [104]. AG copy numbers in bulls and humans have been shown to be positively correlated with fertility [106–108] and in mice it has been proposed that increased AG copy numbers are favoured to escape transcriptional repression in spermatids [109]. This has led some authors [96, 110, 111] to speculate that species with intense sperm competition are more likely to exhibit AG copy number polymorphism and have faster rates of copy number evolution.

Secondly, the two simulation studies [102, 104] that investigated AG copy number evolution under gene conversion and duplication did not consider advantageous mutations. Theoretical models of sequence evolution at multi-copy loci undergoing gene conversion find that the N_e of these loci is increased proportional to the number of gene copies [112].

As a result, we would expect that AG sequences experience stronger selection and might postulate that inverted segmental duplications exist because they “accelerate adaptation by increasing the potential targets and fixation rates of incoming beneficial mutations” [105]. Although studies have consistently found higher dN/dS (ratio of the rate of substitutions at non-synonymous sites (dN) vs the rate of substitutions at synonymous sites (dS)) in AGs as compared to X-degenerate genes [54, 98], this finding is confounded by the difference in breadth of gene expression between these two categories [105].

Finally, the hypothesis that was initially proposed to explain the AG [94] comes from the observation that human AG tend to be arranged in inverted segmental duplications. This observation led some authors [94, 102, 113, 114] to postulate that the organization of the msrY in humans has evolved to promote the rapid gene conversion rates necessary to circumvent the degeneration of this non-recombining region. However, high rates of gene conversion may be disfavoured since crossing-over between segmental duplications can result in isodicentric Y chromosomes, depending on how the Holliday junctions are resolved [115, 116]. Since isodicentric Y chromosomes are correlated with significantly decreased fertility in humans, frequent crossing-over would favour slower or inhibited gene conversion [104].

Goals

The goal of my thesis is to better understand the relative contributions of gene conversion and duplication in generating and maintaining AG families on the degenerate, mammalian msrY. The original motivation of my work was to attempt to test some of the hypotheses summarized above by focusing on AG evolution in Old World monkeys, where we have access to high quality msrY sequence data from human, chimp, and rhesus macaque as well as genomic DNA samples from multiple species. The questions that my work has actually been able to address are: How common is gene conversion among AGs in Old World primates other than humans and chimps?; How frequent are the duplication and deletion events that drive copy number variation of AGs among Old World primates?; and, Do copy numbers evolve differently between AGs, Y-linked single copy genes, and autosomal genes? (see Chapter 2).

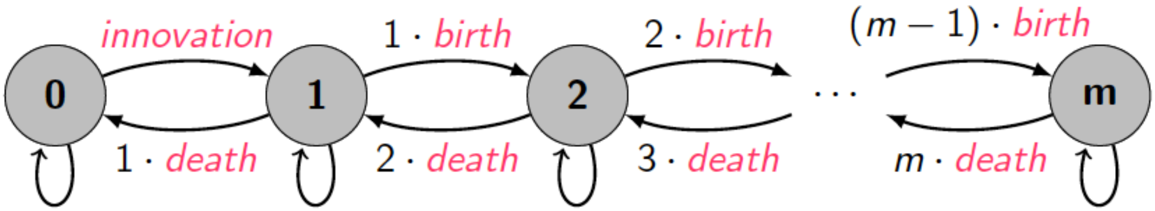


Figure 1.1: **Schematic of the general-case of our gene copy number evolution model.** Shaded circles show discrete copy number states and arrows indicate possible transitions between states. “m” indicates the arbitrary, finite maximum allowed number of copies. Arrows are labeled with the names of the different parameters in pink. Transitions occur at rates that are dependent on the current state (black numbers on arrow labels).

CHAPTER 2

MULTICOPY GENE FAMILY EVOLUTION ON PRIMATE Y CHROMOSOMES

Ghenu, A.-H., B.M. Bolker, D.J. Melnick, B.J. Evans, *BMC Genomics* **2015**. (In Review)

2.1 Abstract

Background: The primate Y chromosome is distinguished by a lack of inter-chromosomal recombination along most of its length, extensive gene loss, and a prevalence of repetitive elements. A group of genes on the male-specific portion of the Y chromosome known as the "ampliconic genes" are present in multiple copies that are sometimes part of palindromes, and that undergo a form of intra-chromosomal recombination called gene conversion, wherein the nucleotides of one copy are homogenized by those of another. With the aim of further understanding gene family evolution of these genes, we collected nucleotide sequence and gene copy number information for several species of papionin monkey. We then tested for evidence of gene conversion, and developed a novel statistical framework to evaluate alternative models of gene family evolution using our data combined with other information from a human, a chimpanzee, and a rhesus macaque.

Results: Our results (i) recovered evidence for several novel examples of gene conversion in papionin monkeys, indicate that (ii) ampliconic gene families evolve faster than autosomal gene families and than single-copy genes on the Y chromosome and that (iii) Y-linked singleton and autosomal gene families evolved faster in human and chimps than they do in other Old World higher Primate lineages we studied.

Conclusions: Rapid evolution of ampliconic genes cannot be attributed solely to residence on the Y chromosome, nor to variation between primate lineages in the rate of gene family evolution. Instead other factors, such as natural selection and gene conversion, appear to play a role in driving temporal and genomic evolutionary heterogeneity in primate gene families.

2.2 Background

Gene families are composed of gene copies that were generated by speciation (orthologs) and those that were generated by gene duplication (paralogs). The evolutionary histories of gene families are trimmed by gene loss and intertwined by non-reciprocal recombination (gene conversion), raising the question of whether and how genomic context influences their evolution. One genomic context of interest is the male specific region of the Y chromosome (msrY) of placental and marsupial (therian) mammals. The origin of this region coincides with the ascendancy of the *SRY* gene as the trigger for the male sex phenotype about 180 million years ago [89, 117, 118]. Subsequently, progressively larger portions of the Y and X chromosomes began diverging from one another as large inversions impeded recombination, forming “strata” with differing levels of divergence [57, 119]. During this time, recruitment of alleles with sexually antagonistic function (i.e., alleles that are advantageous to one sex but deleterious to the other) to this region may have been favoured by natural selection [120]. Compared to recombining genomic regions, a lack of recombination rendered the msrY more vulnerable to phenomena that decrease the efficacy of natural selection by Hill-Robertson effects, including Muller's Ratchet, genetic hitchhiking, and background selection [36, 121]. This has had profound consequences over time, including gene loss and the accumulation of repetitive DNA [121]. Today, contemporary eutherian msrYs retain only a small fraction ($\approx 5\%$) of the genes that were present before divergence from the X chromosome [119]. Male-specific inheritance influenced survival of genes in this region, and surviving genes on the msrY often have male-related functions and expression patterns [15, 122] or are subject to natural selection favoring similar dosage of the proteins they encode in males and females [99]. Examples of gene loss of otherwise conserved Y-linked gene exist, but these are often coupled with translocation to the autosomes or X chromosome [123].

Because the msrY is haploid and paternally inherited, this chromosome is more strongly influenced by genetic drift than the autosomes, which are diploid and biparentally inherited. With equal variance in reproductive success between the sexes, the neutral expectation for the msrY is that its effective population size (N_e) is 25% that of the autosomes. This disparity is more pronounced if the variance in reproductive success is higher in males than in females [27, 124], and even more so if the same male individuals monopolize reproduction over multiple generations [125]. Furthermore, in primates, the rate of sequence evolution is faster in males than in females (faster male evolution [9, 126–128]), a factor that could accelerate divergence and deterioration of genes on the msrY.

2.2.1 Multi-copy ampliconic genes on the msrY

Gene families that include paralogs on the msrY are called ampliconic genes (AGs) [94]. Compared to non-duplicated regions of the Y chromosome that are homologous to the X chromosome, AGs reside in regions that have a higher abundance of genes and pseudogenes but a lower abundance of retrotransposons; the latter observation possibly due to purifying selection acting to remove recently integrated retrotransposons [54, 94, 96]. In primates and other mammals, [99], fruit flies [105], and birds [129, 130], intra-chromosomal recombination occurs between AGs. This phenomenon leads to a non-reciprocal transfer where the nucleotide sequence of one duplicate is homogenized by that of another, a process known as gene conversion [131, 132]. On the human msrY, gene conversion occurs frequently – as much as one to four orders of magnitude faster than the nucleotide substitution rate [113, 133–135]. The close proximity on the msrY of direct or inverted ("palindromic") ampliconic repeats probably facilitates gene conversion [115], although it also occurs less frequently among ampliconic regions that are far apart, including between different chromosome arms [116]. As a result of frequent gene conversion, AG paralogs within either humans or chimpanzees (the tribe Hominini [136]) have higher sequence identity ($> 99.9\%$) than orthologous genes [113], even though similarities in copy number and genomic locations across species are consistent with the duplicates having arisen prior to speciation [54].

Gene copy number on the msrY is variable between Old World Primate species [111, 137], and AG copy number polymorphism is also observed within species [108–110], including humans (reviewed in [138]). *TSPY* copy number variation affects male fertility in humans [139–141] and bulls [107] (but see [142]), suggesting that copy number of this locus is subject to natural selection [94, 106].

2.2.2 Goals

In this study, our goal is to better understand the evolutionary mechanisms that drive gene family evolution within and among Old World Primate species, with a particular aim of testing whether the nature of gene family evolution of msrY AGs can be distinguished from that of other gene families on the msrY or autosomes. To this end, we collected and estimated phylogenetic relationships among DNA sequences from single copy genes (singletons) and AGs on the msrY of several closely related species of papionin monkey (tribe Papionini), and used a phylogenetic approach to qualitatively assess the frequency of gene conversion in AGs. We then used quantitative PCR (qPCR) to quantify AG copy number variation among and within various species of macaque monkey (genus *Macaca*). Using these data and other information from complete genome sequences from a human, a chimpanzee, and a rhesus macaque (*Macaca mulatta*), we then evaluated the fit of alternative models in which the rate and nature of gene family evolution is allowed to vary among genomic regions and among lineages of Old World Primates.

2.3 Results

2.3.1 Phylogenetic analysis of the primate msrY

Focusing on Old World Primate msrYs, we estimated phylogenetic relationships among msrY sequences from a human, chimp, and rhesus macaque, as well as new sequence data that we collected from several species of papionin monkey. New DNA sequences from four to 14 genes were collected from an olive baboon (*Papio anubis*), a mandrill (*Mandrillus sphinx*), and 15 macaque individuals (genus *Macaca*) from 9 species, including intra-specific information for four macaque species. We inferred the paternal relationships among samples from concatenated singleton genes from the msrY, as well as the phylogenetic relationships within individual AG families including pseudogene sequences obtained from completely sequenced msrY from a human, a chimp, and a rhesus macaque. Hereafter we define AGs as any msrY-linked, multi-copy gene family that has been previously demonstrated to have undergone gene conversion in human, chimp, or rhesus macaque; singletons are therefore defined as msrY-linked genes that have not been shown to have undergone gene conversion.

Singleton gene tree is consistent with known phylogeny

Unsurprisingly (because it was constructed from a partially overlapping data set), our estimates of phylogenetic relationships among nine concatenated single copy (singleton) genes on the msrY (*AMELY*, *DBY*, *PRKY*, *SMCY*, *SRY*, *TBL1Y*, *USP9Y*, *UTY*, and *ZFY*; Figure 2.1) were similar in topology and statistical support to the analysis of [143]. This phylogeny supports, for example, monophyly of the msrY of the Sulawesi macaques and a sister relationship between the msrY of *M. fascicularis* and *M. mulatta*. We added information from two additional samples of *M. maura* and a sample of *M. arctoides*; the phylogenetic placement of these samples was consistent with other studies [144, 145].

AG trees support frequent gene conversion in catarrhines

The gene trees inferred from AG sequences (Figures A1-A7) provided evidence of gene conversion in terms of (i) the detection of multiple gene sequences (with qPCR or cloning) with lower intraspecific than interspecific sequence divergence, but whose consistent copy number across species suggests ancestral gene duplication, (ii) well supported discordant relationships among putatively orthologous lineages within the gene tree of duplicated genes compared to the gene tree of single copy genes, and (iii) discordant phylogenetic relationships among 5′ and 3′ portions of duplicated genes. The first pattern (i) has been noted previously based on data from complete msrY sequences from rhesus, humans, and chimps [54, 96] and is further supported in our analysis by multiple identical or almost identical copies in various macaque individuals identified using qPCR. Specifically, for *HSFY* and *CDY*, sequences from macaques clustered in one clade with two almost identical sequences from the complete msrY of rhesus, but our qPCR results indicated that each of these macaque species carries at least two distinct copies (Figures A1, A3, A5). Pattern (i) is illustrated by *HSFY* in a baboon (Figure A3), which has two almost identical copies that cluster together in one clade, whereas other papionins each have two more diverged copies that each cluster in different clades, with evolutionary relationships within each clade matching those inferred among singleton genes (Figure 2.1). Pattern (ii) is shown by the analysis of *TSPY* (Figure A5) in that there is strong support for monophyly of all macaques except the Sulawesi macaque *M. nigrescens*, in sharp contrast to the analysis of singleton genes in which all Sulawesi macaques are a clade (Figure 2.1). When the 5′ and 3′ portions of *TSPY* are separately analysed, the role of gene conversion becomes apparent because pattern (iii) is shown by at least two independent aspects of this gene tree (Figure A6) . First, *M. arctoides* and one of the rhesus macaques have almost identical sequences at the 5′ end of *TSPY* but diverged sequences at the 3′ end (Figures 2.3, A6). Second, this same pattern is observed in *M. nigrescens*, presumably due to an independent gene conversion event that altered the 5′ sequence in the same way in this species but not in other closely related species of Sulawesi macaque (Figures 2.3, A6). A chimeric sequence was also observed in the reference sequence from the rhesus msrY, suggesting that laboratory artifacts such as PCR chimeras are an unlikely explanation for our observations. These data from *TSPY* could stem from an isolated event early in macaque evolution whose chimeric gene products were detected only in a subset of the species we examined. Alternatively, this could be an example of convergence via multiple independent gene conversion events.

Although not a focus of our study, these phylogenetic analyses supported a close relationship between *DAZ* copies on the msrY and the autosomal gene *DAZL1*, which is consistent with the proposal of [146] that this gene reached the msrY via transposition (Figure A1). Similarly, a close relationship between one paralogous msrY lineage of *XKRY* with the autosomal gene *XKR3* is also consistent with an inference of transposition from the msrY to the autosomes (Figure A7) [147].

2.3.2 Copy number variation on the macaques' msrY is low compared to apes

We then quantified AG copy number variation in five ampliconic genes (shown in Figure 2.3) from six to eight species of macaque monkey in seven to 13 individuals using qPCR. We assayed copy numbers of all known Old World Primate AGs (including genes found in just one copy in rhesus macaque but in multiple copies in the tribe Hominini), except for *TSPY*. *TSPY* was not analyzed because of high similarity among multiple partially gene converted regions (see above and Figure 2.3), which prevented us from developing a robust qPCR assay.

Results of the qPCR analysis are presented in Figure 2.4. Missing data are either a consequence of failed qPCR assays, as indicated by melt curve analysis (e.g. *RBMY*), or because substitutions in the primer sites prevented the use of a qPCR assay for select species (e.g. the *DAZ*-a assay in *M. ochreata*). We observed considerably less variation in copy number among the macaque species (summed coefficient of variation (CV) among five qPCR assayed AGs $= 0.396$) as compared to that among humans and chimpanzees (summed CV among five AGs $= 2.99$). Assuming constant generation times for all species, a slightly greater amount of time transpired among our sample of macaque species (≈ 3.1 million generations) compared to that between humans and chimps (≈ 2.2 million generations), suggesting that AG copy numbers evolve more slowly in macaques. Because generation time recently became longer in humans, the higher summed CV in the tribe Hominini is even more surprising if rates of gene family evolution were constant across the evolutionary phylogenetic lineages we examined. Nonetheless, as discussed below, these rates of AG family evolution in these lineages are not significantly different.

In general, larger gene families, such as *CDY* and *HSFY*, exhibited more copy number variation among macaque samples (respective CVs $= 0.201$ and 0.148, respectively) than smaller gene families, such as *RBMY* (CV $= 0.0474$) and *XKRY* (CV$< 10^{-6}$). This is consistent with the probability of gene duplication being proportional to the number of copies, which is a central feature of the model of gene family evolution used in our analyses discussed below. For *CDY*, one instance of intra-specific copy number polymorphism is suggested for *M. tonkeana*. However, this polymorphism is weakly supported in that the 95% confidence interval spans the copy number threshold. Although present in multiple copies in humans and chimps, *RBMY* and *XKRY* are single-copy in rhesus [54], and we recovered no evidence of multiple copies of these genes in the other macaque species surveyed.

In the reference msrY sequence for rhesus macaque, the *DAZ* gene is present in two copies and each copy contains tandemly duplicated exons. Based on the rhesus reference msrY, one of our qPCR assays interrogated a triplicated exon in the first *DAZ* gene, *DAZ1* (Figure 2.3; qPCR assay *DAZ*-bcd, see Supplemental Information), and another assayed a duplicated exon in the second *DAZ* gene, *DAZ2* (Figure 2.3; qPCR assay *DAZ*-a, see Supplemental Information) [146]. Our results indicate that, similar to the rhesus reference sequence, all of the macaque species we assayed have two copies of *DAZ* (i.e. one copy of *DAZ1* and one of *DAZ2*). However, within-gene variation in exon number was detected in *DAZ1* in *M. maura* (individual P001 had three copies instead of two), and in *DAZ2* in *M. hecki* and *M. tonkeana* (individuals PM638, PM561, and PM545 had four copies instead of three) (Figure 2.4).

2.3.3 Models of gene family evolution

We developed and evaluated the fit of 1310 models of gene family evolution to previously published data from chimp, human, and rhesus [54, 85, 94, 96] autosomes and msrY, as well as to our new sequence and qPCR data from AG and singleton genes from various species of macaque monkeys. The evolutionary models we considered allowed for unequal rates of gene duplication and deletion (or 'birth' and deletion, abbreviated as BD) or, alternatively, an equal rate of birth and deletion (L model, where $\lambda \equiv b = d$, following [79, 84]).

Information from completely sequenced Y chromosomes indicates that not all gene families were present on the ancestral Old World Primate msrY, including *TGIF2LY*, *PCDH11Y*, and *VCY* [89]. Therefore, some of the models we evaluated allow a gene family to appear through transposition or to reappear after extinction (abbreviated I for 'innovation,' following [81]). We either estimated the innovation rate (i.e. the transition probability from $0 \to 1$ copy, Pr$(X_{n+1} = 1|X_n = 0)$) independently from the birth/deletion rate(s) (Pr$(X_{n+1} = k + 1|X_n = k), k > 0$; these models are abbreviated LI), or we set the innovation rate equal to the birth rate (Pr$(X_{n+1} = 1|X_n = 0) = b$; these models are abbreviated $B = ID$, if $b = i$, or $L = I$, if $\lambda = i$).

We used a threshold method to assign probabilities to each discrete gene copy number from the continuous qPCR data for each sample (see Methods and Supplemental Information). Our method accommodates uncertainty in copy number inferences based on the qPCR assays, and allows for missing data. Thus we were able to include in our analysis genes for which we had one or more unique sequences but for which we lacked qPCR data. For example, for *AMELY*, we had one unique sequence from each of 13 macaque individuals but we did not have information on copy number variation of this locus: this was considered evidence for one or more *AMELY* copies in each of these individuals. We also lack autosomal gene family data from the eight macaque species whose msrY we investigated with qPCR and sequencing, so these autosomal data were also treated as missing.

These models also considered the possibility of two types of rate heterogeneity. The first type, hereafter "lineage heterogeneity", allows for a different rate of gene family evolution – but the same model of evolution – between the Hominini lineages and the other Old World primate lineages, as previously identified by [85]. The second type, hereafter "gene heterogeneity", allows for different rates and different evolutionary models of gene family evolution among different classes of gene families. The separate categories considered were the gene families in autosomal DNA, singleton gene families on the msrY, and AG families on the msrY, and various combinations of these categories. We chose to exclude the *TSPY* gene family from our analysis because this gene family is a prominent outlier due to its exceptionally high copy number in humans [94]. Inclusion of *TSPY* data in the analyses yielded significantly higher parameter estimates as compared to when this gene family is excluded (data not shown).

Models with an independently estimated innovation parameter are not biologically plausible

Figure 2.5 summarizes representative models that we explored including models in which there is no lineage or gene heterogeneity (Figure 2.5A) and models in which msrY AGs, msrY singletons, and/or autosomal gene families each have a distinct mode of evolution (that is, different evolutionary models and/or different parameter values for each gene category) (Figure 2.5B-E). For each tree topology depicted in Figure 2.5, 10 distinct evolutionary models were initially considered (namely L; LI; $L = I$; BD; $B = ID$; and each of these with or without lineage heterogeneity yielding ten models in total, see Methods. A total of 1310 models were considered including all gene categories pooled (10 models); autosomes and msrY singletons pooled but AGs separate (100 models); msrY singletons and AGs pooled but autosomes separate (100 models); msrY AGs and autosomes pooled but msrY singletons separate (100 models); and autosomes, msrY singletons, and msrY AGs each separate (1000 models).

In order to check whether inferences from our models were biologically plausible, we compared the msrY gene maximum *a posteriori* copy numbers predicted at ancestral nodes to information from completely [54, 94, 96] or partially sequenced [89, 148] Y chromosomes in primates in order to determine if the predicted gene family presence/absence is consistent with this external information. For example, the existence of a pseudogenized copy of *USP9Y* in the chimp and functional copies in the human and rhesus Y chromosomes indicate that this gene was present ancestrally in Old World Primates [54], but the LI model incorrectly inferred that this gene was absent in the ancestor and transposed independently in both human and rhesus.

Overall, this exercise indicated that models where the innovation rate is independently estimated tended to overestimate the number of innovation events to the msrY. Specifically, Figure A19 illustrates that, compared to other models without the innovation parameter, the LI and LI+lineage heterogeneity models have poor sensitivity in that they fail to identify gene families that were present in the ancestors and instead infer them to be instances of innovation. However, models that allowed innovation were able to correctly identify the autapomorphic human transposition of two singleton gene families (*TGIF2LY* and *PCDH11Y*) that were absent in the ancestor of Old World Primates (true negatives) [54, 94], indicating higher specificity relative to models without an innovation parameter (Figure A19). But none of the models with innovation were able to correctly identify *VCY* as absent on the most recent common ancestor (MRCA) of Old World Primates, as was proposed by [54, 89]. For these reasons, we excluded these models from our analysis, leaving 840 models out of the original 1310 for consideration, including all gene categories pooled (8 models); autosomes and msrY singletons pooled but AGs separate (64 models); msrY singletons and AGs pooled but autosomes separate (64 models); msrY AGs and autosomes pooled but msrY singletons separate (64 models); and autosomes, msrY singletons, and msrY AGs each separate (640 models).

msrY AG families evolve faster than msrY singletons and autosomes, and msrY singleton may evolve faster than autosomes

Figure 2.6 illustrates parameter estimates from the six best models, which together comprise $> 95\%$ of the cumulative Bayesian information criterion (BIC) weights. These six models share several consistent features. Each supports a separate model of evolution of AGs from singletons and autosomes. In all six models, the estimated rates of AGs are significantly higher ($\approx 5 - 1500$ fold for λ and $\approx 5 - 10$ fold for i) than for the autosomes, and have confidence intervals that do not overlap with those of the autosomal or singleton and autosomal gene categories. Four out of the six models, which correspond to 92.0% cumulative BIC weights across 840 models, support a separate model of evolution for singletons and autosomes. In each of these six models, the birth/deletion rate, λ, for AGs is significantly higher than λ for singletons in the Old World Primate lineages other than those of the tribe Hominini, and than the deletion rate of singletons in Hominini. Parameter values for msrY singletons and AGs are presented in Tables A11 and A12, respectively.

Despite the difference in the CV of AG copy number between Hominini and other Old World Primate lineages discussed above, we did not recover significant support for lineage heterogeneity for AGs. However, lineage heterogeneity was significantly supported for the autosomes, or for the combination of the autosomes and msrY singletons, with much faster deletion rates ($\approx 60 - 160$ fold) in Hominini than in other Old World Primate lineages (birth/innovation have rates $\approx 0.82 - 0.86$ fold smaller in Hominini than Old World Primates). Lineage heterogeneity is supported for singletons in two out of the top six models, corresponding to 80.7% cumulative BIC weights across 840 models, in which this gene class is considered separately from the autosomes.

In three out of the six most preferred models, corresponding to 29.2% cumulative BIC weights across the 840 models, the preferred model for AGs evolution did not include an innovation parameter. In contrast, a strong preference for models for the msrY singletons with innovation equal to the birth rate is probably explained by the presence of two X-transposed gene families (*PCDH11Y* and *TGIF2LY*) within this category.

Gene family evolution is best explained by a single model for msrY singletons but not AGs

The analysis of autosomes, singletons, and AGs discussed above universally favor models in which AGs evolve under a different model from the rest of the genome, including singletons on the msrY. Considering just the AGs independently from the other gene categories, the top three ($L = I$, $L = I+$lineage heterogeneity, and $B = ID+$lineage heterogeneity) of the total eight models have just 61.1% of the cumulative BIC weights, suggesting that there was little power to distinguish between models. For a given parameter or suite of related parameters, estimates across the models tended to be similar within each gene category. For example, over most models, the rate of AG birth/deletion (λ) or AG birth and AG deletion tended to be around 0.13 events per million generations, or even higher in human and chimpanzee lineages when lineage heterogeneity is allowed.

When we considered just the msrY singletons independently from the other gene categories, the $L = I+$lineage heterogeneity model was preferred with 77.7% of the BIC weight, and the second best of the eight models, $B = ID+$lineage heterogeneity, was supported by 21.1% of the BIC weight. When the rates of birth and deletion were allowed to differ, the deletion rate in the Old World Primate lineages other than Homonini was inferred to be nearly zero, suggesting that genes present in the ancestor of these Old World Primates are also generally still present in macaques.

2.4 Discussion

In order to better understand gene family evolution of duplicated ampliconic genes on the primate msrY, we collected qPCR and sequence data from various species of papionin monkey and we analyzed copy number information and DNA sequences from published autosomal and Y chromosomes. We built gene trees to qualitatively evaluate evidence for gene conversion in new and previously available sequence data, including pseudogenes. We then evaluated alternative scenarios of gene family evolution that either imposed or relaxed assumptions of equal rates of gene copy birth, deletion, and innovation; rate consistency over time (lineage homogeneity); and consistency of the model of evolution across gene families of msrY singletons, msrY AGs, and autosomes (gene homogeneity). We recovered strong evidence of gene conversion in many AGs within the msrY, including several novel examples. In *TSPY*, we recovered possible evidence for multiple independent partial gene convergence events in the same gene, each of which resulted in a chimeric gene product with the 5′ and 3′ ends having originated from the same ancestral copies.

We also found that gene families evolve significantly faster in msrY AGs than in autosomes, and generally faster than msrY singletons, or perhaps similarly to the birth/deletion rate of msrY singletons in the tribe Hominini when this rate is allowed to vary among lineages (Figure 2.6). These results highlight the distinctive nature of AG family evolution, and suggest that this distinctiveness is not solely a consequence of residence on the msrY (because they evolve differently from singletons on the msrY) or genome-wide variation among evolutionary lineages (because they also evolve differently from autosomal gene families).

Another finding that emerged from our analysis is that the inclusion of an independently estimated innovation parameter resulted in biologically unrealistic estimates of other model parameters. Because relatively few genes have been introduced to the msrY since the diversification of Old World Primates [54], it is unsurprising that there were insufficient innovation events to inform the innovation rate for msrY genes in the species that we investigated. The inclusion of an independently estimated innovation parameter may therefore prove more useful in studying gene family evolution across a broader phylogenetic scope in primates, or in other clades.

2.4.1 What determines AG evolution?

Our finding of a higher rate of gene family evolution of msrY AGs compared to autosomal gene families matches population genetic expectations if duplicated copies are mildly deleterious and more likely to be observed as intraspecific polymorphisms or fixed differences between species in genomic regions with a small N_e. However, this fails to explain why msrY AGs evolve faster than msrY singletons, because both of these gene categories reside on the msrY. For msrY singletons but not AGs, a deletion event represents extinction of the entire gene family within a species, and a birth event leads to a doubling of gene dosage. Thus changes in singleton copy number presumably have a more substantial biological effect than in AGs. That singleton gene families evolve more slowly than those of AGs suggests that singletons are under tighter dosage constraints, and thus more resistant to variation in copy number [89, 99]. Consistent with this speculation, there are multiple examples (including independent examples from the same gene) of msrY-linked loci being lost after a copy is translocated to the autosomes [123].

We classified AGs based on whether they were observed to have experienced gene conversion in humans, chimps, or macaques; AGs thus each have at least two copies of a gene in at least one of these taxa. However, several lines of evidence argue that the separate categories of singletons and AGs are warranted for reasons beyond their copy number in these three taxa. First, msrY singletons and AGs differ substantially in their expression patterns. Singletons are broadly expressed across tissue types and developmental stages [94], and frequently have a counterpart on the X chromosome that escapes X-inactivation in females [55], resulting in a similar protein stoichiometry across both sexes. AGs, in contrast, have testis-specific expression and no counterpart on the X [94, 99]. That all AGs have testis-specific expression suggests that this specificity arose prior to the amplification of AG copy number [99]. Indeed, at the nucleotide level, genes in the msrY singleton and AG categories evolve at significantly different rates, with the rate ratio of nonsynonymous to synonymous substitutions per site being higher in AGs [54]; this disparity is expected based on the different expression patterns [149–151].

Second, msrY singletons do indeed undergo duplication as evidenced by duplications of the *SRY* gene in European rabbits [152] and rats [153], although if the duplication results in a palindrome that undergoes gene conversion, as is the case in the *SRY* gene in rabbits, the nature of gene family evolution may more closely match that of AGs (see our discussion of gene conversion below). Evolutionary history also partially distinguishes singletons and AGs [89, 94, 109, 154]. In primates, all msrY protein coding singleton genes were either present on the mammalian ancestral X and Y chromosomes prior to divergence, or delivered to the Y chromosome from the X (the so called "X-degenerate" and "X-transposed" gene classes, respectively) [94]. However, only about half of the AGs (*HSFY*, *RBMY*, *TSPY*, *VCY*, and *XKRY*) evolved in this fashion [89, 94, 100]. Two AG families in humans are of unknown origin (*BPY2* and *PRY*), one (*CDY*) was retrotransposed to the msrY from an autosome prior to the diversification of Eutherian mammals [100, 155], and one (*DAZ*) was transposed from an autosome prior to the diversification of Old World Primates [100, 146]. Of the AGs with ancient ancestry on the msrY [89, 100, 155], four (*CDY*, *RBMY*, *TSPY*, and *XKRY*), experienced a pronounced expansion in copy number in the ancestor of catarrhines, and two (*HSFY* and *XKRY*) were subsequently lost in chimps [96, 100, 156].

And third, there are msrY-linked, multi-copy genes which we did not consider to be AGs. *RPS4Y1* and *RPS4Y2*, for example, are diverged in protein sequence and have not undergone gene conversion for over 35 million years (My) [157, 158] and were thus considered to each be a msrY singleton. Another unusual pair of genes is *CYorf15A* and *CYorf15B* (known together as *TXLNGY*), which code for the msrY-linked paralogs of the 5′ and 3′ portions, respectively, of the X-linked gene *TXLNG* [18]. This gene probably split up into two singleton genes prior to the divergence of Old World Primates [89]. For our analysis, we followed the annotation of [54] for these genes as each functional msrY singletons.

Thus multiple variables distinguish singletons from AGs beyond gene conversion including their expression patterns, molecular evolution, and aspects of their evolutionary history, and these are not strictly a consequence of copy number. Arguably differences in expression patterns and molecular evolution were present ancestrally and thus potentially causal to the observed differing nature of gene family evolution, rather than being consequences of this difference.

2.4.2 Lineage heterogeneity in autosomes and msrY singletons

The support of our study for lineage heterogeneity of gene family evolution in autosomes is consistent with the study by [85], from which we obtained the autosomal copy number data. This consistency was recovered even though that study used a different statistical approach, with the likelihood approximated using the maximum *a posteriori* (MAP) ancestral state [159] instead of our approach which calculates the likelihood by summing across all ancestral states. Our findings that the birth rate outpaces the deletion rate in autosomes also agrees with previous findings that included data from other Great Apes [66]. In particular, for some of the preferred models with lineage heterogeneity, the estimated deletion rate of autosomes was much higher in Hominini than the estimate of the deletion rate in the other primate lineages.

In the autosomes, the rate of segmental duplications is slower in orangutans than other Great Apes [62], but higher in gorillas and chimps than in humans [65, 66]. However, in rhesus, although small structural variants are abundant [69], these may be a result of mobile element insertion by retrotransposition, as opposed to the larger scale duplication events that are generally responsible for the expansion of gene families [68]. Heterogeneity in the rate of gene family evolution might also arise if this rate was influenced by the extent of sperm competition, as has been proposed in chimps [96, 110, 111, 148, 160]. Inconsistent with this hypothesis, however, are the observations that macaques also have high sperm competition [161–165] yet we recovered relatively low variation in AG copy number among and within macaque species compared to the Hominini.

Some clues toward a mechanistic explanation for lineage heterogeneity in msrY singletons may be gleaned from the three available completely sequenced Old World Primate Y chromosomes, which are distinguished from one another in chromosomal structure. In terms of nucleotide sequences, the rhesus macaque msrY comprises about half (≈ 11 megabase pairs) of the euchromatin and about one twentieth of AGs (≈ 0.5 megabase pairs) compared to humans and chimps [54]. A higher content of inverted repeats and repetitive sequences in the Great Apes may promote chromosome fragility, and increase opportunities for duplication or deletion through non-allelic homologous recombination or microhomology-mediated events [60, 61].

Polymorphism in copy number variation in the autosomes appears to be influenced by demographic changes such as bottlenecks. The Western Chimpanzee, for example, has a high level of polymorphism in duplications and deletions, and also has genomic signatures of a population bottleneck [66]. By comparison, there is also evidence for a dynamic demography, including recent population decline in the western population of the tonkean macaque [11], but our analyses failed to recover compelling evidence of copy number polymorphism based on a limited sample (three individuals; Figure 2.4). This disparity could be attributed to a lack of statistical power of our small data-set. We did, however, discover polymorphism at an exon repeat within a paralog of *DAZ* in this population.

2.4.3 Gene conversion in primates and beyond

In general, gene conversion occurs more rapidly in palindromes on the msrY than among palindromic sequences in the autosomes [133], suggesting that the nature of natural selection on duplicates on the Y chromosome may differ from that on duplicates elsewhere in the genome. Relatively few cases of gene conversion within genes on the Y (or W) chromosomes are known beyond those identified in primates based on the complete Y chromosome sequences of a human, a chimp, and a macaque [105]. Known examples include genes that are also arranged in palindromes, including duplicated *SRY* genes in the European rabbit [152], copies of the *HINTW* and *CHD1W* genes in various birds [129, 130], and several genes in cows [108, 166]. Thus our discovery of several clear examples of gene conversion add to a relatively small list of examples from species whose Y chromosomes have yet to be completely sequenced.

Previous studies have considered the evolution of AGs from a theoretical perspective. Ancillary genes that do not themselves undergo gene conversion could catalyze gene conversion of other duplicated genes; these theoretical genes are called recombination modifiers [167]. Using population genetic parameter estimates from humans, simulations indicate that the fixation rate of msrY-linked recombination modifiers can be faster than that for a neutral variant [102]. Simulations and analytical models that jointly consider the phenomena of gene conversion and gene duplication suggest that gene conversion can promote the persistence of gene duplicates on the msrY by resuscitating copies that have undergone deleterious mutation [104]. In this case, gene conversion is not influenced by the fixation probability of a newly arisen duplicate. The theoretical findings of [104] may be supported by the empirical finding of a longer average lifespan of multi-copy genes on the mammalian msrY compared to single-copy genes [99]. The effects of mildly advantageous mutations on duplicates that undergo gene conversion has also been explored, with the conclusion that gene conversion can increase the rate of adaptive evolution [112]. This study noted that gene conversion can be biased, for example, by favoring GC over AT base pairs, and that this bias becomes important when the rate of gene conversion is high [112], which it is on the Y chromosome [113]. Moreover, there is significant evidence of GC biased gene conversion in macaques [11] and other primates [168–171]. Thus, while gene conversion is uniquely possible in multi-copy gene families, theoretical studies suggest that this phenomenon may promote the persistence of gene families on the msrY. In the background of the msrY, whose evolution is dominated by deleterious mutations and strong linkage effects, the role of gene conversion as a conservative force can also lead to greater adaptive evolution in AG families [105]. Overall, gene conversion is a plausible causative factor – in addition to being a consequence – of the distinctive nature of AG family evolution.

2.4.4 Caveats and future directions

qPCR assays

An advantage of studying closely related species is that we were able to use multiple gene copy-specific assays to quantify copy number variation across a protracted period of evolutionary time. We anticipate that our qPCR assays accurately identified copy numbers for orthologs that have high sequence identity to the rhesus AG sequences, based on (i) comprehensive sequencing of qPCR primer sites during the development of our qPCR assays, (ii) the high number of technical replicates per individual assay ($n = 4 - 36$ for the experimental samples and $n = 11 - 34$ for the rhesus reference sample), and (iii) the conservative measures we took to identify and exclude replicates with inconsistent reaction efficiencies. However, a drawback (that is difficult to overcome without complete Y chromosome sequences from each species) is that some paralogs may have gone undetected by our qPCR assays if orthologous data were not available in the rhesus macaque due to deletion in an ancestor of rhesus after divergence from the other macaques we assayed.

However, there are two reasons to suspect that our assays did in fact evaluate most or all of the gene families on the macaque msrY. First, if we assume that this rate is similar to that estimated from the available data, the posterior distribution of AG ancestral copy numbers under each of the models, and the relative probability of each model, we would expect, using model averaging, only 0.613 autapomorphic deletion events along the rhesus lineage among all AG families (or, 0.620 deletion events under the preferred model, $L = I$). Thus we do not anticipate major differences between the rhesus macaque and the other macaques we surveyed in gene content on the msrY. Second, in our model gene birth and deletion occur at rates that are proportional to the number of copies. For this reason, even if our qPCR assays did systematically fail to identify AG paralogs because they were deleted in the rhesus lineage, this should not bias our estimate of the per copy birth and/or deletion rates. It would, however, mean that we have less information in our data and therefore the confidence intervals on the parameter estimates are larger than they would be if we had more complete data. Further characterization of inter- and intra-specific variation in copy number of primate AGs will undoubtedly increase our understanding of these inferences and increase their phylogenetic precision. At this time, however, accurate quantification of copy number variation on the msrY is hampered by the repetitive nature of this genomic region, a dearth of completely sequenced Y chromosomes in primates, and by technical complexities associated with assaying copy numbers of genes that are frequently homogenized by gene conversion.

Evolutionary models

Our models made several simplifying assumptions that may poorly reflect the actual biological events that occurred during the evolution of gene families on the msrY. For instance, we assumed independence among gene families even though gene families on the msrY are genetically linked. This assumption was made in order to simplify the likelihood calculation. We also assumed that copy number changes of AG families proceed in a stepwise fashion, as has been previously supported [102]. However, it is conceivable that a few rare events may be responsible for multiple duplications happening at the same time – even among different gene families – for example, via crossing over [115, 116, 137, 172] or chromothripsis [80, 173] of the msrY. We also did not include a role of evolutionary "strata" in msrY gene family evolution. However, because the most recent msrY arose prior to the diversification of our species of interest [54], this seems like a reasonable simplifying assumption for our data. Another factor not considered by our models is the possibility that epistatic interactions between these genes and genes encoded elsewhere in the genome could influence gene family dynamics in unique ways. In particular, if there are favorable combinations of Y-linked and non-Y-linked alleles across genes whose protein products interact, this could favor the translocation from the autosomes or the X chromosome to the Y chromosome in order to prevent these associations from being lost due to recombination. Support for this possibility has been found in fruit flies in which the same Y chromosome exhibits considerable heterogeneity in fitness in different genetic backgrounds [174], and is associated with differential expression of autosomal genes [175].

2.5 Conclusions

This study found multiple novel examples of gene conversion among AGs on the Old World Primate msrY, including one gene that appears to have undergone multiple independent gene conversion events in different species and with similar recombination margins. These independent events yielded chimerical gene products whose evolutionary histories differ between the $5'$ and $3'$ ends of the affected exon. Using data from qPCR, gene sequences, and completely sequenced msrY of a human, chimpanzee, and rhesus macaque, we also demonstrated that AGs on the msrY evolve significantly faster than msrY singletons and autosomal gene families, and that AGs are perhaps better approximated by an altogether distinct model of evolution than those that best approximate the other gene categories. We speculate that the distinctive nature of msrY AGs is a consequence both of the high frequency of gene conversion and natural selection acting on male-specific function of these genes.

2.6 Methods

2.6.1 Genomic DNA extraction & Sequencing

The origins and genomic DNA (gDNA) extraction of samples used in this study are summarized in [143] with the exception of one rhesus macaque, a baboon, and a mandrill sample which were obtained from the Toronto Zoo. Genetic samples for this project were obtained using methods approved by the Institutional Animal Care and Use Committee (IUCAC) at Columbia University.

Sequencing of *TSPY* and *SRY* loci confirmed species identity as determined by [144, 145] and argued against the possibility of inter-specific contamination of Y-chromosome DNA. AG exons in papionins were amplified and sequenced by polymerase chain reaction (PCR) using primers designed from rhesus macaque Y-chromosome bacterial artificial chromosome (BAC) sequences with high similarity to human AG exons. For all AG loci, multiple primers for at least two exons and/or at different sites were created whenever possible to minimize the possibility of false negative (failed) amplifications due to divergence of primer sites.

2.6.2 Phylogenetic estimation

A Y-chromosome phylogeny for 14 male macaques (Table A2), human, chimpanzee, and marmoset was estimated using concatenated nucleotide sequences from up to nine msrY singletons. This analysis included novel sequences for three macaque samples (one *M. arctoides* and two *M. maura* samples), sequences from a marmoset that were identified using BLAST [176], and several other species from a previous study [143] (GenBank accessions in Table A4). Primers for single-copy, msrY-linked exons and GenBank accessions for the remaining 11 macaque samples, human, and chimpanzee are listed in [143]. The total alignment length was 6 185 bp, and the alignment length after excluding positions with gaps was 6 167 bp. The time-calibrated phylogeny was built in BEAST v1.7.5 [177], assuming mean divergence times of 6 My and 30 My for the ancestor of the tribe Hominini and other Old World Primates [178–180], respectively, with model selection, molecular clock calibration, and other analytical details provided in the Supplemental Information.

A similar procedure was used for generating the AG trees. Pseudogenes were identified in the completed human, chimp, and rhesus macaque msrY using the functional gene sequences as BLAST queries. In addition to the two calibration dates listed above, a mean divergence time of 8.5 My was assumed for papionins [179, 181, 182] when a putative functional AG ortholog was identified for either mandrill or baboon.

2.6.3 qPCR

Quantitative PCR was performed in accordance with the minimum information for publication of quantitative real-time PCR experiments (MIQE) guidelines [183, 184]. Raw data in RDML format for all assays can be found at RDML database (`http://rdmldb.org`, [accession TBD]). Further details of macaque sample processing and gDNA extractions are available in the Supplemental Methods. Standard curves are shown in Figures A17-A18. gDNA from rhesus macaque was used as a reference sample and assumed to have the same ampliconic gene copy numbers as the individual sequenced for the rhesus macaque Y-chromosome project [54]. Since ampliconic gene copy numbers are unknown for macaque species other than the rhesus macaque and qPCR primers are specific for the genus *Macaca*, no other controls with known copy numbers were available. Therefore, the remaining 13 male macaque samples representing eight species are all part of the experimental group (Table A2). Samples with divergent sequences at qPCR primer sites (namely DAZa, see Figure A9), poor assay specificity as determined by melt curve analysis, or assay efficiencies consistently different from the median assay efficiency were excluded from the analysis for the problematic gene family.

qPCR was used to determined the mean expression in each experimental macaque sample relative to rhesus macaque. The known single-copy Y-linked gene *SRY* was used as a reference to confirm the invariant, single-copy status of *TSPY1* and *XKRY* (Figure 2.3) for all of the experimental samples. Then, since *TSPY1* and *XKRY* had satisfactory mean stability (geNorm M-value) measures and coefficients of variation (Table A9), all three loci were used as reference genes to calculate the relative expression for the remaining six loci. Finally, the relative expression for each experimental sample was rescaled to gene copy number using the copy numbers from the rhesus macaque Y-chromosome project [54]. Additional details on qPCR are provided in Supplemental Methods.

2.6.4 Copy number estimation from qPCR and sequence data

In order to be able to put the qPCR copy number data into an evolutionary perspective, we needed to generate estimates of the discrete copy numbers for each gene and sample. We assumed that the estimated copy numbers are Normally distributed with a standard deviation equal to the estimated standard error. We assigned a probability to each copy number integer (> 0) by calculating the cumulative probability under the density curve for intervals at $(0, 1.5, 2.5, 3.5, \ldots)$. These probabilities were used as the likelihood for the extant taxa, which allowed us to incorporate the uncertainty from the qPCR estimate into our models.

Similarly, for genes and/or samples for which we did not perform qPCR, we used the number of unique sequences observed as an estimate of the minimum number of gene copies present. All copy numbers smaller than the number of unique sequences observed were assumed to have a likelihood of zero, while copy numbers equal to or greater than the number of unique sequences have a likelihood of one.

2.6.5 Gene family evolution

A homogeneous time Markov process with an arbitrary finite number of states was used to model gene family evolution along the primate phylogeny. Gene duplication and deletion events were modeled using a continuous-time Poisson process where the probability per unit time of an event is proportional to the number of copies. Models BD and $B = ID$ allowed unequal rates of gene duplication, ‘birth,’ and deletion. Models $L = I$, LI, and $B = ID$ had an innovation parameter that describes the probability of a gene family moving from zero copies to one copy. Although an innovation parameter has previously been used to model lateral gene transfer of gene families in Prokaryotes [82], innovation may be of particular importance to msrY-linked gene families since it can be used to describe events such as the acquisition of novel gene families on the msrY to the autosomes, the acquisition of a gene family to the msrY, the suppression of recombination in part of the pseudo-autosomal region resulting in novel msrY-linked genes, and the putative resuscitation of an extinct gene family by gene conversion of complementary pseudogenes. In models without the innovation parameter, a copy number of zero is an absorbing state; therefore models without innovation assume that each gene family was present in at least one copy in the MRCA of all taxa, while models with innovation do not make this assumption. Furthermore, models without innovation have a limiting distribution at zero and a quasi-stationary distribution at the largest copy number state; both of these are biologically unreasonable distributions for the ancestral state at the MRCA of any gene. We assumed a generation time of five years for all primates [185, 186].

2.6.6 Missing values and heterogeneous rates

We did not have complete copy number estimates (or minimum values) for all gene families and all macaque species investigated. Therefore, in order to fit the complete data to a single model, we had to accommodate missing data by assigning a likelihood of 1 at all states for genes and taxa with missing data. We implemented rate heterogeneity among lineages in a way that is analogous to its implementation in CAFE [85].

2.6.7 Analysis of whole genome and msrY data

We downloaded the autosomal gene family size data from [85] and kept only the data from human, chimp, and rhesus macaque. For computational efficiency, we excluded one gene family that has a copy number of > 400; this left a total of 9904 autosomal gene families. We supplemented these data with the complete msrY gene family size data from [54]. We also included in the dataset all of the macaque species by inputting NAs for the autosomal data and the copy numbers determined as described above for the msrY-linked gene data.

All model fitting was performed in `R` v3.1.0 [187] using custom functions that were based on the function `ace` from the `R` package `ape` v3.0-6 [188]. These functions are available upon request and will be distributed as an `R` package.

2.7 Figures

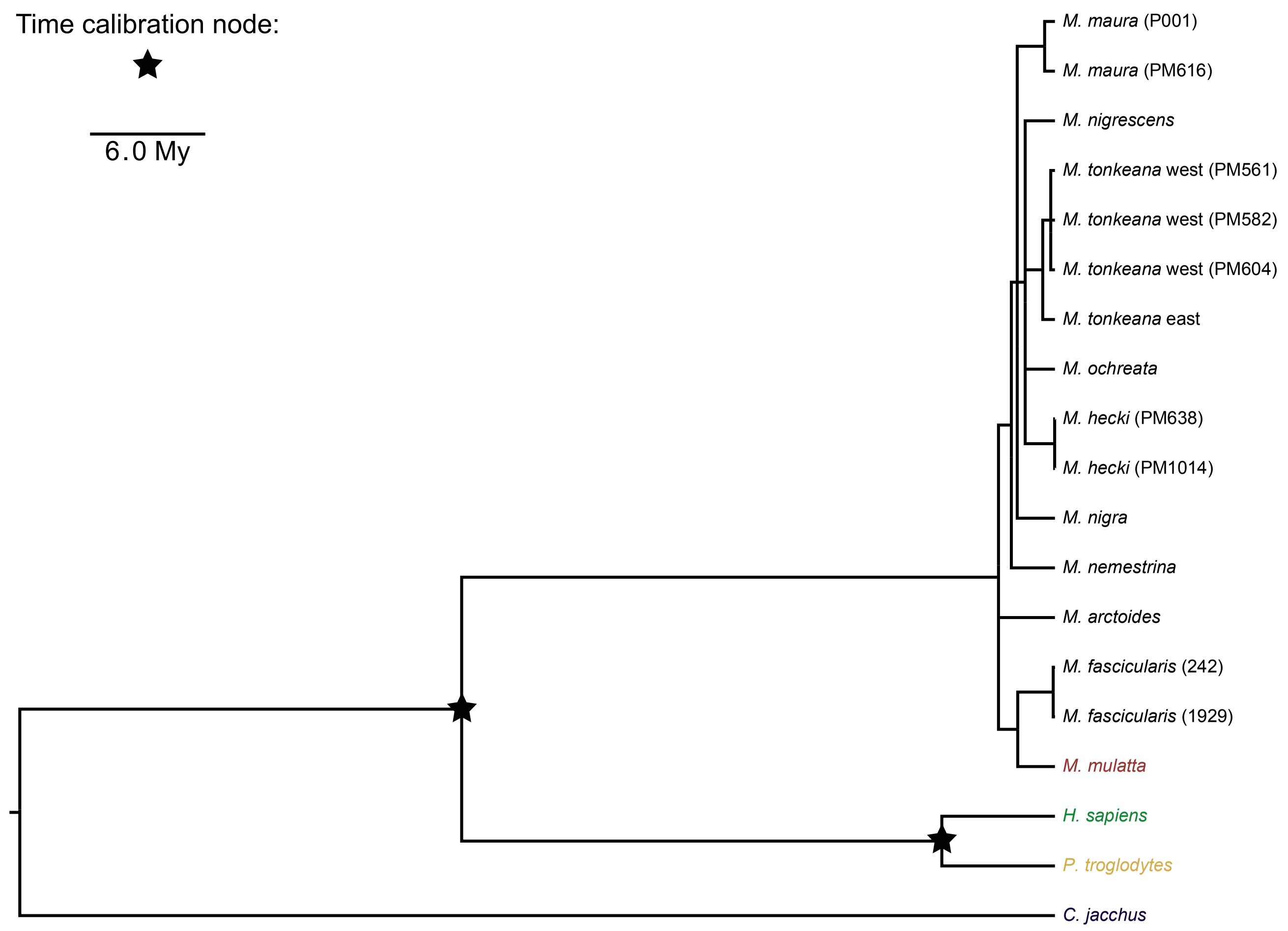


Figure 2.1: **Maximum clade credibility tree of nine concatenated msrY-linked single copy genes.** Nodes with less than 95% posterior probability are collapsed. The nodes used for time calibration are labeled with stars. The total alignment length is 6 185 nt. The tree was built in BEAST.

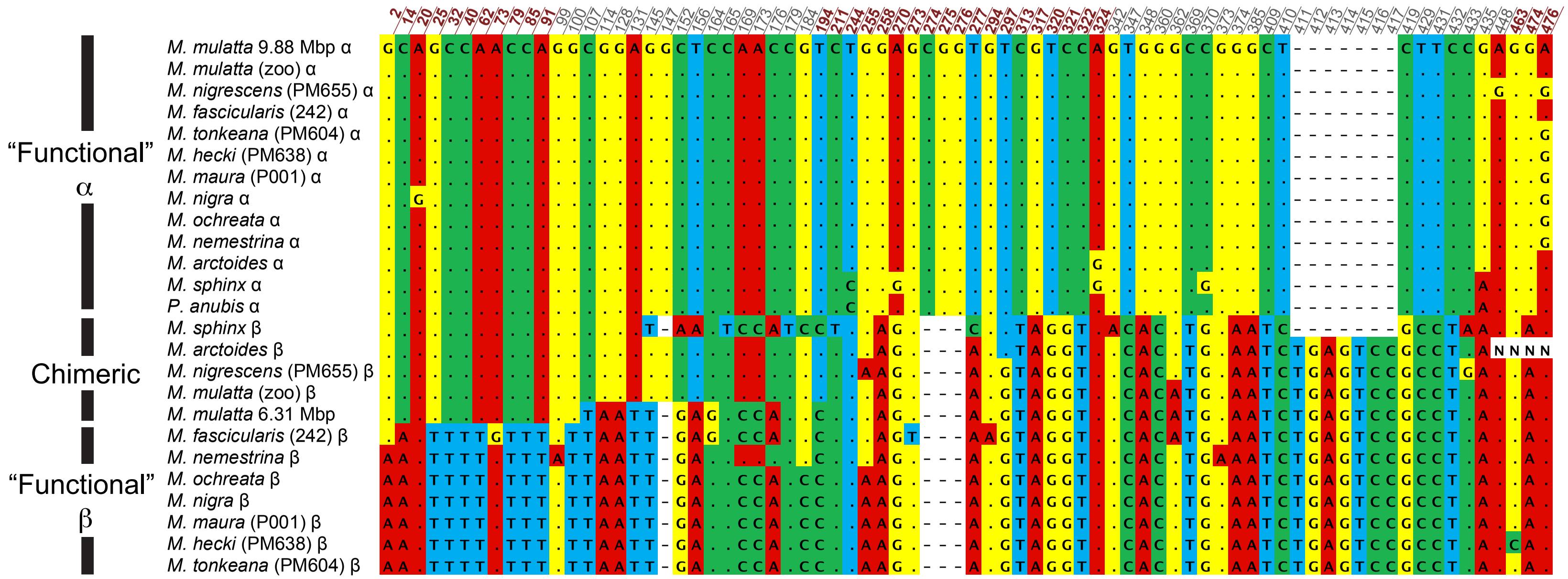


Figure 2.2: **Papionin monkey variable sites in the *TSPY* multiple sequence alignment.** Paralogs labeled "α" correspond to sequences of *TSPY2-5*, paralogs labeled "β" correspond to sequences of *TSPY1*, and paralogs labeled "chimeric" have sequences similar to *TSPY2-5* at the 5′ end and similar to *TSPY1* in the 3′ end. Dots represent sites that are not different from the rhesus macaque *TSPY2-5* Y-chromosome sequence; letters represent sites that differ. The location (in bp) after primer TSPYex3-5For (Table A3) are indicated above the alignment; numbers in dark red and gray font show exons and introns, respectively.

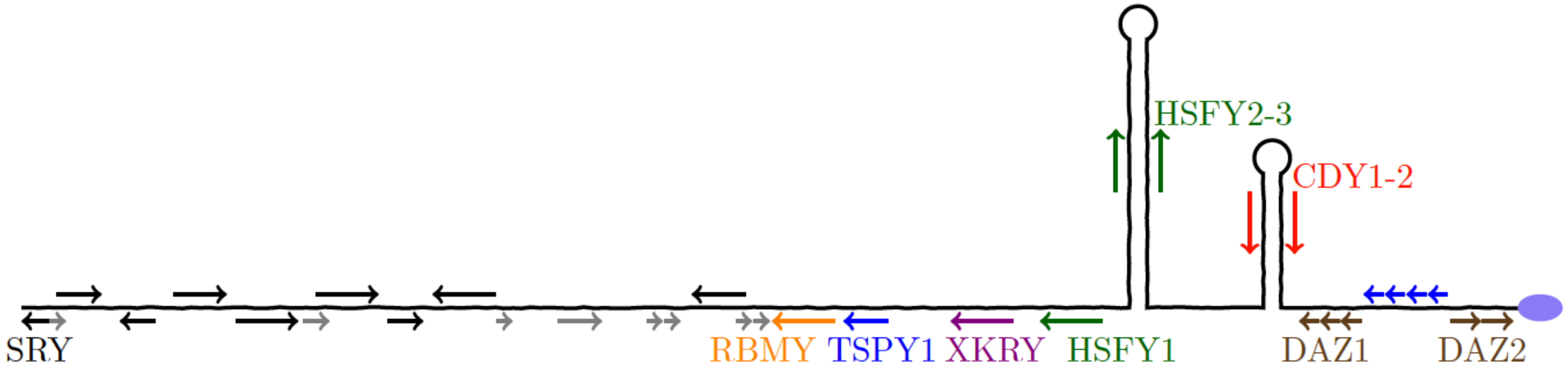


Figure 2.3: **Locations of qPCR-assayed genes on a schematic of the rhesus macaque msrY.** Arrows indicate the orientation of protein coding genes. Black arrows indicate singleton loci sequenced in this study, coloured arrows indicate AG loci sequenced in this study, and gray arrows indicate loci not sequenced in this study. Labeled loci indicate genes whose copy numbers were assayed using relative qPCR; *SRY*, *TSPY1*, and *XKRY* were used as reference genes while the others as well as *XKRY* are experimental genes. The blue unlabeled loci correspond to *TSPY2-5*, which was not assayed by qPCR. The multiple arrows for *DAZ1* and *DAZ2* illustrate the exon duplicates within each gene. The two stem-loop structures on the right illustrate palindromes (i.e. inverted repeats) in the ampliconic region and the purple ellipse on the far right illustrates the centromere.

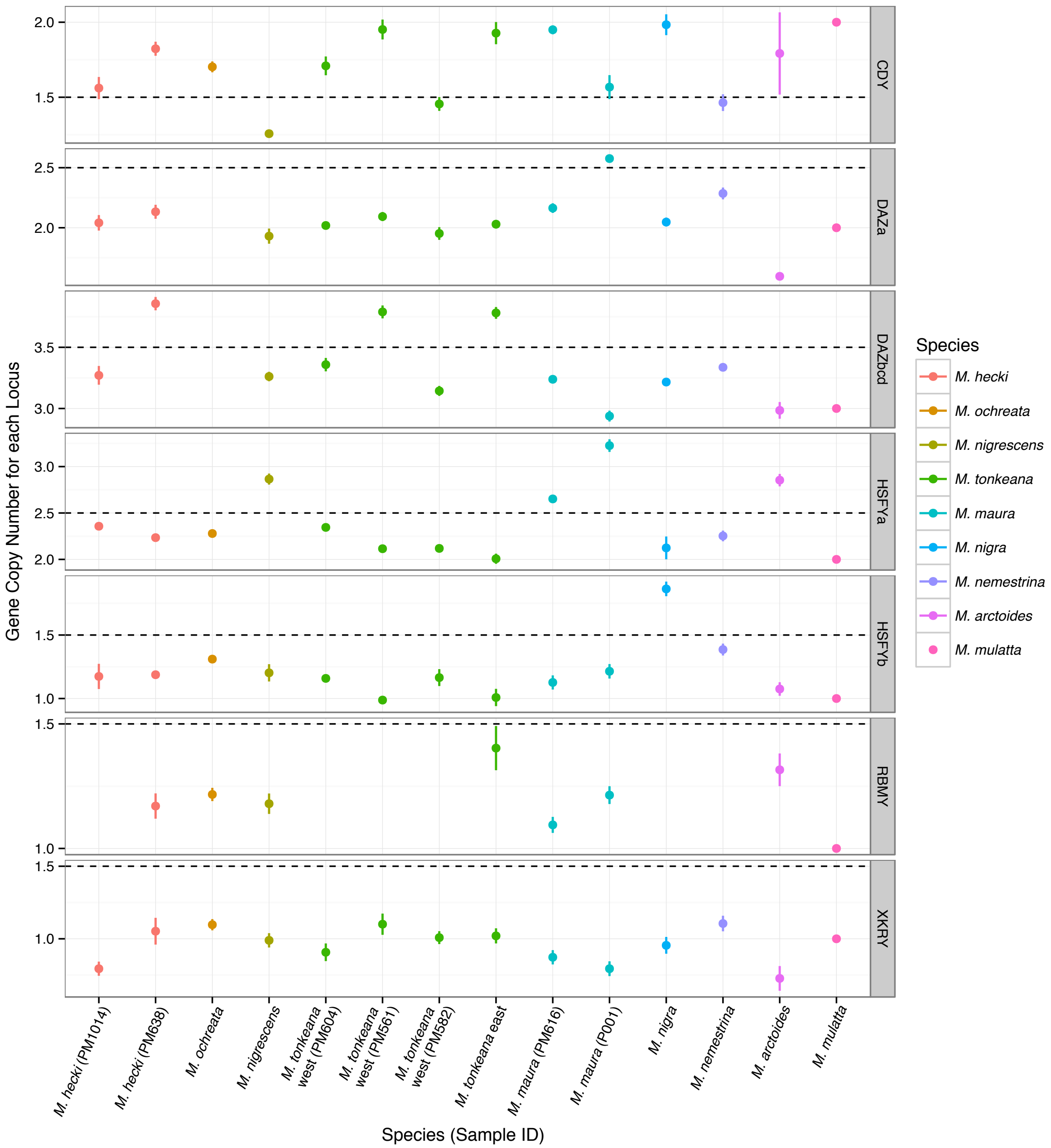
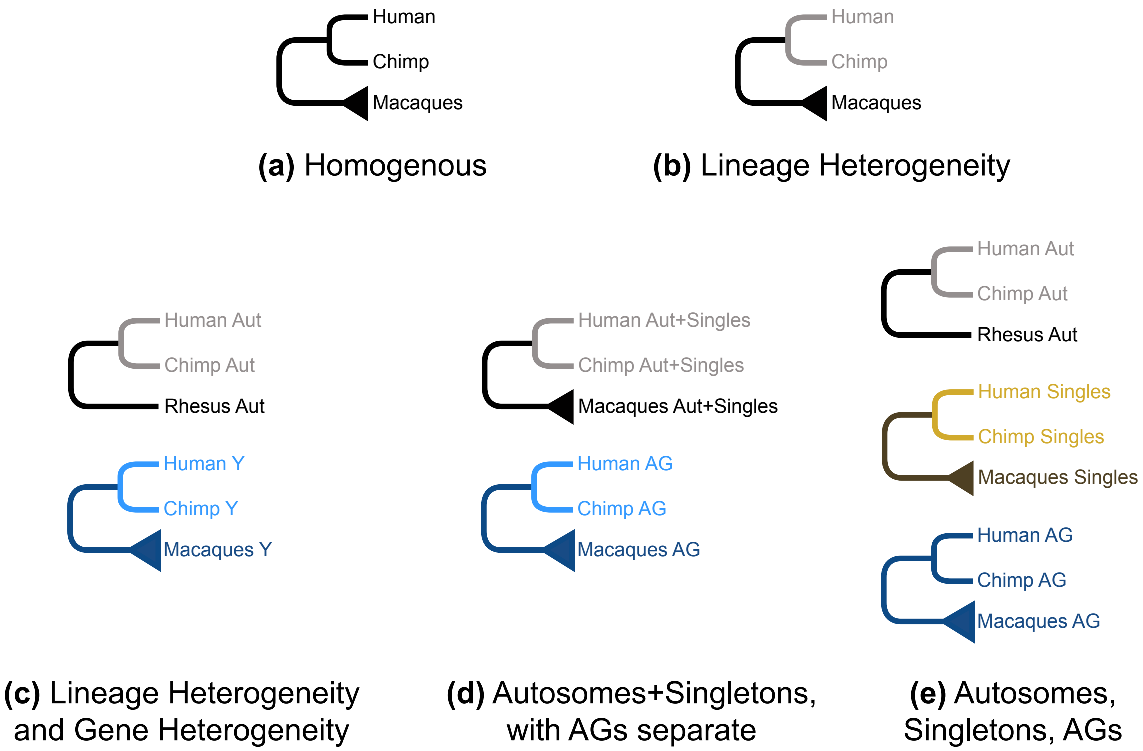


Figure 2.4: **Gene copy numbers for each macaque sample among all seven AG loci assayed by qPCR.** The points show the estimated mean copy number and lines depict standard errors, except for the rhesus macaque (*M. mulatta*) where the copy numbers from the Y-chromosome project [54] have been included as a reference. The black dashed lines represent relevant threshold values used for inferring discrete copy numbers from the continuous qPCR data. The lines and points are coloured by species, as indicated in the symbol key on the right.



(a) Homogenous

(b) Lineage Heterogeneity

(c) Lineage Heterogeneity and Gene Heterogeneity

(d) Autosomes+Singletons, with AGs separate

(e) Autosomes, Singletons, AGs

Figure 2.5: **Diagrams of examples of models fitted to gene copy number data for human, chimp, and macaques.** **(a)** Homogeneous: autosomal genes, msrY-linked singletons, and AGs of all species evolve at the same rate(s) for all lineages. **(b)** Lineage heterogeneity: Hominini lineage evolves differently from the Old World Primate lineages; autosomal and msrY-linked genes evolve at the same rate(s). **(c)** Gene heterogeneity and lineage heterogeneity: autosomes (abbreviated "Aut") and msrY-linked genes (abbreviated "Y") evolve differently from each other; there is also lineage heterogeneity between Hominini and Old World Primates. **(d)** Similar to (c) but autosomes and singletons (abbreviated "Aut+Singles") evolve differently from AGs. **(e)** Autosomal, singleton, and AGs all evolve separately from each other. There is lineage heterogeneity for the autosomal and singleton genes but lineage homogeneity for the AGs.

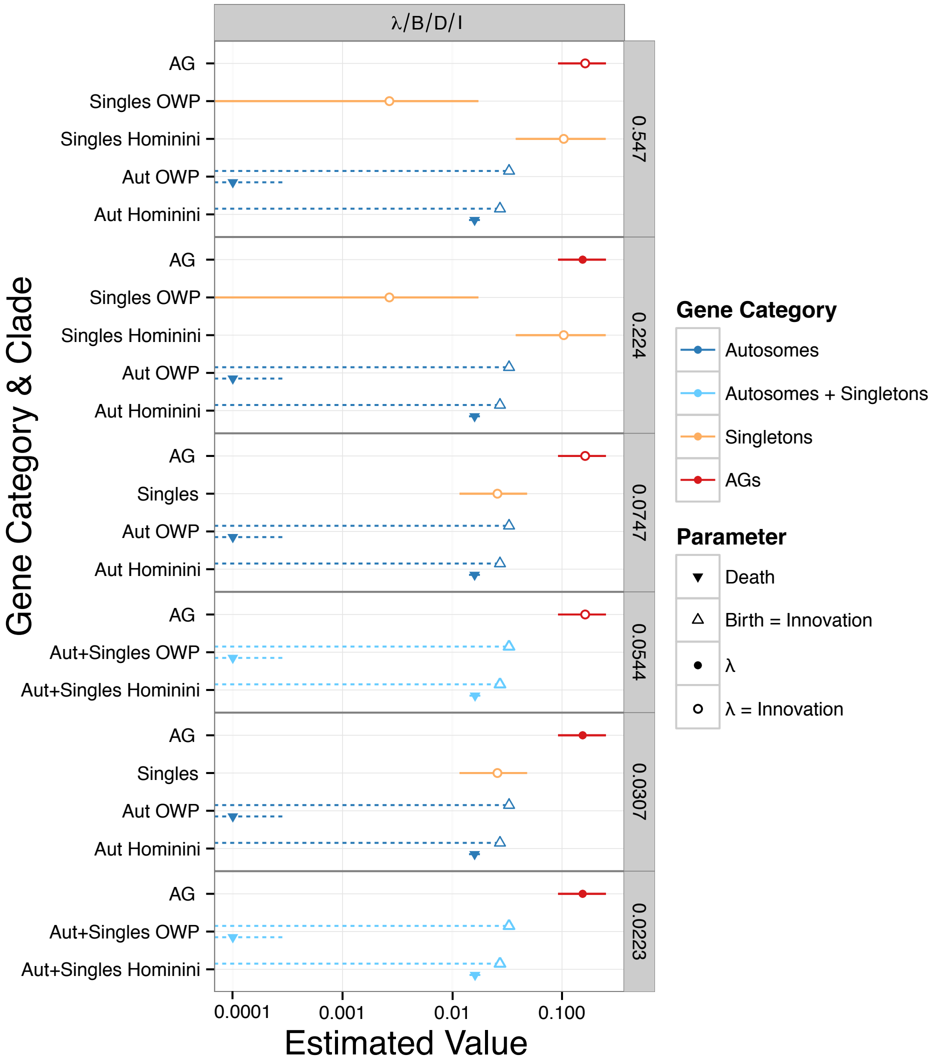


Figure 2.6: **Maximum likelihood estimated (MLE) values for the top 95% cumulative BIC weight models.** Symbols show the MLE and bars indicate univariate 95% confidence intervals. The solid points show the birth=deletion (λ) rate estimates and the outlined points show the birth=deletion=innovation (λ=i) rate estimates. The solid triangles with dashed lines show the birth=innovation (b=i) rate estimates and the outlined triangles with dashed lines show the deletion (d) rate estimates. The BIC weight of each model is indicated on the far right. On the y-axis, the autosomal gene category is abbreviated as “aut,” the msrY-linked singleton category is abbreviated “singles,” and the gene category where autosomal genes and singletons evolve at the same rates is abbreviated “aut+singles.” The colours of the symbols highlight gene heterogeneity in all of the models and correspond to different gene categories as indicated in the symbol key on the right. Models with lineage heterogeneity are indicated on the y-axis when a gene category has the lineage heterogeneity between Hominini and other Old World Primates (“OWP”).

2.8 Availability of supporting data

The sequence data supporting the results of this article are available in the GenBank repository [sequence accessions TBD]. The qPCR data sets supporting the results of this article are available in the RDMLdb repository [data set accessions TBD].

2.9 List of Abbreviations

AG: ampliconic gene; b: duplication (or 'birth') rate; BAC: bacterial artificial chromosome; BIC: Bayesian Information Criterion; CV: coefficient of variation; d: deletion rate; gDNA: genomic DNA; i: innovation rate; λ: a rate of copy number evolution where birth and deletion are equal; MAP: maximum *a posteriori*; MRCA: most recent common ancestor; msrY: male-specific region of the Y chromosome; My: million years; N_e: effective population size; PCR: polymerase chain reaction; qPCR: quantitative polymerase chain reaction.

2.10 Competing interests

The authors declare that they have no competing interests.

2.11 Authors' contributions

This study was designed by AHG, BMB, and BJE, laboratory work was performed by AHG, statistical analyses were performed by AHG and BMB, and the manuscript was written by AHG and BJE. All authors then provided comments on this draft.

2.12 Acknowledgements

We thank Brian Golding and members of the Evans and Golding labs for feedback on this project. We also thank Michael Wakem from Bio-Rad Laboratories Canada for his help optimizing the qPCR assays.

2.A Supplementary Methods

Sample details, gDNA extraction, & sequencing

Sequencing of ampliconic exons was accomplished prior to the publication of the complete rhesus macaque Y-chromosome; therefore BAC sequences from the rhesus macaque Y-chromosome project were used for primer design (GenBank accessions in Table A1). Rhesus macaque sequences with high similarity to human ampliconic exons were used with Primer3 [189] to design primers for PCR. For all genes, except *XKRY* and *RBMY*, primers were initially designed to co-amplify putative paralogs and used for direct sequencing. In order to resolve the heterozygous sites resulting from co-amplification of paralogs, the PCR products from at least three Indonesian macaque samples were cloned and sequenced, then, when possible, paralog-specific primers were designed.

Rhesus macaque, mandril, and baboon whole blood samples were obtained from the Toronto Zoo using a needle draw. Blood samples were frozen to -20°C almost immediately after being drawn, shipped on ice to Hamilton, Ontario, then stored continuously at -80°C until DNA extraction. All other macaque samples were obtained from pet animals in Indonesia as detailed in [190–192]. Whole blood was drawn with a needle and fixed almost immediately by mixing with equal parts buffer containing SDS, EDTA, and Tris, as detailed in [190]. Fixed blood samples were stored at room temperature for one to eight weeks before being frozen, then stored continuously at -80°C until DNA extraction.

gDNA was extracted from whole blood for all samples using a DNeasy Blood and Tissue Kit (Qiagen, cat.#69504) with the spin protocol. A modification was made to the manufacturer's protocol for the final step: the incubation was done for five minutes with $20 - 80\mu$L of either distilled water or Buffer AE (as listed in Table A2). The nucleic acid concentration and purity of each gDNA extraction (listed in Table A2) was determined using a NanoDrop ND-1000 Spectrophotometer and the program ND1000 v3.8.1 (Thermo-Scientific). No DNase or RNase treatment was performed on the gDNA extracts; nor was a contamination assessment carried out on the gDNA extracts. PCR was performed using FailSafe PCR 2x PreMix D (Epicenter, cat.#FSP995D), and Taq DNA polymerase (Life Technologies); the annealing temperatures are given in Table A3. Primers were tested using genomic DNA from a male rhesus macaque (positive control) and a female *M. ochreata* (negative control) to confirm Y-linkage. PCR reactions were visualized using agarose gel electrophoresis and ethidium bromide. Cloning was done using the TOPO TA kit (Invitrogen, cat.#K4500-40). Sequencing was done using the BigDye Terminator v3.1 Cycle Sequencing kit (Life Technologies, cat.#4337456). Sequence chromatograms were analyzed using Sequencher v4.7 (Gene Codes Corp., Ann Arbour, MI), primer sequences were trimmed in MacClade v4.08 (Sinauer Associates Inc., Sunderland MA, [193]), and FASTA files were aligned using MUSCLE v3.8 [194].

Phylogenetic estimation

The total alignment length for the concatenated single-copy, msrY-linked, protein coding genes was 6 185 bp, and the alignment length after excluding positions with gaps was

6 167 bp. jModelTest v2.1.3 [195] was used to estimate the best nucleotide substitution model among 88 candidate models. The model favoured by BIC was TPM3uf+G. Using this model, BEAST v1.7.5 [196] was used to create a time-calibrated tree under the following assumptions: a strict molecular clock, that Hominins and macaques each form monophyletic clades, and that the times, in years, to the MRCA have priors of $N(\mu = 6.0 \cdot 10^6, \sigma = 3.0 \cdot 10^5)$ and $N(\mu = 3.0 \cdot 10^7, \sigma = 1.5 \cdot 10^6)$ for Hominins and Catarrhines, respectively. These values were chosen for the time to the MRCA since the mean values are commonly accepted [178–180] and a standard deviation of 5% of the mean seems to reasonable given the estimated speciation times from other studies [197]. BEASTMC3 v1.7.5 was used to run three chains at default settings. After inspecting the chain for convergence in Tracer v1.5 [198], a burn-in of 1 million generations was applied, yielding a total chain length of 19 million generations.

A similar procedure was used for generating the AG trees. Pseudogenes were identified in the completed human, chimp, and rhesus macaque msrY by using the functional exon sequences as BLAST queries for all three pair-wise combinations. The longest, non-redundant BLAST hits with $> 80\%$ sequence identity were used to generate the multiple-sequence alignment. The same assumptions as above were used for building BEAST trees except that a relaxed log Normal molecular clock was used instead of a strict molecular clock, since the alignments include pseudogenes as well as functional genes. In addition, the previously reported speciation times [179, 181, 182] for papionins were used as calibration points (prior $= N(\mu = 8.5 \cdot 10^6, \sigma = 4.25 \cdot 10^5)$) in gene families where baboon and/or mandrill presumed functional ortholog sequences were available. Finally, no assumptions (e.g. of monophyly) were made about the tree topology.

qPCR

The annotation from [54] was used to identify functional paralogs for qPCR assay in macaques. qPCR primers were designed according to the recommendations of [184] using Primer3Plus [189] and the AG sequence alignments obtained from macaques. To ensure a consistent reaction efficiency across samples, primers were placed at sites with 100% identical sequence for all paralogs targeted and in all species assayed (Figures A8-A16). To ensure primer specificity to the targeted paralog(s), primers were designed in regions that had distinguishing sequences in all macaque species as compared to other functional or pseudogenized paralogs; in particular, each pair of primers was required to have at least two distinguishing substitutions within the first five nucleotides on the $3'$ end (Figures A8-A16). Any potential secondary structure was excluded using the mfold web server [199] and primer specificity was screened in silico using Primer3Plus [189] and Primer-BLAST [200]. Amplicon lengths and locations on the rhesus macaque Y-chromosome (GenBank accession PRJNA253406) are given in Table A5. Primers were ordered from Sigma-Aldrich with reverse-phase cartridge purification and without any modifications to the sequences listed in Table A3.

The rhesus macaque blood sample obtained from a single male individual at the Toronto Zoo was defined as the control group. We assumed that this individual has the same copy numbers for all AGs and *SRY* as the individual sequenced by [54]. The 13 other macaque samples were defined as the experimental groups because their AG copy numbers are unknown. gDNA extraction methods and purity are described above.

Primer specificity in males was confirmed as detailed above. Specificity was validated using PCR amplification in rhesus macaque followed by either direct sequencing or digestion with a restriction enzyme specific to the mis-primed product (see Table A5). Restriction enzyme digested PCR products were visualized using 8% acrylamide gel electrophoresis and ethidium bromide. Finally, to ensure that primer specificity is consistent across all samples, a melt curve analysis was performed for all reactions to exclude any non-specific reactions.

Primer annealing temperatures were optimized using an eight-point temperature gradient from $55 - 65$°C. Primer efficiency and linear dynamic range was determined using an eight-point dilution gradient of rhesus macaque gDNA (Figures A17-A18 and Table A7; slope, y-intercept, and r^2 of calibration curves are given in Figures A17-A18). The rhesus macaque gDNA concentration and quantification cycle (C_q) variation at the limit of detection (LOD) for each assay is given in Table A7.

All qPCR reactions were set-up manually and run using a CFX96 Touch Real-Time PCR Detection System (Bio-Rad Laboratories, cat.#185-5196), 96-well PCR plates (Bio-Rad Laboratories, cat.#MLL-9651), and Microseal 'B' Adhesive Seals (Bio-Rad Laboratories, cat.#MSB-1001). To minimize the effects of sample-specific inhibition and differential amplification efficiencies between samples, the gDNA concentrations of the experimental samples were normalized relative to the gDNA concentration of the reference sample as recommended by [201] (Table A2). Preliminary assays were done using 8.077μL iTaq Fast SYBR Green Supermix with ROX (Bio-Rad Laboratories, cat.#172-5100), 1μL of gDNA, 0.323μL of 10μM each forward and reverse primers, and UltraPure DNase/RNase-free distilled water (Life Technologies, cat.#10977-015) to a total reaction volume of 15μL. Cycling conditions were: enzyme activation at 95°C for 60s, then 40 cycles of 95°C for 5s, primer specific annealing temperature for 30s, and 70°C for 30s with plate read, and finally a melt curve from $65 - 95$°C in 0.5°C increments of 5s. After this reagent was discontinued, reference genes were repeated and further assays were done using 5.00μL SsoFast EvaGreen Supermix (Bio-Rad Laboratories, cat.#172-5202), 1μL of gDNA, for some assays 1μL of $10mg/\mu$L bovine serum albumin (BSA) fraction V (Gibco, cat.#11018-017), assay-specific volumes of 10μM primers, and UltraPure DNase/RNase-free distilled water to a total reaction volume of 10μL. Cycling conditions were: enzyme activation at $98°$C for $120s$, then 40 cycles of $98°$C for 5s and primer specific annealing temperature for 5s with plate read, and finally a melt curve as above. Assay-specific qPCR reaction details (i.e. annealing temperatures, primer concentrations, and whether BSA was added) are listed in Table A6.

C_q values were determined using the baseline-subtracted regression from the CFX Manager Software v3.0 [202]. Average amplification efficiency per sample was estimated using LinRegPCR v2013.1 [203, 204]. Reactions with baseline errors, as identified by CFX Manager, or efficiencies greater than/less than 8% from the median assay efficiency of the plate, as identified by LinRegPCR, were considered to be kinetic outliers and excluded from further analysis. Inter-run calibration was performed using CFX Manager and the average assay efficiencies estimated from LinRegPCR (Table A7) were used to calculate the relative quantities. The results for the no template control (NTC) reactions is shown in Table A8.

SRY was chosen as a reference gene because it is msrY-linked, rarely lost in mammals (but see [205]), and usually single-copy in mammals (but see [99, 152, 206]). *SRY* was used as a reference to confirm the invariant, single-copy status of *TSPY1* and *XKRY*. We found that using these three reference genes yielded more consistent results for all assays. Finally, the target stability values for these three genes (Table A9) are below the acceptable values for stably expressed reference genes (M-value < 0.5 and the CV of normalized reference gene relative quantities < 0.5 [1]).

The normalized relative quantities (NRQs) were calculated in CFX Manager. Technical replicates were performed at the qPCR stage. After exclusion of kinetic outliers, the rhesus macaque reference sample had on average 22.1 (min= 11, max= 34) technical replicates and the experimental samples had on average 16.4 (min= 4, max= 36) technical replicates among all assays. The intraassay repeatability was quite good, as shown in Table A10.

Copy number estimation from qPCR and sequence data

Unfortunately, previous studies have found that qPCR gene copy number data from gDNA does not cluster cleanly around discrete gene copy numbers [207–211]. For this reason, we chose to use a method that incorporates the estimated uncertainty from the qPCR assay to convert the continuous copy number data into discrete copy number data.

Gene family evolution

Two models, L and BD, are similar to those implemented in CAFE [79, 86, 159]. To circumvent the previously documented [87] numerical instability of the probability calculation from [84], we used matrix exponentiation for all of the models, as implemented by the `expm` v0.99-1.1 `R` package [212], which we found to be numerically stable at copy numbers exceeding 100. Because the Markov process does not have a biologically sensible stationary distribution for models without innovation, we assumed for all models that the ancestral state at the MRCA is Poisson distributed and estimated the Poisson characteristic λ value from the observed copy number data at the tips across all gene families. Note that this prior distribution on the ancestral state at the root is different from that used by CAFE, but similar to that used by BadiRate [82].

For a proposed parameter value(s), the likelihood was calculated using the pruning algorithm [213] and summing over all the possible reconstructions at the root as weighted by the prior probability of each ancestral state at the MRCA [214]. The MLE and univariate (profile likelihood) confidence intervals for each model were found using optimization methods in `R` as implemented by `mle2` from the package `bbmle` v1.0.17 [215]. A bounded method, L-BFGS-B, was used for the optimization because parameter values < 0 are nonsensical for all of the models.

Analysis of whole genome and msrY data

We fitted each of the ten evolutionary models to each of the following gene categories: AGs; msrY-linked singletons; autosomes; AGs and singletons; AGs and autosomes; singletons and autosomes; and AGs, singletons, and autosomes. We calculated the maximum likelihood for all possible combinations of the gene categories that encompassed the entire data by summing the log likelihood of the maximum likelihood estimates for both or all three of these individual model fits.

For each full model, the BIC was estimated using the sum of the number of qPCR observations and the number of copy number estimates from both [54] and [85] as the sample size ($n =$10 040). The BIC weights and normalized probabilities were calculated using the formulas from [216].

2.B Supplementary Tables

Table A1: **List of rhesus macaque BACs used for primer design.**

GenBank accessions
AC206800, AC207040, AC207520, AC208129, AC208130, AC208132, AC208133, AC208822, AC209262, AC209263, AC209264, AC212487, AC212790, AC214069, AC215549, AC215550, AC215640, AC216894, AC217105, AC217129, AC217130, AC217138, AC219066, AC225627, AC225636, AC225837, AC231654, AC231831, AC232761, AC234329, AC234330, AC237223

Table A2: **Quality of the gDNA extracts for each sample.**

Sample	**Species**	**Set(s)***	**Eluted in**	**Nucleic Acid Conc.** (ng/μL)	**Dilution**	**Purity** (A_{260}/A_{280})
Zoo	*M. mulatta*	1^{st}	dH$_2$O	20.0	1	1.74
Zoo	*M. mulatta*	2^{nd}	dH$_2$O	20.0	0.125	1.74
143	*M. arctoides*	2^{nd}	dH$_2$O	251.2	0.067	1.75
P001	*M. maura*	2^{nd}	dH$_2$O	184.3	0.067	1.84
PM616	*M. maura*	2^{nd}	dH$_2$O	16.6	0.5	1.73
PM545	*M. tonkeana* east	1^{st}	dH$_2$O	8.8	1	1.87
PM545	*M. tonkeana* east	2^{nd}	Buf. AE	13.4	0.067	1.97
PM561	*M. tonkeana* west	1^{st}	dH$_2$O	10.0	1	1.73
PM561	*M. tonkeana* west	2^{nd}	Buf. AE	17.4	0.067	1.90
PM582	*M. tonkeana* west	1^{st}	dH$_2$O	1.6	1	1.32
PM582	*M. tonkeana* west	2^{nd}	Buf. AE	7.6	0.5	1.60
PM604	*M. tonkeana* west	1^{st}	dH$_2$O	4.8	1	1.01
PM604	*M. tonkeana* west	2^{nd}	dH$_2$O	18.8	0.25	1.95
PM638	*M. hecki*	1^{st}	dH$_2$O	5.1	1	1.83
PM638	*M. hecki*	2^{nd}	Buf. AE	7.8	0.5	1.78
PM1014	*M. hecki*	$1^{st} + 2^{nd}$	dH$_2$O	4.6	1	2.03
PM655	*M. nigrescens*	1^{st}	dH$_2$O	8.3	1	1.38
PM655	*M. nigrescens*	2^{nd}	dH$_2$O	13.8	0.5	1.81
PM661	*M. nigra*	$1^{st} + 2^{nd}$	dH$_2$O	1.4	1	1.76
PM665	*M. nemestrina*	1^{st}	dH$_2$O	7.8	1	1.80
PM665	*M. nemestrina*	2^{nd}	dH$_2$O	9.4	0.5	1.91
PM704	*M. ochreata*	1^{st}	dH$_2$O	2.9	1	1.73
PM704	*M. ochreata*	2^{nd}	dH$_2$O	11.5	0.25	1.97

* 1^{st} refers to the set of preliminary assays performed using the iTaq Fast SYBR Green Supermix with ROX; 2^{nd} refers to the set of assays performed using the SsoFast EvaGreen Supermix (see Supplementary Methods); $1^{st} + 2^{nd}$ refers to both sets of assays.

Table A3: **Sequences of sequencing and qPCR primers and their annealing temperatures for PCR.**

Primer Name*	**Gene(s)†**	**Sequence**	**Annealing Temp.** (°C)
CDY_For340	*CDY* 1-2	GCCAGCAAGAACGTTAGGAG	58
CDY_Rev892	*CDY* 1-2	TCTGGGTGAATCCATCCTCT	58
CDY_For1005	*CDY* 1-2	CAGTGCAGCTGGAAGTGTGT	58
CDY_Rev1308	*CDY* 1-2	CCTTTTCTGTTGCCCACACT	58
CDY_Rev1536	*CDY* 1-2	TCTCTCATTGGCCTTTTCCA	58
CDYps_For1	*CDY* ψ	CCAGTCAGGGATGCTTTCTC	58
CDYps_Rev1	*CDY* ψ	GGCCCTTTCCAACTCAATCT	58
qCDY_For1	*CDY*	GCGGTCTTGATTTTGGGTAT	58
qCDY_Rev1	*CDY*	ACTGATACAACAATAGGCTTTTAAACT	58
DAZex4-6_For1	*DAZ2*	CCTTTATAGCTATGGATTTGTTTCA	57
DAZex4-6_Rev1	*DAZ2*	GCAGTTCTCACCTGAACGTACT	57
DAZex4-6_For2	*DAZ1*	CTGGACTATGTGCTGTATGATGG	58
DAZex4-6_Rev2	*DAZ1*	TTACAGGATTCAGCGTTATTGG	58
DAZex4-6_For3	*DAZ1*	CTTCACCTTTTCTCTGCCTTT	57
DAZex4-6_Rev3	*DAZ1*	ACAGGATTCAGCGTTATTGG	57
qDAZ_A_Rev2	*DAZ*	TCCAGACATTCTGAAACTGC	58
qDAZ_A_For1	*DAZ2*	TCAGTCACAGATCCATATCCA	59
qDAZ_A_Rev1	*DAZ2*	CACGTGTCAAAAAGAACAATG	59
qDAZ_BCD_For2	*DAZ1*	CTGCAATCAGGAAACAAAAA	58
qDAZ_BCD_Rev2	*DAZ1*	CGAGCACCTTATAAAAAGCA	58
HSFYa_For1	*HSFY2-3*	CTGGAACAGCGGCTAAAGA	57
HSFYa_Rev1	*HSFY2-3*	CTTGTTGGAACAGCAGGTGA	57

Continued on next page...

Table A3 – Continued from previous page

Primer **Name***	**Gene(s)**†	**Sequence**	**Annealing** **Temp.** (°C)
HSFYb_For1	*HSFY1*	GCCTGGAAGAGTAGCTCAGG	57
HSFYb_For2	*HSFY1*	CATGCAGCCTGGAAGAGTAG	57
HSFYb_Rev1	*HSFY1*	GGGCCAGATGAATTAGCAGT	57
HSFYintron_For	*HSFY1-3*	GGGATGAGAATGGAACTTGC	57
HSFYintron_Rev	*HSFY1-3*	CCATGTTAGCCCCTGCTCTA	57
HSFYex1_For	*HSFY1-3*	AAGCYTCCAMTAGGTCTCCA	55
HSFYex1_Rev	*HSFY1-3*	GAAAGGTGGSTASAAAGGCAGA	55
HSFYex1_For2	*HSFY2-3*	AGGTCTCCATTGTGTGAGCA	58
HSFYex1_Rev2	*HSFY2-3*	CAGCCAGAAAGGTGGGTAGA	58
HSFYex1_For3	*HSFY1*	GCTTCCAATAGGTCTCCATTGT	58
HSFYex1_Rev3	*HSFY1*	CTGCCAGAAAGGTGGCTACA	58
HSFYex1ps1For	*HSFY* ψ	TCCCCTAGGTCTCCATTGTG	55
HSFYex1ps1Rev	*HSFY* ψ	GTTGGCCAAAGAAGCAGAAG	55
HSFYex1ps2For	*HSFY* ψ	TCTCCATTGCGTGAACACAT	55
HSFYex1ps2Rev	*HSFY* ψ	GCCAGAAAGTTGACTAGAAAAGC	55
HSFYex2For	*HSFY1-3*	TGGCTGTCCCCAACTTTTAG	58
HSFYex2Rev	*HSFY1-3*	GTTGARKAGCTGGCCTGGAA	58
HSFYex2_Rev2	*HSFY2-3*	TCCAGTGGTGATGGTTGAGT	58
HSFYex2_Rev3	*HSFY1*	TGTCCAGTAGTGATGGTTGAAGA	58
HSFYex2psFor	*HSFY* ψ	TTCAAATGTGGCTGTCTCCA	58
HSFYex2psRev	*HSFY* ψ	TGGTTGAAGAGTTGGCCTGT	58
qHSFYex2_A_For1	*HSFY2-3*	AGAATCACCTGCTGTTCCAA	58
qHSFYex2_A_Rev1	*HSFY2-3*	CCTGGAAAGAGGGCTAGATG	58

Continued on next page...

Table A3 – Continued from previous page

Primer Name*	Gene(s)[†]	Sequence	Annealing Temp. (°C)
qHSFYex2_B_For1	*HSFY1*	TGGCCCAATTAGAAGTGGTT	58
qHSFYex2_B_Rev1	*HSFY1*	CCCCTTGCTTGCATATAGGT	58
PRYex3F	*PRY*ψ	TGGGAAGGTTGGCTCTATTT	55
PRYex3R	*PRY*ψ	CACGAGGTACCCTGAAAACA	55
PRYex3_For2	*PRY*ψ	TTCAAGGTATGGGAAGGTTGA	57
PRYex3_Rev2	*PRY*ψ	GGTGTCCCAAAGCCACAG	57
PRYex3_For3	*PRY*ψ	AGGTGTCCCAAAGCTGTGAT	57
PRYex3_Rev3	*PRY*ψ	TCCAGTGGTGATGGTTGAGT	57
RBMYex1-2_For2	*RBMY*	GCAGCACAATGGTAGAAGCA	58
RBMYex1-2_Rev3	*RBMY*	GGCATTTTACTATCTTCAAAAGTTACA	58
RBMYex1-2_For1	*RBMY* ψ	AGCTTTTCATTGGTGGGCTA	58
RBMYex1-2_Rev1	*RBMY* ψ	ATCTGCAGGGTTCTCAAACG	58
RBMYex1-2_For4	*RBMY* ψ	TGGCAAGCTTTTCATTAGCA	58
RBMYex1-2_Rev4	*RBMY* ψ	GCTTTGGCAGCATTCTTAGC	58
RBMYex3_For3	*RBMY*	GAACAAGCCAAGAAACCATCA	58
RBMYex3_Rev3	*RBMY*	AAACATTACCCAAGTGTCCTTCA	58
RBMYex3_For1	*RBMY* ψ	AACAAGCCAACAAACCATCC	58
RBMYex3_Rev1	*RBMY* ψ	CATTACCCAAGTGTCCTCCA	58
RBMYex3_For2	*RBMY* ψ	AGTCTTTGGATGGAAAAGCAA	58
RBMYex3_Rev2	*RBMY* ψ	GTCCTTCATGTGAGGGATGC	58
RBMYex3_For4	*RBMY* ψ	ATGCCAATGTTATTGATAGTATGC	58
RBMYex3_Rev4	*RBMY* ψ	TTCATGTGAGGGAAGCCATC	58
RBMYex3_For5	*RBMY* ψ	TGCTTAGGTTAAATATGCCAATG	58
RBMYex3_Rev5	*RBMY* ψ	GTGAGGGAAGCCATCCTCTT	58

Continued on next page...

Table A3 – Continued from previous page

Primer	**Gene(s)**[†]	**Sequence**	**Annealing**
Name[*]			**Temp.** (°C)
RBMYex3_For6	*RBMY* ψ	CCACAGGTGACATAAATCTGCT	58
RBMYex3_Rev6	*RBMY* ψ	ATCATGTGAGGGAAGCCACT	58
qRBMYex3_For1	*RBMY*	CATTTCAAAGTGGTGGTAGGC	58
qRBMYex3_Rev3	*RBMY*	CCTCCACTACTTCCTTTTGCAG	58
qSRY_For1	*SRY*	CTCAAGAATGCAGCACCAGT	58
qSRY_Rev1	*SRY*	GCTTTGTCGAGTGGCTGTAG	58
TSPYex3-5For	*TSPY1-5*	GTGAAAGAAGCGAAGCATCC	58
TSPYex3-5Rev	*TSPY1-5*	CTCTTCAGGYGGCTTCATC	58
TSPYex3-5_Rev_a1	*TSPY2-5*	TATCCCGGGTATCAGACAGC	58
TSPYex3-5_Rev_b1	*TSPY1*	CCTCATGTAGCATTGCATGG	58
TSPYex3-5_For_b2	*TSPY1*	CACCTCAGCCAAAAAGGTGT	58
TSPYex3-5_Rev_b2	*TSPY1*	TCAGGGGAATCAATCGAGAG	58
qTSPY_B_For3	*TSPY1*	TCCAATTCAGTGGTGTCAGG	58
qTSPY_B_Rev3	*TSPY1*	CCACCTCAGCAATCCTATTACC	58
XKRY_For28	*XKRY*	CCTGATGACATGTTCCCTGTT	55
XKRY_Rev354	*XKRY*	AGCATCAGTACTGTACCCACCA	55
XKRY_For24	*XKRY* ψ	CATTGCTGATGACATGTTCTCTC	55
XKRY_Rev339	*XKRY* ψ	CCCAACATGCTGGAATTATTTT	55
qXKRY_For1	*XKRY*	ATATGGCGTTTTCTGGAGGT	58
qXKRY_Rev1	*XKRY*	ATGGTGCCAACAATGGTACA	58

[*] Primers whose names begin with “q” were used for qPCR; all other primers were used only for PCR amplification and sequencing.

[†] ψ indicates that a msrY-linked, pseudogenized version of the gene is targeted by this primer.

Table A4: **Accession numbers for sequences used to build gene trees but not sequenced in this study.**

Gene	Species	Genbank accession number
AMELY	*H. sapiens*	NC_000024.10: 6 868 563-6 867 999
AMELY	macaques	HM071684-HM071693
AMELY	*P. troglodytes*	AB091782.1
CDY	*C. jacchus*	FJ526998.1
DAZ1	*H. sapiens*	NC_000003.12
DAZ1	*M. mulatta*	AF053608.1
DAZ1	*P. troglodytes*	AF053606.1
DBY	*C. jacchus*	AC245293.3: 42 530-42 764, 60 146-60 340
DBY	*H. sapiens*	NC_000024.10: 12916651-12916854
DBY	macaques	HM071810-HM071817
DBY	*P. troglodytes*	AC146254.2: 47 245-47 042
PRKY	*H. sapiens*	NC_000024.10: 7 325 834-7 326 451
PRKY	macaques	HM071789-HM071798
PRKY	*P. troglodytes*	NC_006492.3: 25 186 998-25 186 381
RBMY	*C. jacchus*	AC220987.3
SMCY	*C. jacchus*	AC226175.3: 2 297-3 036
SMCY	*H. sapiens*	NC_000024.10: 19 707 809-19 707 071
SMCY	macaques	HM071647-HM071656
SMCY	*P. troglodytes*	AC144429.1: 66 765-66 039
SRY	*C. jacchus*	AC221052.5: 165 915-166 709

Continued on next page...

Table A4 – Continued from previous page

Gene	Species	Genbank accession number
SRY	*H. sapiens*	NM_003140.2
SRY	macaques	HM071767-HM071776
SRY	*P. troglodytes*	JF293177.1
TBL1Y	*H. sapiens*	NC_000024.10: 7 070 164-7 070 931
TBL1Y	macaques	HM071704-HM071713
TBL1Y	*P. troglodytes*	AC146484.3: 22 321-21 554
TSPY	*C. jacchus*	AC234250.2
TSPY	*M. nigrescens*	AF284268.2
TSPY	*M. nigra*	AF284267.2
TSPY	*M. hecki*	AF284256.2
TSPY	*M. tonkeana* east	AF284236.2
TSPY	*M. tonkeana* west	AF284235.2
TSPY	*M. ochreata*	AF284269.2
TSPY	*M. maura*	AF284257.2
USP9Y	*C. jacchus*	AC234102.2: 7 5193-7 5979
USP9Y	*H. sapiens*	NM_004654.3
USP9Y	macaques	HM071667-HM071676
USP9Y	*P. troglodytes*	NM_001009110.1
UTY	*C. jacchus*	AC231992.2: 2 967-3 707
UTY	*H. sapiens*	XM_011531442.1
UTY	macaques	HM071726-HM071733
UTY	*P. troglodytes*	AC146194.2: 47 615-46 872

Continued on next page...

Table A4 – Continued from previous page

Gene	Species	Genbank accession number
XKR3	*H. sapiens*	NM_175878.3
ZFY	*C. jacchus*	AC221058.4: 151 603-152 290
ZFY	*H. sapiens*	NM_003411.3
ZFY	macaques	HM071746-HM071755
ZFY	*P. troglodytes*	NM_001009003.1

Table A5: **qPCR target information.**

Primer Name	**Length** (nt)	**Primer Location*** (bp)	**Exon/ Intron**	**Specificity Check**
qCDY_For1 qCDY_Rev1	132	8813798-79, 8903851-70 8813693-66, 8903956-83	ex1	direct sequencing
qDAZ_A_For1 qDAZ_A_Rev1	103	10637794-814, 10650686-706 10637877-97, 10650769-89	ex4-ex6	restriction enzyme: CviAII
qDAZ_BCD_For2 qDAZ_BCD_Rev2	117	9166093-74, 9177331-12, 9186075-56 9165995-76, 9177233-14, 9185977-58	ex4-ex6	direct sequencing
qHSFY_A_For1 qHSFY_A_Rev1	135	8386429-48, 8553596-77 8386545-64, 8553480-61	ex2	restriction enzyme: CviAII
qHSFY_B_For1 qHSFY_B_Rev1	146	8006912-893 8006785-66	ex2	restriction enzyme: MboI
qRBMYex3_For1 qRBMYex3_Rev3	93	6278470-50 6278377-98	ex3	direct sequencing
qSRY_For1 qSRY_Rev1	91	81 721-02 81 649-30	ex1	none
qTSPY_B_For3 qTSPY_B_Rev3	124	6308032-13 6307929-08	ex3-ex5	restriction enzyme: Fnu4HI
qXKRY_For1 qXKRY_Rev1	120	7642396-77 7642295-76	ex1	restriction enzyme: DdeI

* On the rhesus macaque Y-chromosome.

Table A6: **Table of qPCR conditions.**

Assay	**Set***	**Annealing Temp.** (°C)	**BSA Added?**	**For. Primer Vol.** (μL)	**Rev. Primer Vol.** (μL)
CDY	2^{nd}	59.0	yes	0.600	1.000
DAZa	1^{st}	61.4	no	0.323	0.323
DAZa	2^{nd}	61.4	no	0.833	0.833
DAZbcd	2^{nd}	59.0	yes	1.000	0.600
HSFYa	1^{st}	63.4	no	0.323	0.323
HSFYa	2^{nd}	63.4	yes	0.833	0.833
HSFYb	1^{st}	63.4	no	0.323	0.323
HSFYb	2^{nd}	63.4	yes	0.500	0.500
RBMY	2^{nd}	56.0	yes	0.300	0.300
SRY	1^{st}	63.8	no	0.323	0.323
SRY	2^{nd}	63.4	yes	0.833	0.833
TSPYb	1^{st}	63.4	no	0.323	0.323
TSPYb	2^{nd}	63.4	yes	0.300	0.300
XKRY	1^{st}	63.4	no	0.323	0.323
XKRY	2^{nd}	63.4	yes	0.300	0.500

* 1^{st} Refers to the set of preliminary assays performed using the iTaq Fast SYBR Green Supermix with ROX; 2^{nd} refers to the set of assays performed using the SsoFast EvaGreen Supermix (see Supplementary Methods).

Table A7: **Table of qPCR validation.**

Assay	**Set***	**Estimated % Eff.**†	**LOD**‡	C_q **St. Dev. at LOD**	**Mean Eff.**[1]	**Among Sample Eff. Variation**[2]
CDY	2^{nd}	109.3	$2.06 \cdot 10^{-3}$	0.554	1.756	0.022
DAZa	1^{st}	93.2	$1.95 \cdot 10^{-3}$	0.213	1.822	0.061
DAZa	2^{nd}	107.7	$1.13 \cdot 10^{-3}$	0.0917	1.683	0.034
DAZbcd	2^{nd}	101.0	$2.06 \cdot 10^{-3}$	0.198	1.634	0.027
HSFYa	1^{st}	90.2	$1.95 \cdot 10^{-3}$	0.119	1.714	0.044
HSFYa	2^{nd}	102.7	$2.06 \cdot 10^{-3}$	0.199	1.791	0.035
HSFYb	1^{st}	103.0	$1.95 \cdot 10^{-3}$	0.670	1.781	0.054
HSFYb	2^{nd}	109.9	$5.50 \cdot 10^{-3}$	0.193	1.778	0.018
RBMY	2^{nd}	106.8	$1.03 \cdot 10^{-3}$	1.45	1.701	0.048
SRY	1^{st}	90.5	$6.91 \cdot 10^{-4}$	NA	1.825	0.070
SRY	2^{nd}	99.4	$2.06 \cdot 10^{-3}$	0.161	1.815	0.030
TSPYb	1^{st}	91.1	$6.91 \cdot 10^{-4}$	1.27	1.761	0.025
TSPYb	2^{nd}	102.9	$1.03 \cdot 10^{-3}$	0.320	1.658	0.037
XKRY	1^{st}	97.7	$6.91 \cdot 10^{-4}$	1.51	1.873	0.020
XKRY	2^{nd}	106.3	$3.92 \cdot 10^{-2}$	0.391	1.666	0.035

* 1^{st} Refers to the set of preliminary assays performed using the iTaq Fast SYBR Green Supermix with ROX; 2^{nd} refers to the set of assays performed using the SsoFast EvaGreen Supermix (see Supplementary Methods).

† Reaction efficiency as estimated from an eight-point dilution gradient using CFX Manager Software.

‡ LOD is expressed as the dilution of the *M. mulatta* gDNA sample listed in Table A2.

[1] Mean reaction efficiency as estimated from averaging the efficiencies determined for each qPCR reaction performed on experimental and reference samples using LinRegPCR.

[2] Standard deviation of the mean LinRegPCR reaction efficiencies of each sample for each assay.

Table A8: **Table of C_q values for NTC reactions for all assays performed.**

Assay	**Set**[*]	C_q **Values**[†]
CDY	2^{nd}	35.45, 36.55, 37.18, 37.93, NA, NA, NA
DAZa	1^{st}	NA, NA, NA
DAZa	2^{nd}	37.21, 37.79, 38.11, Na
DAZbcd	2^{nd}	NA, NA, NA, NA, NA, NA, NA, NA, NA
HSFYa	1^{st}	2.17, NA, NA, NA, NA
HSFYa	2^{nd}	35.82, 36.50, 36.53, 39.59, NA, NA
HSFYb	1^{st}	NA, NA, NA, NA, NA
HSFYb	2^{nd}	NA, NA, NA, NA
RBMY	2^{nd}	36.37, 36.42, 36.43 36.47, 37.82
SRY	1^{st}	NA, NA, NA, NA, NA, NA, NA
SRY	2^{nd}	NA, NA, NA
TSPYb	1^{st}	NA, NA, NA
TSPYb	2^{nd}	NA, NA, NA, NA, NA, NA, NA
XKRY	1^{st}	NA, NA, NA
XKRY	2^{nd}	NA, NA, NA, NA, NA, NA, NA

[*] 1^{st} Refers to the set of preliminary assays performed using the iTaq Fast SYBR Green Supermix with ROX; 2^{nd} refers to the set of assays performed using the SsoFast EvaGreen Supermix (see Supplementary Methods).

[†] "NA" values indicate a C_q that was undetectable.

Table A9: **Table of target stability values for reference genes used in qPCR.**

Assay	**CV**	**M-value**
SRY	0.175	0.358
TSPYb	0.123	0.303
XKRY	0.100	0.285
Mean	0.133	0.315

Table A10: **Table of intraassay variability, shown as standard errors of NRQs, for each sample and the mean across all samples.**

Sample	**CDY**	**DAZa**	**DAZbcd**	**HSFYa**	**HSFYb**	**RBMY**	**XKRY**
PM1014	0.074	0.065	0.077	0.035	0.099	–	0.049
PM638	0.047	0.058	0.055	0.020	0.015	0.051	0.092
PM704	0.036	–	–	0.033	0.031	0.027	0.039
PM655	0.021	0.063	0.039	0.059	0.068	0.041	0.049
PM604	0.063	0.019	0.055	0.031	0.028	–	0.061
PM561	0.066	0.038	0.053	0.034	0.029	–	0.073
PM582	0.046	0.053	0.041	0.039	0.067	–	0.044
PM545	0.074	0.035	0.048	0.055	0.068	0.089	0.052
PM616	0.030	0.040	0.031	0.038	0.056	0.032	0.049
P001	0.080	0.033	0.044	0.068	0.057	0.036	0.051
PM661	0.069	0.037	0.032	0.123	0.058	–	0.058
PM665	0.056	0.049	0.034	0.058	0.047	–	0.054
143	0.274	0.024	0.068	0.067	0.053	0.066	0.085
Zoo	0.218	0.036	0.042	0.028	0.030	0.028	0.035
Mean	0.082	0.042	0.048	0.049	0.050	0.046	0.057

Table A11: **Summary of the model fits to the singleton data set alone.**

Model	logLik	df	Estimate(s)*	95% CI	BIC	w(BIC)
L	-57.1	1	$\lambda = 0.0252$	0.012–0.048	116	0
L+het	-51.8	2	$\lambda_R = 0.00875, \lambda_{HC} = 0.101$	0.002-0.025, 0.038-0.218	108	0
$L = I$	-49.8	1	$\lambda' = 0.0258$	0.01-0.059	102	0.003
$L = I$+het	-43.2	2	$\lambda'_R = 0.00267, \lambda'_{HC} = 0.104$	0-0.017, 0.038-0.25	90.9	0.777
BD	-51.1	2	$b = 0, d = 0.051$	0-0.03, 0.023-0.098	106	0
BD+het	-45.7	4	$b_R = 0, d_R = 0.0166$ $b_{HC} = 0, d_{HC} = 0.193$	0-0.039, 0.003-0.05 0-0.167, 0.075-0.394	100	0.007
$B = ID$	-49.7	2	$b' = 0.0178, d = 0.032$	0-0.074, 0.004-0.082	104	0.001
$B = ID$+het	-42.3	4	$b'_R = 0.00279, d_R = 0$ $b'_{HC} = 0.0618, d_{HC} = 0.173$	0-0.031, 0-0.026 0.007-0.202, 0.046-0.397	93.4	0.211

* The units of estimated parameter values are “per gene copy per million generations” for b, d, and λ, or “per gene family per million generations” for i.

Table A12: **Summary of the model fits to the AG data set alone.**

Model	logLik	df	Estimate(s)*	95% CI	BIC	ΔBIC	w(BIC)
L	-71.4	1	$\lambda = 0.154$	0.092-0.251	145	1.60	0.117
L+het	-70.5	2	$\lambda_R = 0.128, \lambda_{HC} = 0.347$	0.067-0.226, 0.126-1.084	146	2.34	0.081
$L = I$	-70.6	1	$\lambda' = 0.163$	0.094-0.277	144		0.260
$L = I$+het	-69.7	2	$\lambda'_R = 0.134, \lambda'_{HC} = 0.357$	0.066-0.248, 0.102-1.687	144	0.744	0.179
BD	-70.9	2	$b = 0.122, d = 0.185$	0.046-0.233, 0.099-0.3	147	3.14	0.054
BD+het	-68.7	4	$b_R = 0.0161, d_R = 0.161$ $b_{HC} = 0.714, d_{HC} = 0.355$	0-0.158, 0.086-0.267 0.127-1.476, 0.089-1.074	148	3.83	0.038
$B = ID$	-70.3	2	$b' = 0.14, d = 0.187$	0.054-0.267, 0.089-0.319	146	1.94	0.098
$B = ID$+het	-67.2	4	$b'_R = 0.0184, d_R = 0.176$ $b'_{HC} = 0.762, d_{HC} = 0.24$	0-0.147, 0.089-0.306 0.238-2.393, 0-1.981	145	0.832	0.172

* The units of estimated parameter values are “per gene copy per million generations” for b, d, and λ, or “per gene family per million generations” for i.

2.C Supplementary Figures

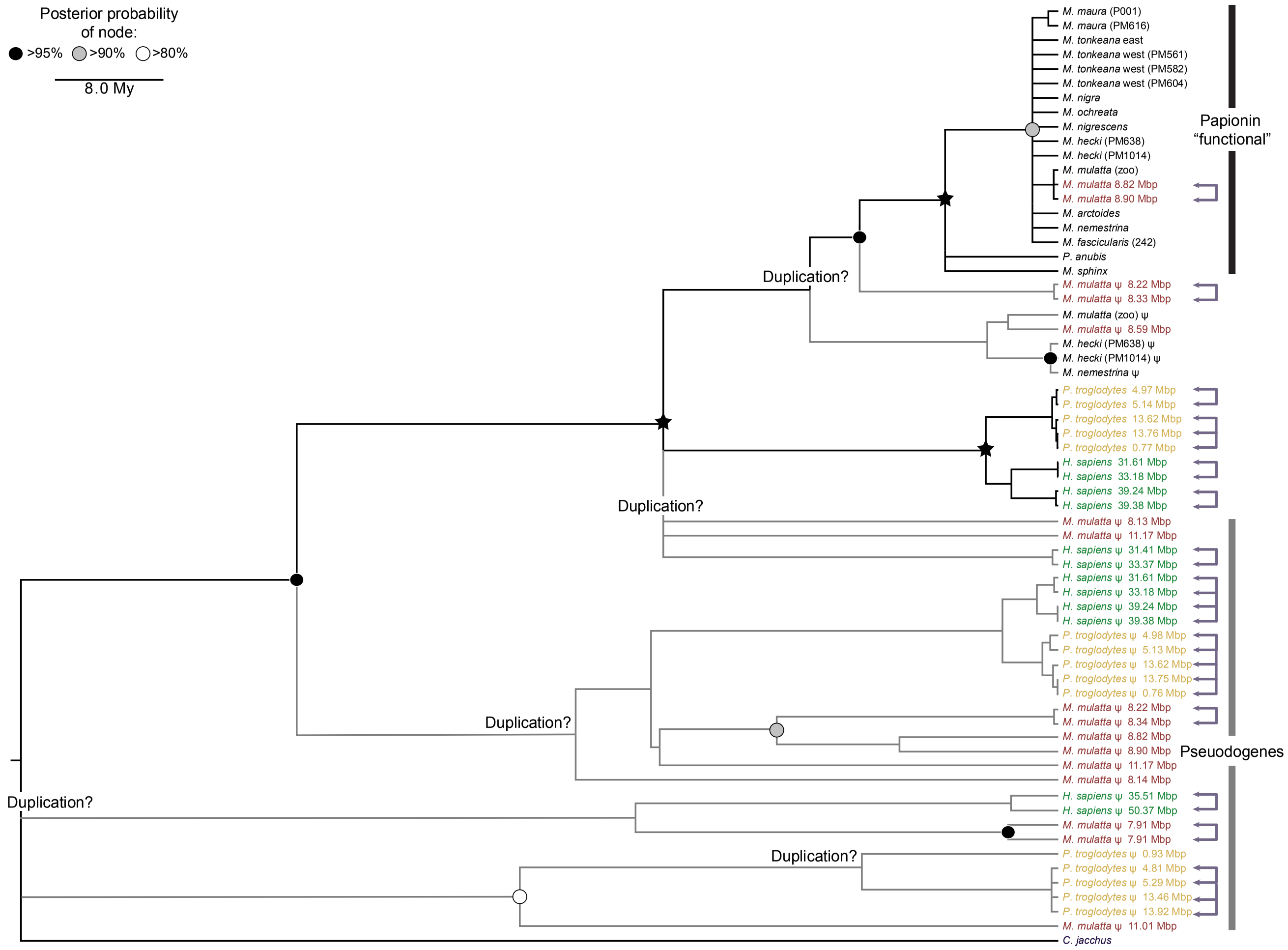


Figure A1: **Maximum clade credibility tree of *CDY*.** Known functional genes (in rhesus, human, and chimp) and presumed functional genes (in marmoset and papionin species) are connected by thick branches. The approximate position, in megabases, on the completed Y-chromosome is given for each rhesus macaque, human, and chimp sequence. All unlabeled nodes have $> 99\%$ posterior probability. Sample ID's can be found in Table A2

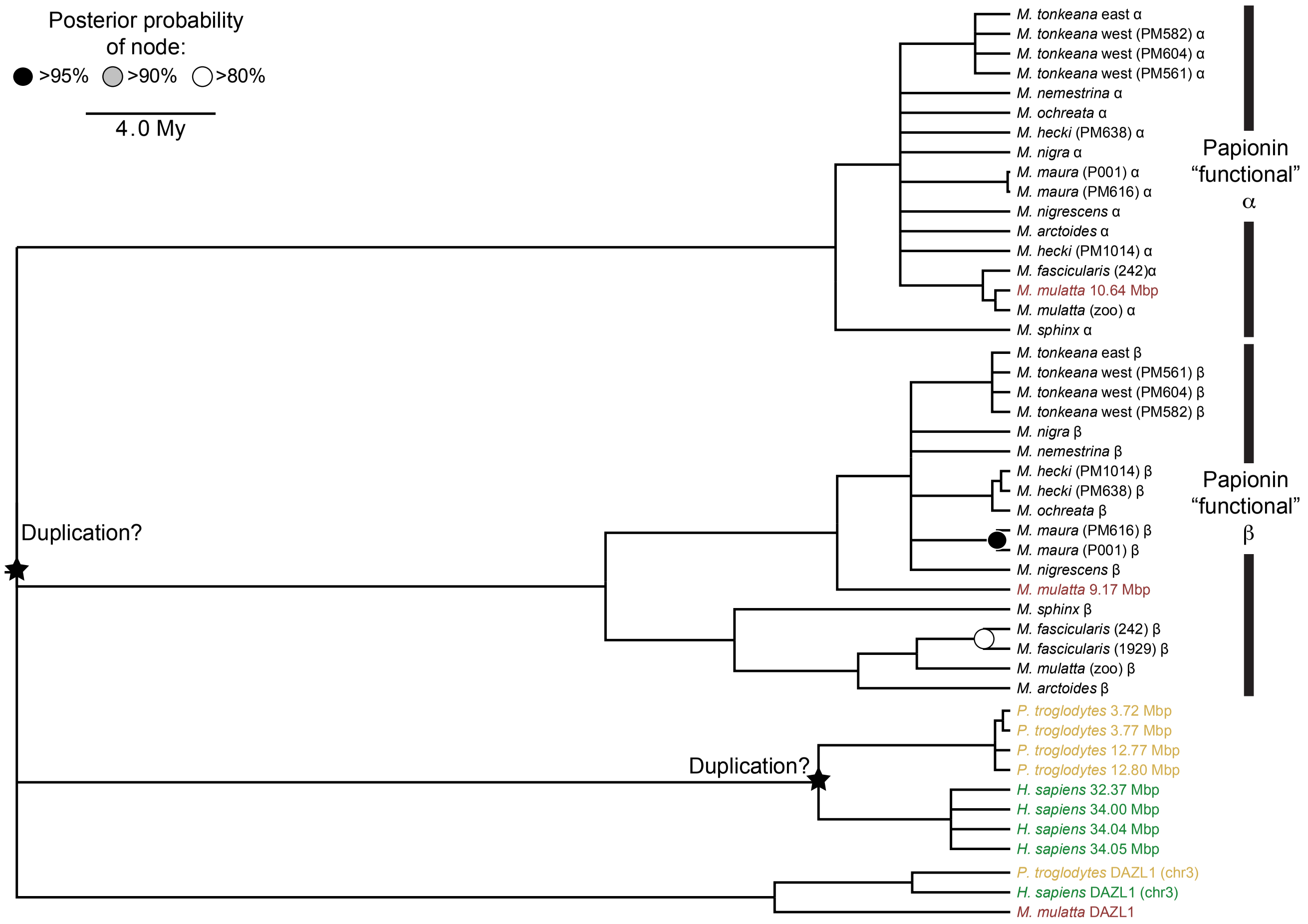


Figure A2: **Maximum clade credibility tree of *DAZ*.** Three orthologs of the autosomal paralog *DAZL1*, which was transposed to the msrY in the Old World primate ancestor [146], are included as an outgroup. Papionin orthologs of rhesus *DAZ1* are labeled "α" and orthologs of *DAZ2* are labeled "β," corresponding to the primer set ("a" or "bcd," respectively, in Table A3) that successfully amplified the sequence. Other features of the tree are drawn as detailed in Figure A1.

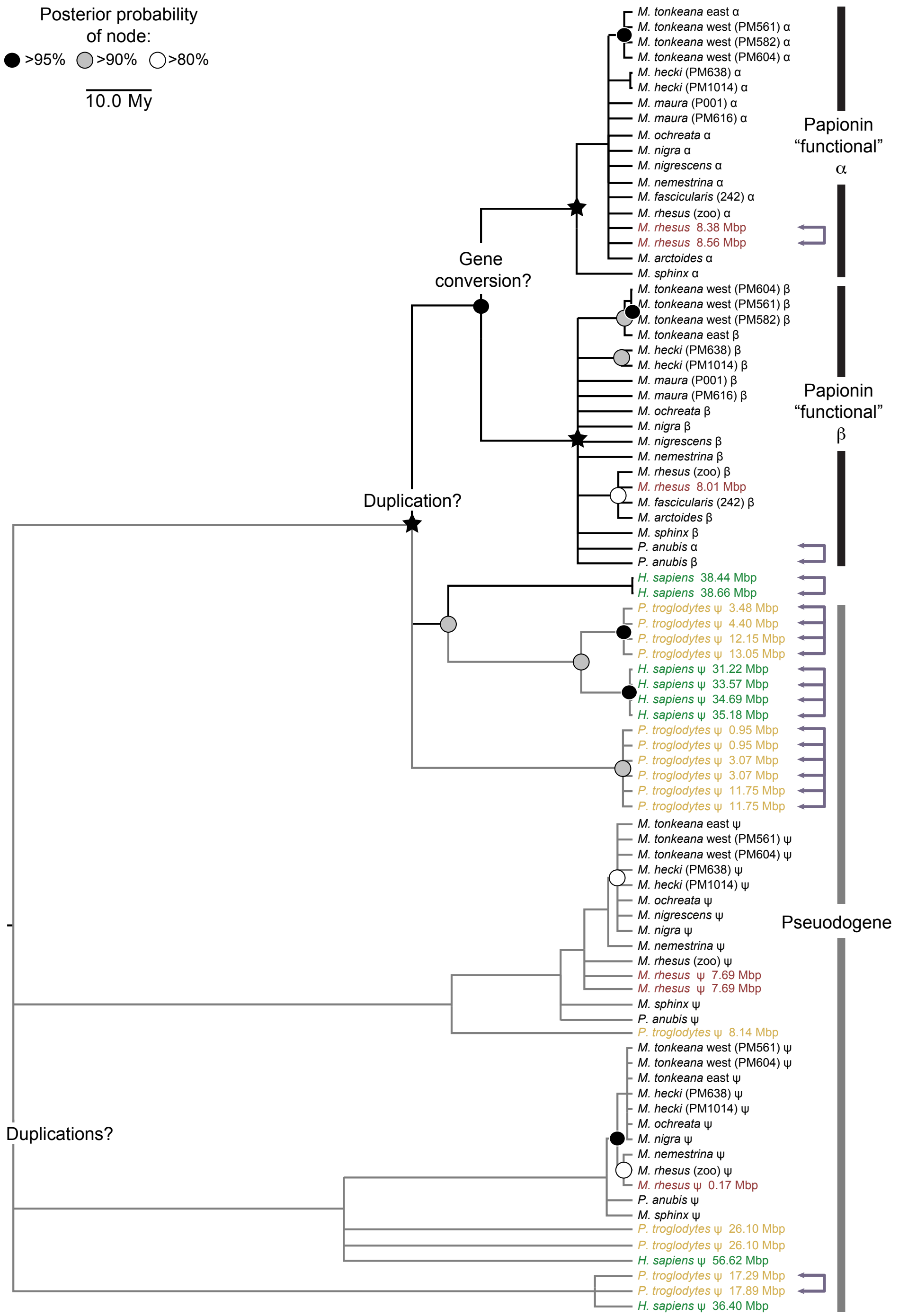


Figure A3: **Maximum clade credibility tree of *HSFY*.** Papionin orthologs of rhesus *HSFY2-3* are labeled "α" and orthologs of *HSFY1* are labeled "β," corresponding to the primer set ("a" or "b," respectively, in Table A3) that was used for sequencing. Other features of the tree are drawn as detailed in Figure A1.

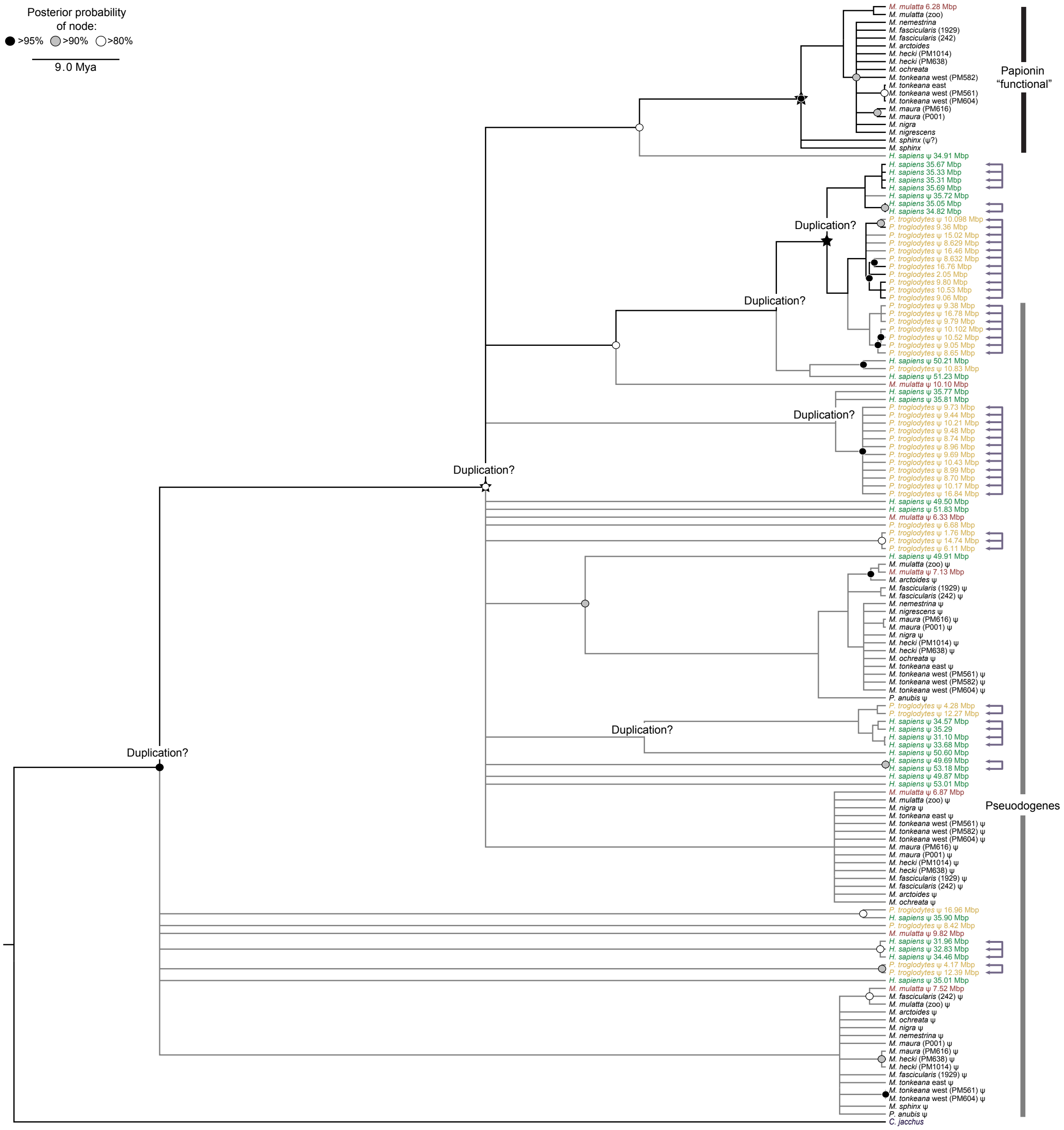


Figure A4: **Maximum clade credibility tree of *RBMY*.** Features of the tree are drawn as detailed in Figure A1.

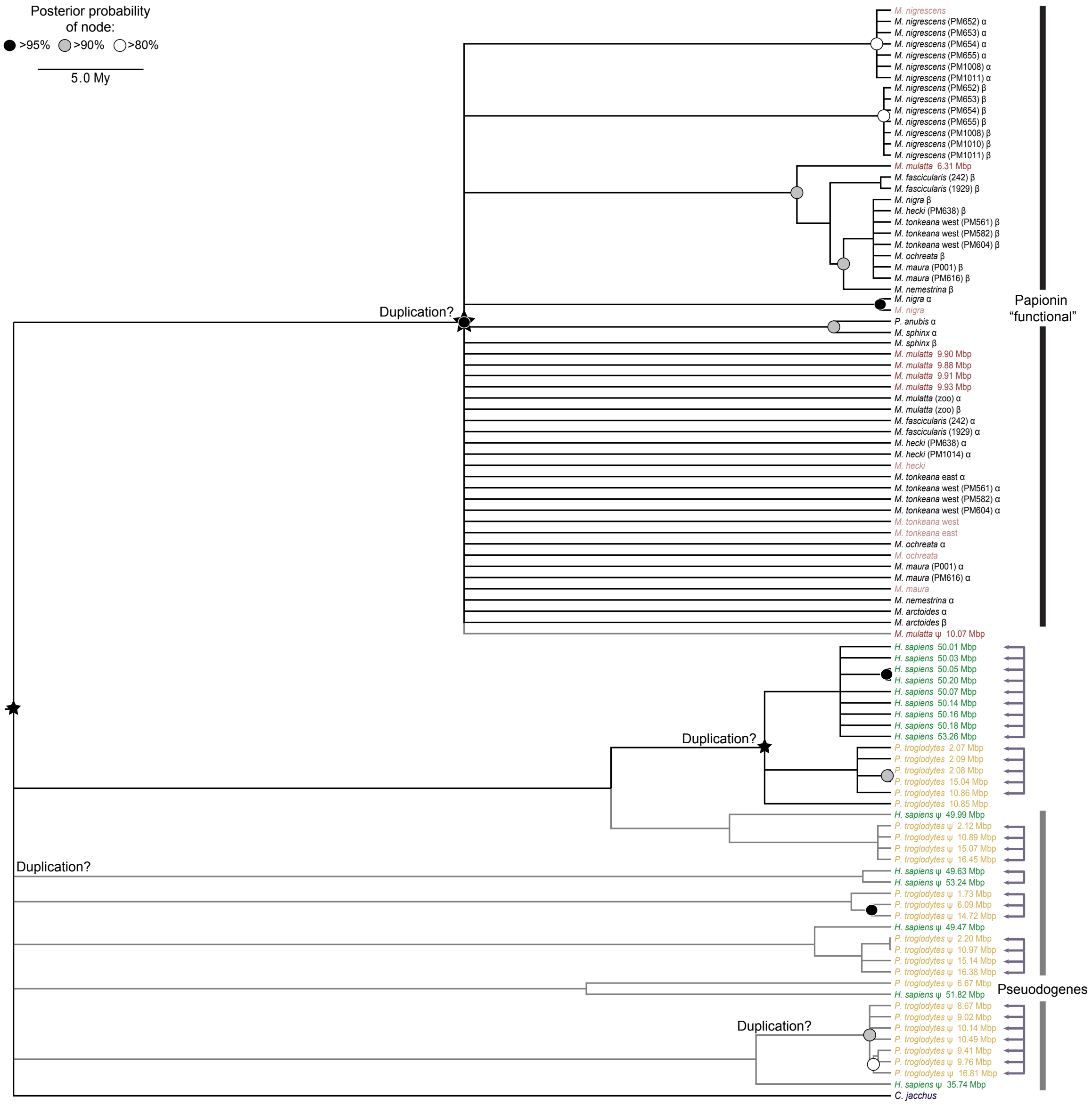


Figure A5: **Maximum clade credibility tree of *TSPY*.** Sulawesi macaque species written in pink were sequenced by [144] (see Table A4). Papionin orthologs of rhesus *TSPY2-5* are labeled “α” and orthologs of *TSPY1* are labeled “β,” corresponding to the primer set (“a” or “b,” respectively, in Table A3) that was used for sequencing. Other features of the tree are drawn as detailed in Figure A1.

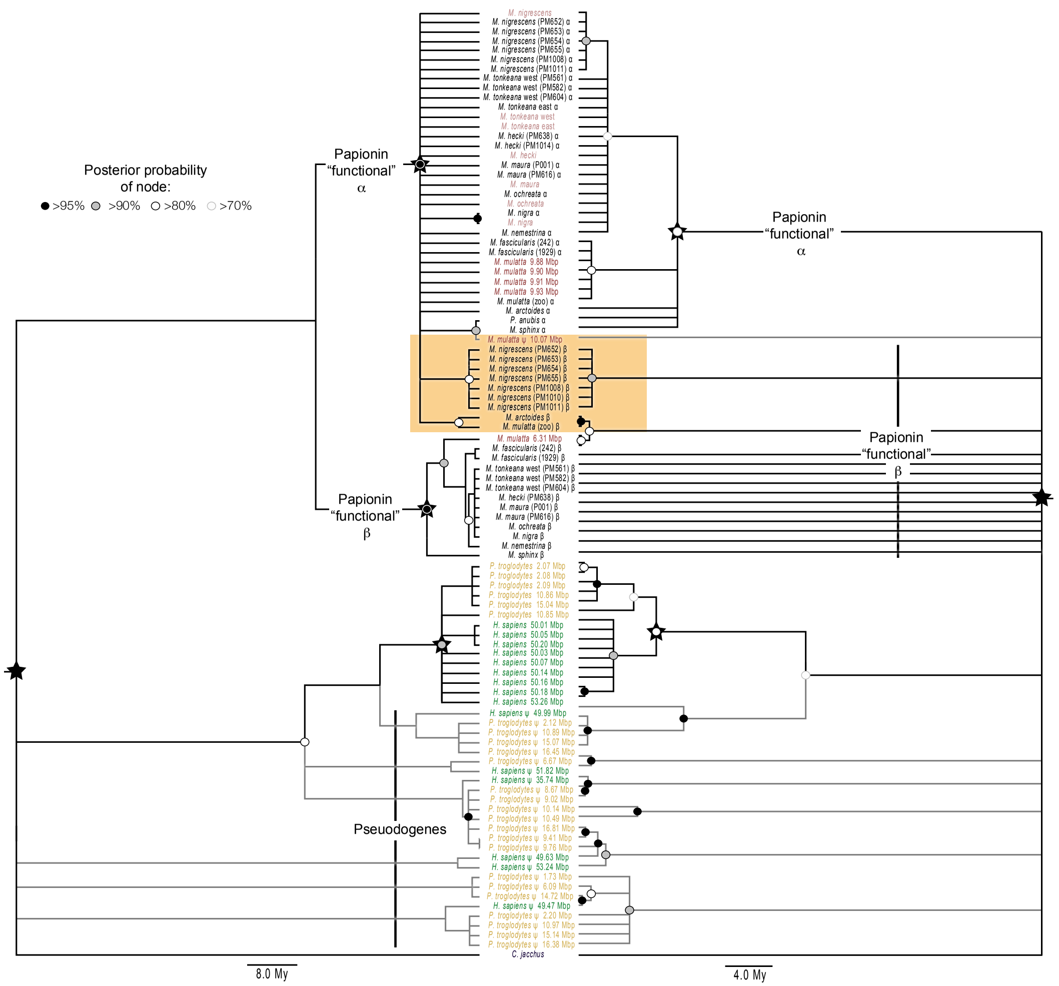


Figure A6: **Maximum clade credibility trees of the 5′ half (left) and 3′ half (right) of *TSPY*.** The trees are labeled as in Figure A5 except that nodes marked with gray outlined open circles have a posterior probability of $> 70\%$. Partial gene conversion events where the relationships between leaves differ between the 5′ and 3′ end are highlighted by the yellow box.

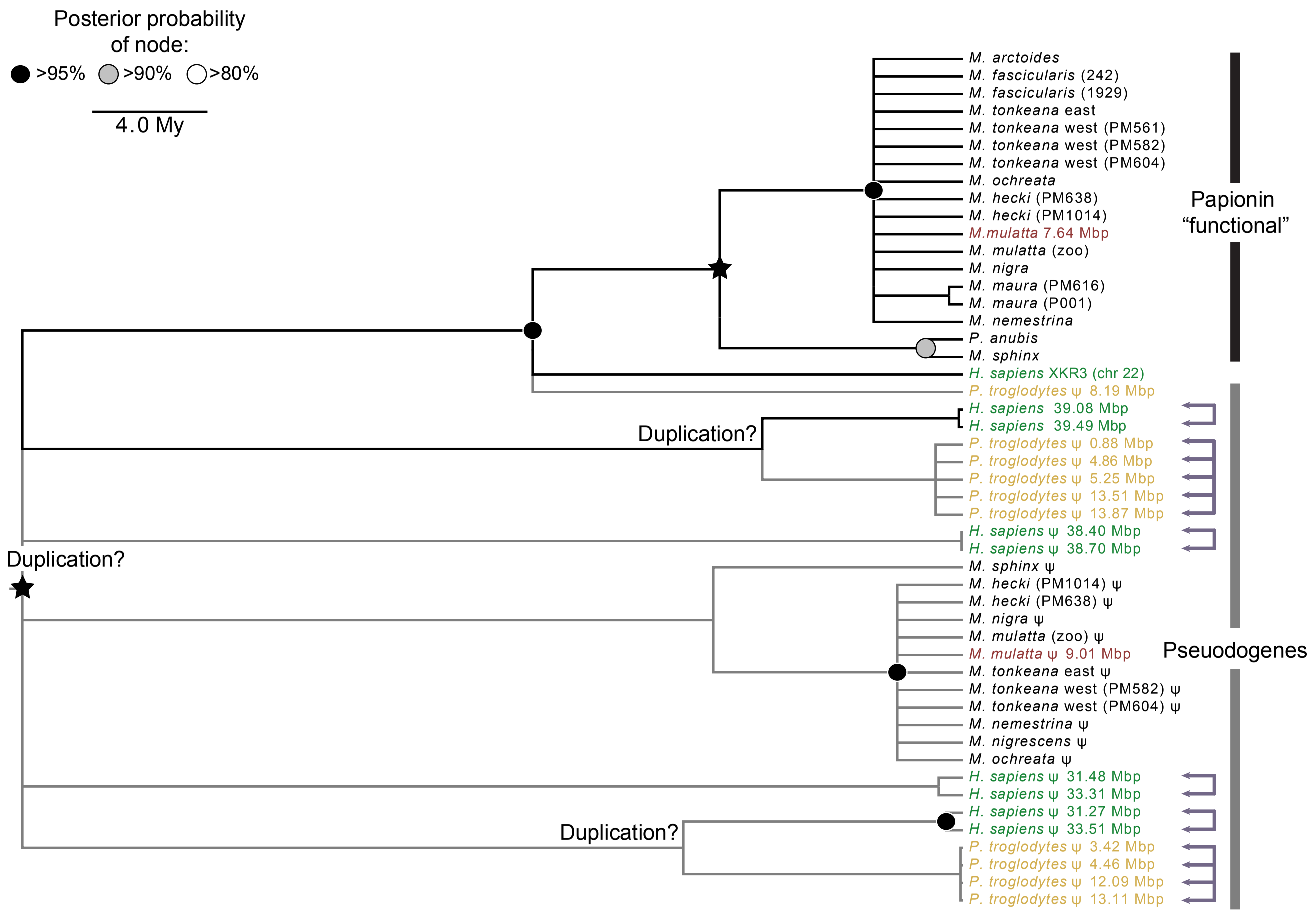


Figure A7: **Maximum clade credibility tree of *XKRY*.** The human autosomal paralog *XKR3*, which was transposed from the msrY to chromosome three in the Old World primate ancestor [100], is also included. Other features of the tree are drawn as detailed in Figure A1.

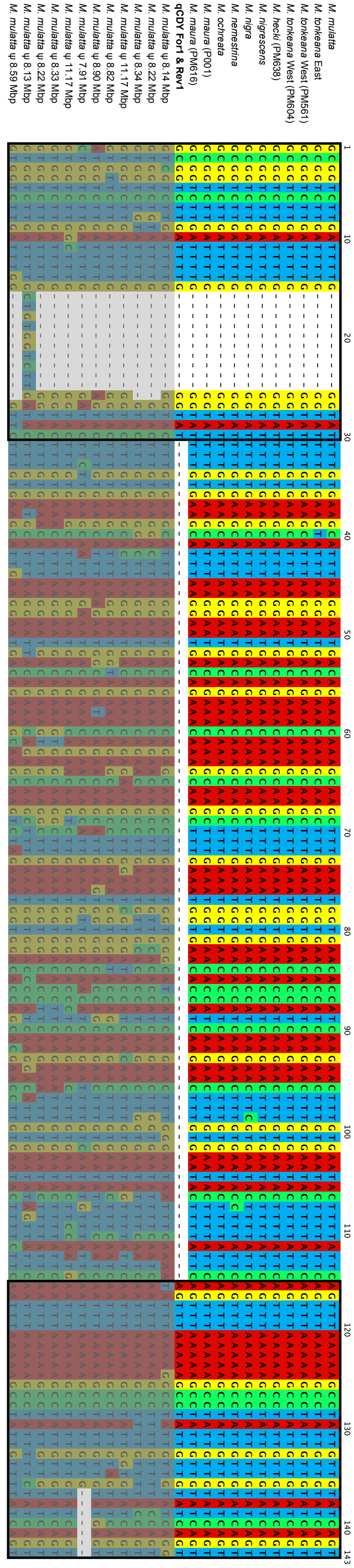


Figure A8: **Alignment of the qPCR amplicon sequence of *CDY* and its primers for select macaque samples.** The sites corresponding to the forward and reverse primer sequences are enclosed in black boxes on the left and right, respectively. Greyed-out sequences below the primers show *CDY* pseudogene sequences from *M. mulatta*, which the primers are designed to *not* amplify.

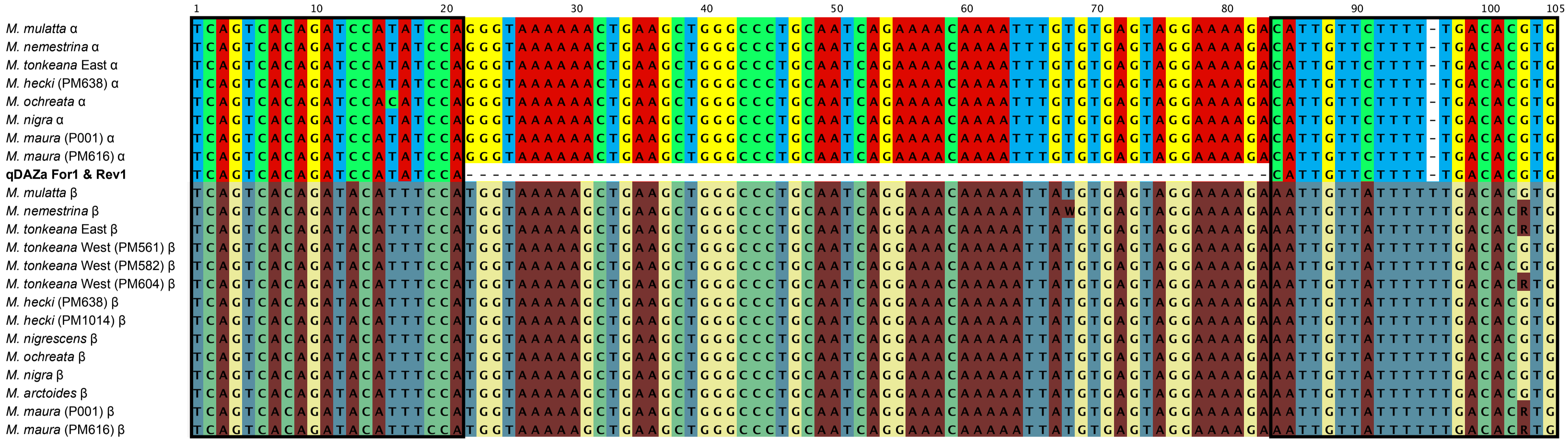


Figure A9: **Alignment of the qPCR amplicon sequence of *DAZ*-a and its primers for select macaque samples.** The sites corresponding to the forward and reverse primer sequences are enclosed in black boxes on the left and right, respectively. Washed-out sequences below the primers show *DAZ*-bcd sequences from selected macaques, which the primers are designed to *not* amplify. The *M. ochreata* sample has a substitution (T→C) at a site six nucleotides from the 5′ end of the forward primer that prevented efficient amplification of this assay in this species.

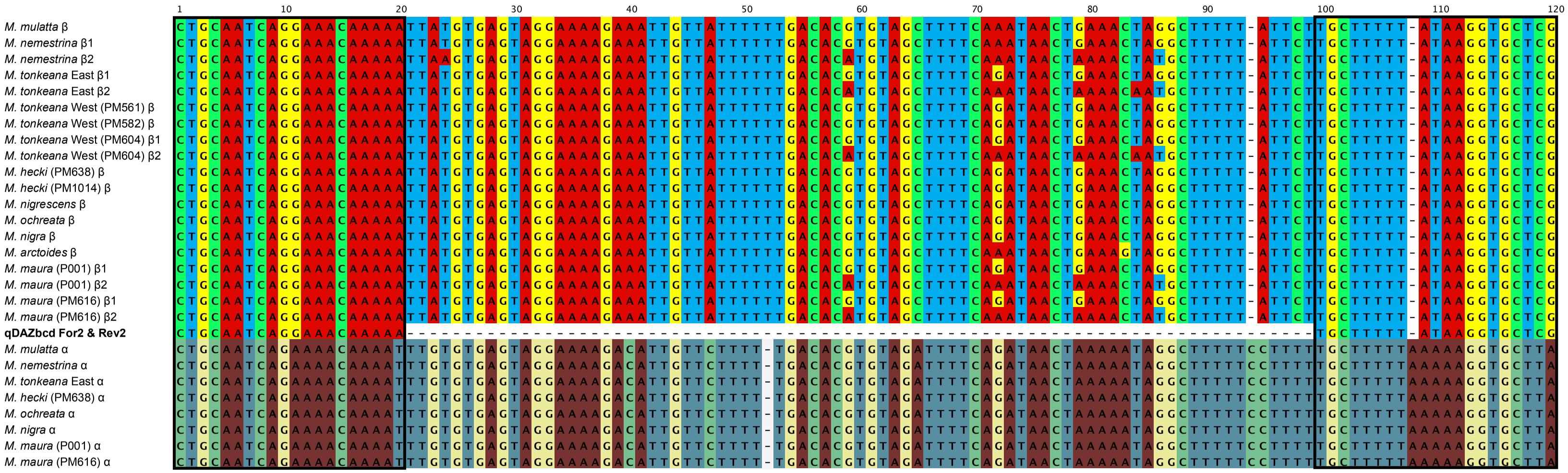


Figure A10: **Alignment of the qPCR amplicon sequence of *DAZ*-bcd and its primers for select macaque samples.** The sites corresponding to the forward and reverse primer sequences are enclosed in black boxes on the left and right, respectively. Washed-out sequences below the primers show *DAZ*-a sequences from selected macaques, which the primers are designed to *not* amplify. Some samples (e.g. *M. nemestrina* and *M. tonkeana* East) exhibited sequence differences between exon repeats; different sequenced haplotypes are shown for these samples and labeled “$\beta 1$” and “$\beta 2$.”

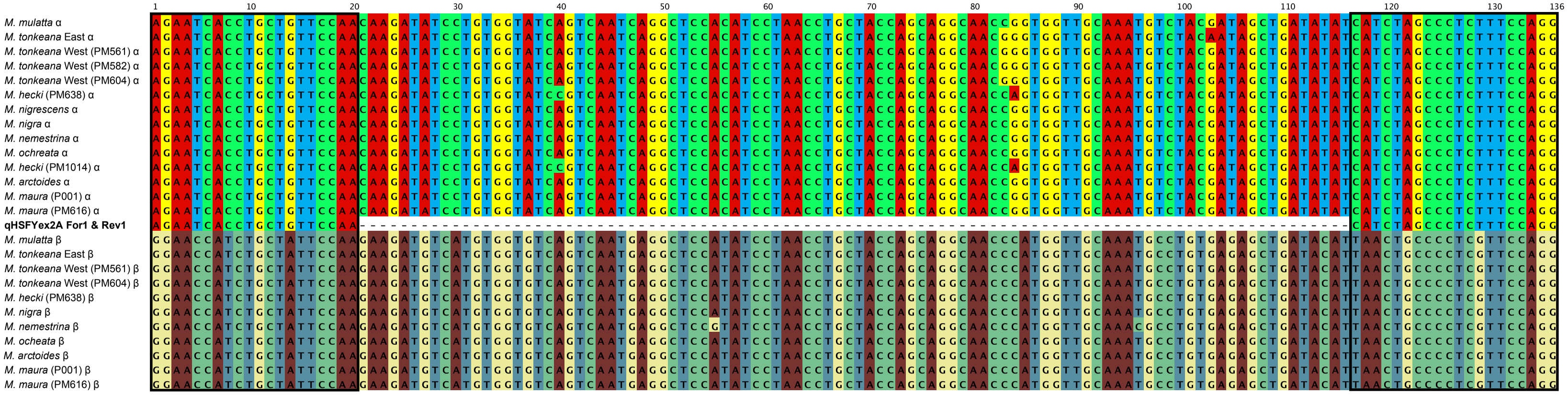


Figure A11: **Alignment of the qPCR amplicon sequence of *HSFY*-a and its primers for all of the macaque samples.** The sites corresponding to the forward and reverse primer sequences are enclosed in black boxes on the left and right, respectively. Washed-out sequences below the primers show *HSFY*-b sequences from selected macaques, which the primers are designed to *not* amplify.

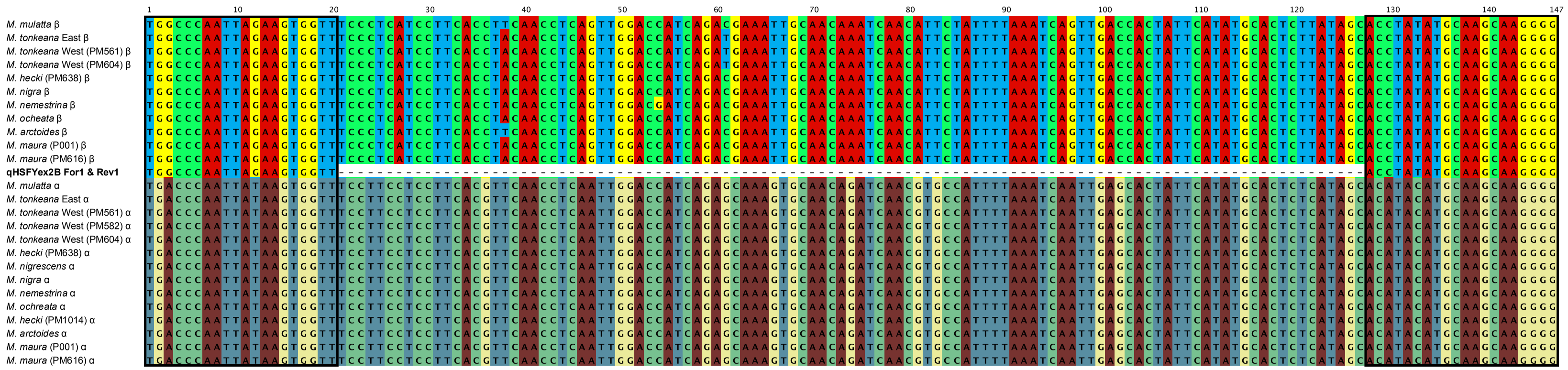


Figure A12: **Alignment of the qPCR amplicon sequence of *HSFY*-b and its primers for select macaque samples.** The sites corresponding to the forward and reverse primer sequences are enclosed in black boxes on the left and right, respectively. Washed-out sequences below the primers show *HSFY*-a sequences from macaques, which the primers are designed to *not* amplify.

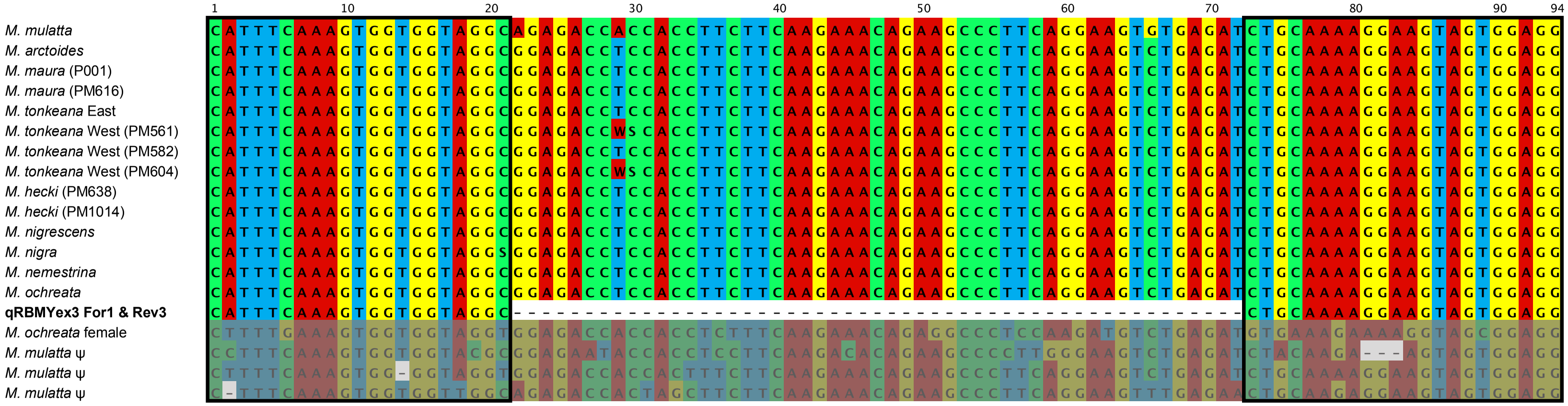


Figure A13: **Alignment of the qPCR amplicon sequence of *RBMY* and its primers for all of the macaque samples.** The sites corresponding to the forward and reverse primer sequences are enclosed in black boxes on the left and right, respectively. Greyed-out sequences below the primers show a highly similar sequence isolated from a *M. tonkeana* female (sample PF559; as well as macaque males) and *RBMY* pseudogene sequences from *M. mulatta*; primers were designed to avoid amplification of any of these non-target sequences.

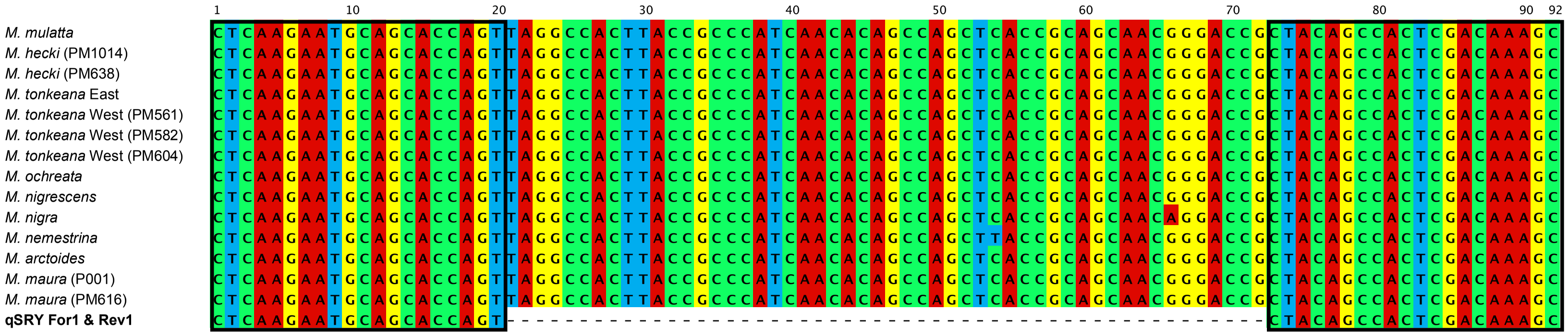

1
10
20
30
40
50
60
70
80
90
92
M. mulatta CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. hecki (PM1014) CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. hecki (PM638) CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. tonkeana East CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. tonkeana West (PM561) CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. tonkeana West (PM582) CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. tonkeana West (PM604) CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. ochreata CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. nigrescens CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. nigra CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACAGGACCGCTACAGCCACTCGACAAAGC
M. nemestrina CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTTACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. arctoides CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. maura (P001) CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
M. maura (PM616) CTCAAGAATGCAGCACCAGTTAGGCCACTTACCGCCCATCAACACAGCCAGCTCACCGCAGCAACGGGACCGCTACAGCCACTCGACAAAGC
qSRY For1 & Rev1 CTCAAGAATGCAGCACCAGT--CTACAGCCACTCGACAAAGC

Figure A14: **Alignment of the qPCR amplicon sequence of *SRY* and its primers for all of the macaque samples.** The sites corresponding to the forward and reverse primer sequences are enclosed in black boxes on the left and right, respectively.

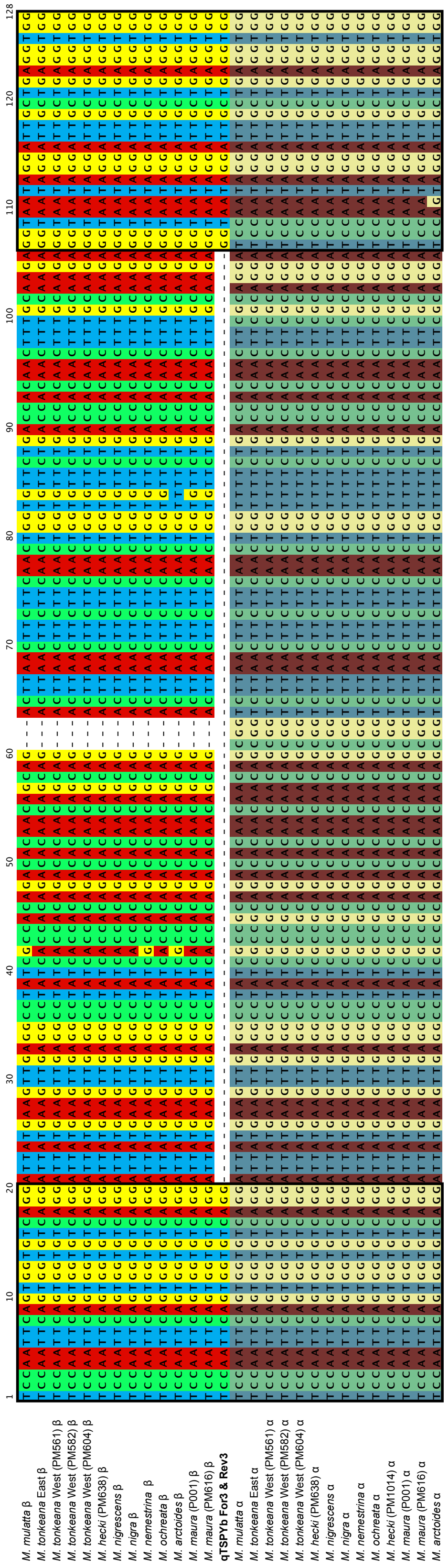


Figure A15: **Alignment of the qPCR amplicon sequence of *TSPY*-b and its primers for select macaque samples.** The sites corresponding to the forward and reverse primer sequences are enclosed in black boxes on the left and right, respectively. Washed-out sequences below the primers show *TSPY*-a sequences from macaques, which the primers are designed to *not* amplify.

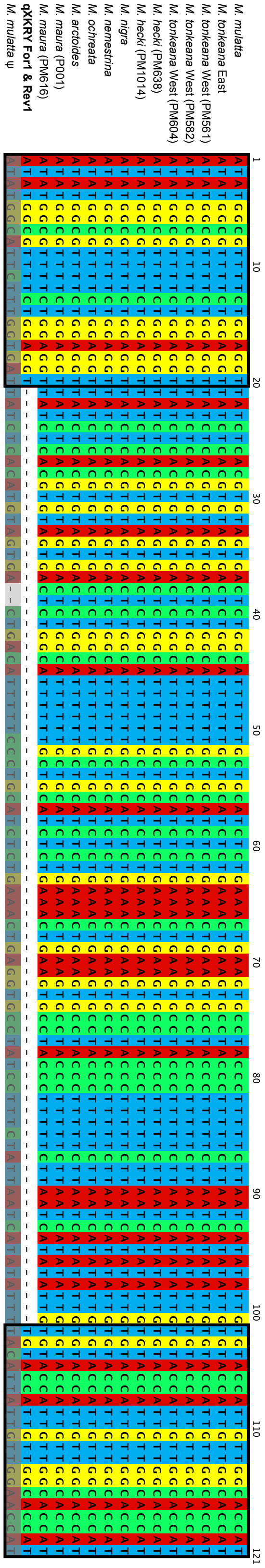
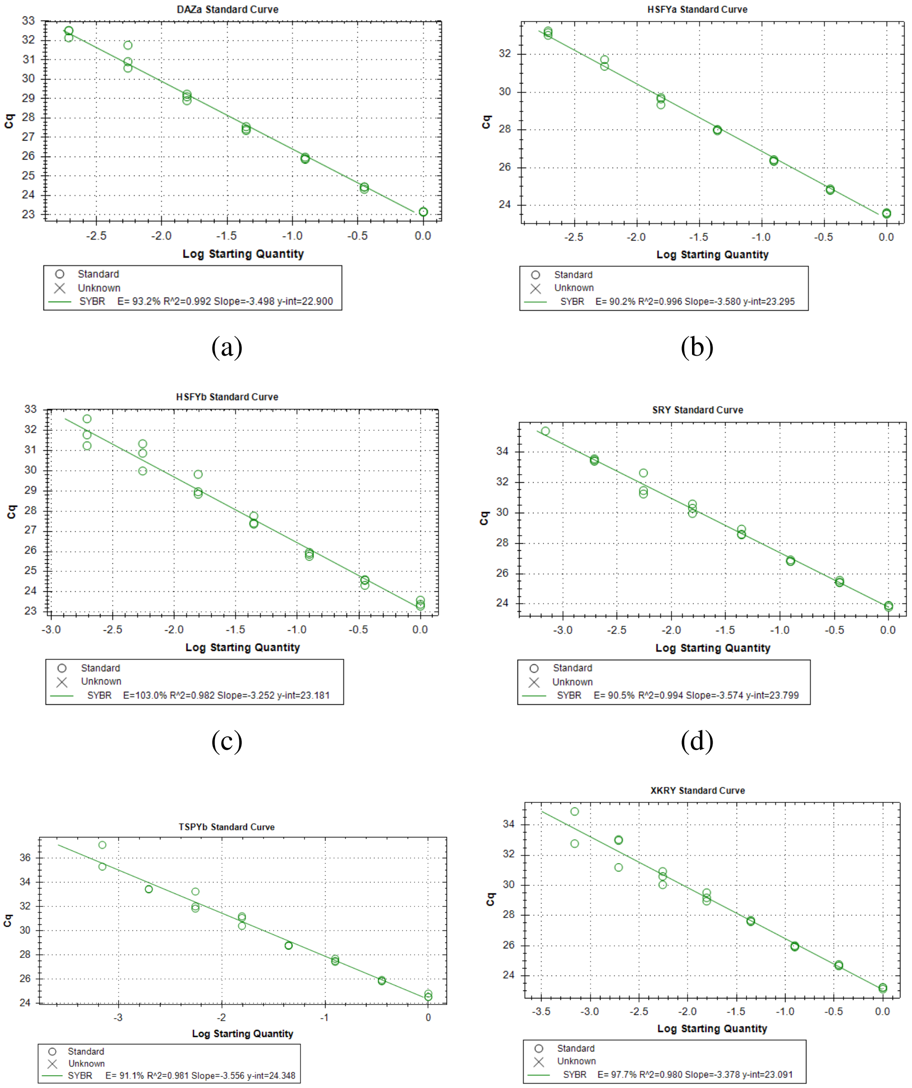


Figure A16: **Alignment of the qPCR amplicon sequence of *XKRY* and its primers for select macaque samples.** The sites corresponding to the forward and reverse primer sequences are enclosed in black boxes on the left and right, respectively. Greyed-out sequences below the primers show a *XKRY* pseudogene sequence from *M. mulatta*, which the primers are designed to *not* amplify.

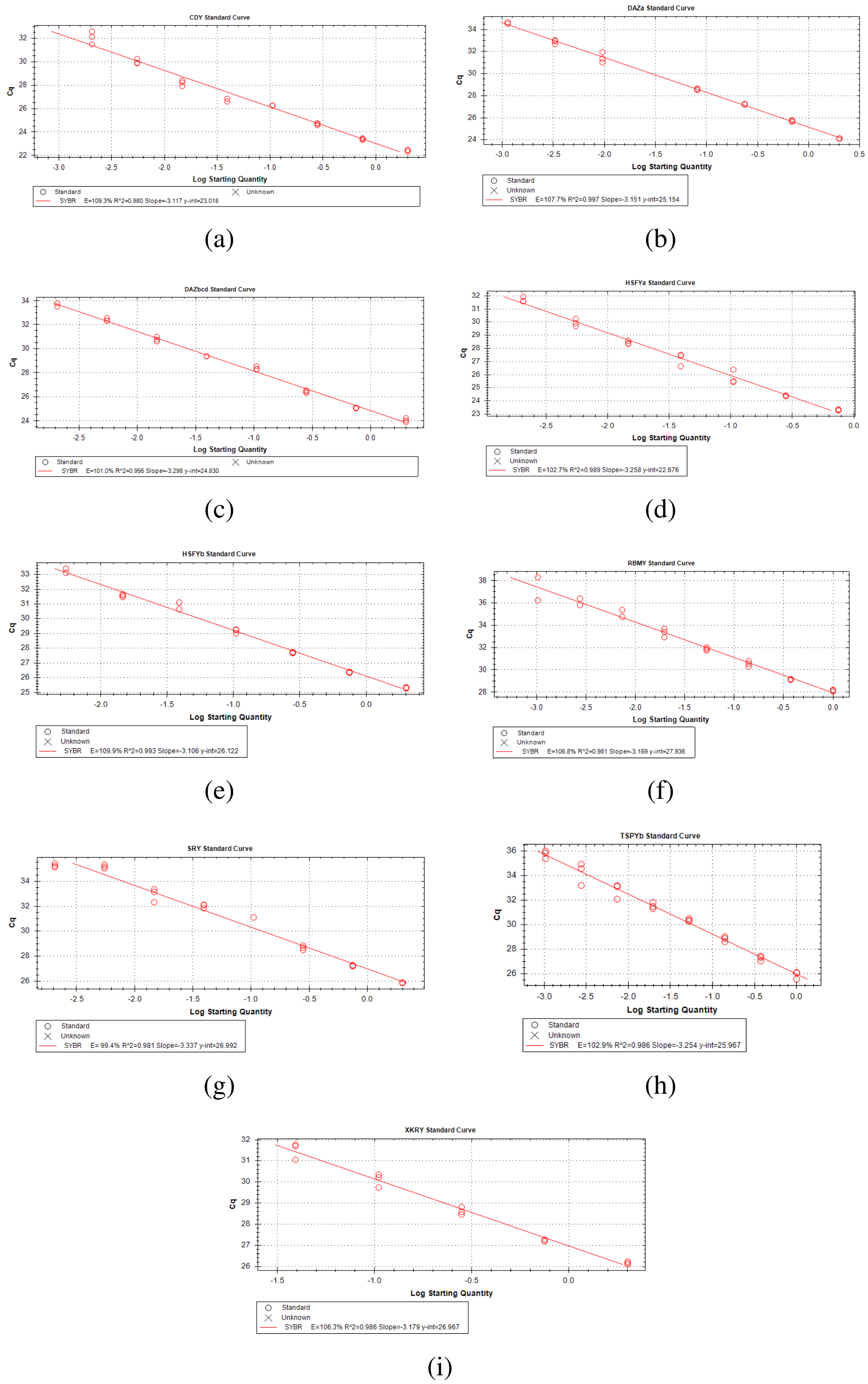


(a) (b)

(c) (d)

(e) (f)

Figure A17: **Standard curves for the set of preliminary assays performed using the iTaq Fast SYBR Green Supermix with ROX.** (a) shows assay DAZa, (b) shows assay HSFYa, (c) shows assay HSFYb, (d) shows assay SRY, (e) shows assay TSPYb, and (f) shows assay XKRY.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure A18: **Standard curves for the assays performed using the SsoFast EvaGreen Supermix.** (a) shows CDY, (b) shows DAZa, (c) shows DAZbcd, (d) shows HSFYa, (e) shows HSFYb, (f) shows RBMY, (g) shows SRY, (h) shows TSPYb, and (i) shows XKRY.

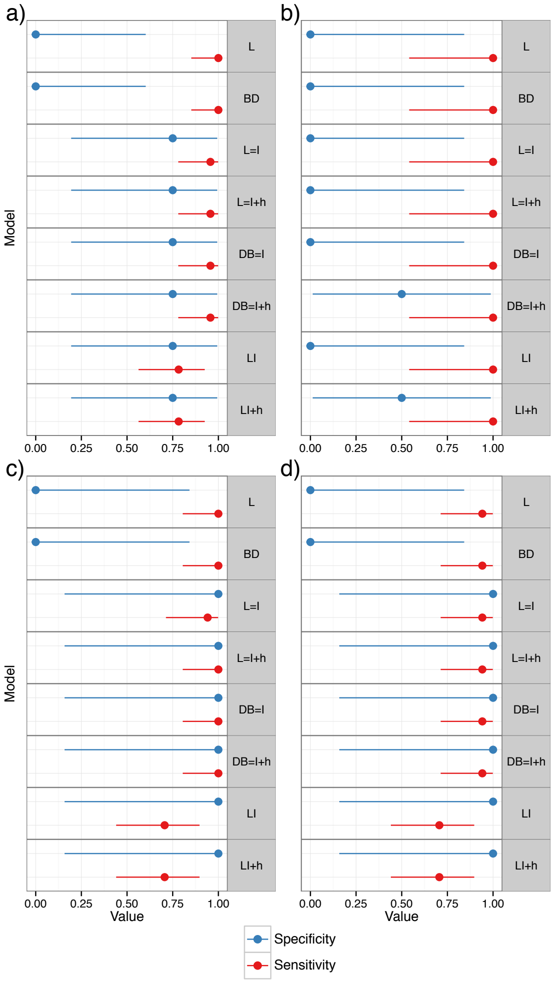


Figure A19: **The specificity and sensitivity of models in identifying the ancestral state of msrY-linked gene families.** The true negative rate (i.e. specificity) and the true positive rate (i.e. sensitivity) of each model to correctly identify gene families that were absent (i.e. true negatives) or present (i.e. true positives) in the MRCAs of Old World primates (a-c) and Hominini (d) as determined by [54, 89, 94]. Points show the estimated values and lines show 95% confidence intervals estimated from a Binomial distribution with $p = 0.5$ and n equal to the number of gene families. **a)** Models are fit to the complete Y-chromosome data. There are 27 gene families, of which four are assumed to be absent in the MRCA: the two X-transposed gene families, *PCDH11Y* and *TGIF2LY*, and the AGs *PRY* and *VCY*. **b)** Models are fit to AG data. There are eight gene families, of which the same two AGs are absent as in a). **c)** Models are fit to the Y-linked singleton data. There are 19 gene families, of which the same two singletons are absent as in a). **d)** Models are fit as in c) but compared against the marginal ancestral reconstruction at the Hominini MRCA, where the same two singletons are absent as in a)

CHAPTER 3

CONCLUSIONS AND PERSPECTIVES

As promised at the end of the introductory chapter, my work has been able to address three questions regarding AGs evolution on Old World primate msrYs: **(i)** How common is gene conversion?; **(ii)** How frequent are duplication and deletion events?; and, **(iii)** Do copy numbers evolve differently between AGs, Y-linked single copy genes, and autosomal genes? The brief answers to these questions are below.

(i) Qualitatively, gene conversion appears to be about as frequent among macaque monkeys as it is in the rhesus macaque. This is in agreement with the conclusion of [54] that humans and chimpanzees have the fastest rates of gene conversion among the Old World primate species that have been investigated. **(ii)** AG copy numbers evolve at a rate of ≈ 1 event per copy per 5 My. Surprisingly, this estimate is quite similar to the best available estimate for the rate of spontaneous nucleotide mutations on the human msrY [217]. This seems very fast given that our estimate was determined from phylogenetic data and is therefore not a measure of the spontaneous mutation rate of copy number variants. It would be interesting to estimate the within-population AG family deletion/duplication rate in the primate species investigated because this could assay whether purifying or directional selection is acting on the AG copy number, as suggested by [104]. **(iii)** AG copy numbers may evolve about an order of magnitude faster than Y-linked singletons and autosomal gene families. This slower rate of copy number evolution for Y-linked singletons and autosomal genes is in agreement with the expectations of very strong purifying selection [99] and efficient selection, respectively, for these gene families.

The remainder of this chapter is a contemplation of the topics that, in my opinion, need to be further elaborated in order for evolutionary biologists to make sense of AG evolution. As such, I have included a personal wish-list of future directions.

The linkage problem

As previously explained, linkage is probably the most important constraint acting on the msrY. What makes the mammalian msrY particularly intriguing is that non-ampliconic regions experience (almost) no recombination, while the nearby ampliconic regions experience abundant recombination (i.e. through gene conversion), and both of these regions together experience clonal inheritance. This means that the msrY as a whole has a much smaller N_e than autosomal regions but also that ampliconic regions within the msrY can vary greatly in their local N_e [112].

Despite its known significance, linkage continues to pose a problem for most evolutionary biologists empirically investigating the msrY. For example, [52] is the most comprehensive study to have looked at how selection is operating on the msrY in primates but the authors did not investigate all of the msrY-linked genes. By looking at msrY-linked singleton genes in human populations and accounting for population history, they concluded that the observed nucleotide variation is *too low* to be explained by purifying selection acting on these genes alone. The authors invoke a very dissatisfying *deus ex machina*-like construction at the end of the paper by suggesting that the ampliconic genes – which they did not investigate – are responsible for their results. Regardless of how dissatisfied we may feel as readers, [52] is right in stressing that the msrY operates under strong linkage constraints. Therefore, all genes on the msrY must be investigated concurrently in order to begin to make sense of its evolution.

In our study, we considered the msrY as a whole by looking at both the singleton genes and AGs. However, a major drawback of our models is that they assume independent evolution of gene families, and therefore, do not accommodate for linkage among gene families. This is because we are not sure how to feasibly include this constraint. Further development of models that are computationally efficient and do away with the assumption of independence between sites or gene families is a requirement for enriching our understanding of msrY evolution.

Identifying the target(s) of selection

Another lesson from [52] is that strong linkage makes it incredibly difficult for evolutionary biologists to detect the targets of selection on the msrY. This is because selection acting on any target of the msrY will produce a selective sweep in the population, wherein a whole chromosome will either rise to a higher frequency or decrease to a lower frequency in the population, depending on the direction of selection. As explained in Chapter 1, selection in old msrYs may be acting on other features than just functional genes; for example, the repetitive ampliconic regions [52, 113], inversions that prevent recombination between the X and Y chromosomes [58, 59], or regions implicated in chromosome fragility [60, 61]. For this reason, future studies that aim to identify the target(s) of selection on the msrY should sequence the entire genic and intergenic regions of the msrY, or accompany targeted gene sequencing with methods that assay the whole structure of the chromosome.

One group has previously attempted to study the evolution of msrY ampliconic genic and intergenic duplicate structures in great apes, both within populations and between species [110, 111]. However, these studies did not use an explicit model of evolution to analyze their results, instead relying on verbal arguments that invoke maximum parsimony. Unfortunately, since chromosome rearrangements and gene duplications are common in primate msrYs ([111] and Chapter 2), likelihood methods with explicit models of evolution must be used instead of parsimony methods in order to accurately interpret the evolutionary data [213]. Our study provides an example for future work of how likelihood methods can be used with an explicit model of gene family evolution while taking into account the phylogenetic relationships among species.

The gene conversion problem

Gene conversion is another limitation in the study of AG evolution because it erases the information found in nucleotide substitutions among intra-specific paralogs. Not only does this homogenization of nucleotide sequences mean that new models must be developed in order to study these gene families (e.g. Chapter 2), it also means that there is less information available with which to inform those models. For example, in order to distinguish between *CDY* paralogs in a rhesus macaque individual, one would need to sequence through as much as 50 kbp of continuous DNA sequence, of which less than 1% would contain nucleotide variation [54]. Even in the age of high through-put genomics, that is a lot of buck for not much bang!

In order to understand the micro-evolutionary forces that currently govern AGs, one would need to do a population study. However, most methods of detecting selection within a population do so by identifying reduced levels of genetic diversity [218]. This is problematic for msrY-linked genes because the levels of genetic diversity are already very low. The high frequency of gene conversion among paralogs makes the study of AG families interesting because this process leads to an overall increase in the genetic diversity in the population. Nevertheless, the amount of sequencing that would be required for such an undertaking would probably prove prohibitive with current short-read sequencing methods. One way to get around this problem would be to identify a study system where gene conversion is infrequent relative to the mutation rate.

Identifying the causes of heterogeneity

As explained in the introduction (Chapter 1), there are many potential biological methods that can lead to heterogeneity among lineages and between different genomic regions. One frustration that I have with our results is that, although we were able to detect two different types of heterogeneity, we were not able to state why these heterogeneities are present. This inability to identify the causal factors is an inherent problem of correlational studies. One way to understand the role of factors like N_e and mating system on AG family evolution, respectively, would be to look at the msrYs of *Drosophila* [219] inbred lines kept at different N_e or under different mating systems over multiple generations.

Another draw-back of our results is that we had insufficient power to detect lineage heterogeneity among AGs, if such heterogeneity does indeed exist. Since msrYs are gene depauperate and have few completed sequences available, low power is a real limitation that evolutionary biologists currently face in studying AG family evolution. Hopefully this problem will be alleviated as more msrY sequences are completed.

In spite of our inability to make strong conclusions about the causes of the observed heterogeneity, our study design serves as an excellent model for future investigations into AG evolution. By looking at a rapid allopatric speciation of macaque monkeys [143, 220] we were able to observe multiple independent evolution events from a common ancestor without worrying about the confounding effects that changes in selective or mutational processes have had on AG evolution (i.e. in contrast to studies comparing closely related species with diverged mating systems [96, 110, 111]).

BIBLIOGRAPHY

[1] Hellemans, J., G. Mortier, A. D. Paepe, F. Speleman, J. Vandesompele, “qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data”, *Genome Biol* **2007**, *8*, R19.

[2] Muller, H. J., “Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors”, *Genetics* **1918**, *3*, 422.

[3] Nelson-Rees, W. A., A. J. Kniazeff, R. L. Malley, N. B. Darby Jr, “On the karyotype of the tahr *Hemitragus jemlahicus* and the Y-chromosome of goats and sheep”, *Chromosoma* **1967**, *23*, 154–61.

[4] Baverstock, P., C. Watts, J. T. Hogarth, “Polymorphism of the X-chromosome, Y-chromosome and autosomes in the Australian hopping mice, *Notomys alexis, N. cervinus and N. fuscus* (Rodentia, Muridae)”, *Chromosoma* **1977**, *61*, 243–256.

[5] Potter, W., P. Upton, “Y chromosome morphology of cattle”, *Aust Vet J* **1979**, *55*, 539–41.

[6] Robinson, T., P. Condy, “The chromosomes of the southern elephant seal, *Mirounga leonina* (Phocidae: Mammalia)”, *Cytogenet Genome Res* **1979**, *23*, 157–62.

[7] Huxley, J., *Evolution: the modern synthesis*, George Alien & Unwin Ltd, London, **1942**.

[8] Haldane, J., “The mutation rate of the gene for haemophilia, and its segregation ratios in males and females”, *Ann Hum Genet* **1946**, *13*, 262–71.

[9] Miyata, T, H Hayashida, K Kuma, K Mitsuyasu, T Yasunaga, “Male-driven molecular evolution: a model and nucleotide sequence analysis”, *Cold Spring Harb Symp Quant Biol* **1987**, *52*, 863–7.

[10] Hurst, L. D., H. Ellegren, “Sex biases in the mutation rate”, *TIG* **1998**, *14*, 446–52.

[11] Evans, B. J., K. Zeng, J. A. Esselstyn, B. Charlesworth, D. J. Melnick, “Reduced representation genome sequencing suggests low diversity on the sex chromosomes of tonkean macaque monkeys”, *Mol Biol Evol* **2014**, *31*, 2425–40.

[12] Wilson Sayres, M. A., K. D. Makova, “Genome analyses substantiate male mutation bias in many species”, *Bioessays* **2011**, *33*, 938–45.

[13] Drost, J. B., W. R. Lee, “Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human”, *Environ Mol Mutagen* **1995**, *25*, 48–64.

[14] Arnheim, N., P. Calabrese, "Understanding what determines the frequency and pattern of human germline mutations", *Nat Rev Genet* **2009**, *10*, 478–88.

[15] Noordam, M. J., S. Repping, "The human Y chromosome: a masculine chromosome", *Curr Opin Genet Dev* **2006**, *16*, 225–32.

[16] Jiang, P.-P., D. L. Hartl, B. Lemos, "Y not a dead end: epistatic interactions between Y-linked regulatory polymorphisms and genetic background affect global gene expression in *Drosophila melanogaster*", *Genetics* **2010**, *186*, 109–18.

[17] Mank, J. E., "Small but mighty: the evolutionary dynamics of W and Y sex chromosomes", *Chromosome Res* **2012**, *20*, 21–33.

[18] Navarro-Costa, P., C. E. Plancha, J. Gonçalves, "Genetic dissection of the AZF regions of the human Y chromosome: thriller or filler for male (in)fertility?", *J Biomed Biotechnol* **2010**, *2010*.

[19] Griffin, R. M., D. Le Gall, H. Schielzeth, U. Friberg, "Within-population Y-linked genetic variation for lifespan in *Drosophila melanogaster*", *J Evol Biol* **2015**.

[20] Dean, R., B. Lemos, D. K. Dowling, "Context-dependent effects of Y chromosome and mitochondrial haplotype on male locomotive activity in Drosophila melanogaster", *J Evol Biol* **2015**.

[21] Navarro-Costa, P., "Sex, rebellion and decadence: The scandalous evolutionary history of the human Y chromosome", *BBA–Mol Basis Dis* **2012**, *1822*, 1851–63.

[22] Chang, T.-C., Y. Yang, E. F. Retzel, W.-S. Liu, "Male-specific region of the bovine Y chromosome is gene rich with a high transcriptomic activity in testis development", *PNAS* **2013**, *110*, 12373–8.

[23] Moghadam, H. K., M. A. Pointer, A. E. Wright, S. Berlin, J. E. Mank, "W chromosome expression responds to female-specific selection", *PNAS* **2012**, *109*, 8207–11.

[24] Kirkpatrick, M., R. F. Guerrero, "Signatures of sex-antagonistic selection on recombining sex chromosomes", *Genetics* **2014**, *197*, 531–41.

[25] Zhou, Q., D. Bachtrog, "Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*", *Science* **2012**, *337*, 341–5.

[26] Masel, J., "Genetic drift", *Curr Biol* **2011**, *21*, R837–R838.

[27] Hedrick, P. W., "Sex: differences in mutation, recombination, selection, gene flow, and genetic drift", *Evolution* **2007**, *61*, 2750–71.

[28] Bartolomé, C., B. Charlesworth, "Evolution of amino-acid sequences and codon usage on the *Drosophila miranda* neo-sex chromosomes", *Genetics* **2006**, *174*, 2033–44.

[29] Shikano, T., H. M. Natri, Y. Shimada, J. Merilä, "High degree of sex chromosome differentiation in stickleback fishes", *BMC Genomics* **2011**, *12*, 474.

[30] Orr, H. A., Y. Kim, "An adaptive hypothesis for the evolution of the Y chromosome", *Genetics* **1998**, *150*, 1693–8.

[31] Graves, J. A. M., "Sex chromosome specialization and degeneration in mammals", *Cell* **2006**, *124*, 901–14.

[32] Na, J.-K., J. Wang, R. Ming, "Accumulation of interspersed and sex-specific repeats in the non-recombining region of papaya sex chromosomes", *BMC Genomics* **2014**, *15*, 335.

[33] Pennell, M. W., M. Kirkpatrick, S. P. Otto, J. C. Vamosi, C. L. Peichel, N. Valenzuela, J. Kitano, "Y fuse? Sex chromosome fusions in fishes and reptiles", *PLOS Genetics* **2015**, e1005237.

[34] Charlesworth, B., "Model for evolution of Y chromosomes and dosage compensation", *PNAS* **1978**, *75*, 5618–22.

[35] Hill, W. G., A. Robertson, "The effect of linkage on limits to artificial selection", *Genet Res* **1966**, *8*, 269–94.

[36] Bachtrog, D., M. Kirkpatrick, J. E. Mank, S. F. McDaniel, J. C. Pires, W. Rice, N. Valenzuela, "Are all sex chromosomes created equal?", *Trends Genet* **2011**, *27*, 350–7.

[37] McGaugh, S. E., C. S. Heil, B. Manzano-Winkler, L. Loewe, S. Goldstein, T. L. Himmel, M. A. Noor, "Recombination modulates how selection affects linked sites in *Drosophila*", *PLOS Biol* **2012**, *10*, e1001422.

[38] Bachtrog, D, "The temporal dynamics of processes underlying Y chromosome degeneration", *Genetics* **2008**, *179*, 1513–25.

[39] Muller, H. J., "The relation of recombination to mutational advance", *Mutat Res* **1964**, *1*, 2–9.

[40] Felsenstein, J, "The evolutionary advantage of recombination", *Genetics* **1974**, *78*, 737–56.

[41] Charlesworth, B., D. Charlesworth, "Rapid fixation of deleterious alleles can be caused by Muller's ratchet", *Genet Res* **1997**, *70*, 63–73.

[42] Bergero, R, S Qiu, D Charlesworth, "Gene loss from a plant sex chromosome system", *Curr Biol* **2015**, *25*, 1234–40.

[43] Zhou, Q., D. Bachtrog, "Chromosome-wide gene silencing initiates Y degeneration in *Drosophila*", *Curr Biol* **2012**, *22*, 522–5.

[44] Fisher, R. A., *The genetical theory of natural selection: a complete variorum edition*, Oxford University Press, **1930**.

[45] Charlesworth, B, M. T. Morgan, D Charlesworth, "The effect of deleterious mutations on neutral molecular variation", *Genetics* **1993**, *134*, 1289–303.

[46] Peck, J. R., "A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex", *Genetics* **1994**, *137*, 597–606.

[47] Handley, L. L., R. Hammond, G Emaresi, A Reber, N Perrin, "Low Y chromosome variation in Saudi-Arabian hamadryas baboons (*Papio hamadryas hamadryas*)", *Heredity* **2006**, *96*, 298–303.

[48] Charlesworth, B., "Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture", *J Hered* **2013**, *104*, 161–71.

[49] Smith, J. M., J. Haigh, "The hitch-hiking effect of a favourable gene", *Genet Res* **1974**, *23*, 23–35.

[50] Rice, W. R., "Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome", *Genetics* **1987**, *116*, 161–7.

[51] Gerrard, D. T., D. A. Filatov, "Positive and negative selection on mammalian Y chromosomes", *Mol Biol Evol* **2005**, *22*, 1423–32.

[52] Wilson Sayres, M. A., K. E. Lohmueller, R. Nielsen, "Natural selection reduced diversity on human Y chromosomes", *PLOS Genetics* **2014**, *10*, e1004064.

[53] Blackmon, H., J. P. Demuth, "Estimating tempo and mode of Y chromosome turnover: explaining Y chromosome loss with the fragile Y hypothesis", *Genetics* **2014**, *197*, 561–72.

[54] Hughes, J. F. et al., "Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes", *Nature* **2012**, *483*, 82–6.

[55] Sayres, M. A. W., K. D. Makova, "Gene survival and death on the human Y chromosome", *Mol Biol Evol* **2013**, *30*, 781–7.

[56] Wright, A. E., P. W. Harrison, S. H. Montgomery, M. A. Pointer, J. E. Mank, "Independent stratum formation on the avian sex chromosomes reveals inter-chromosomal gene conversion and predominance of purifying selection on the W chromosome", *Evolution* **2014**, *68*, 3281–95.

[57] Lahn, B. T., D. C. Page, "Four evolutionary strata on the human X chromosome", *Science* **1999**, *286*, 964–7.

[58] Bergero, R., S. Qiu, A. Forrest, H. Borthwick, D. Charlesworth, "Expansion of the pseudo-autosomal region and ongoing recombination suppression in the *Silene latifolia* sex chromosomes", *Genetics* **2013**, *194*, 673–86.

[59] Lemaitre, C., M. D. V. Braga, C. Gautier, M.-F. Sagot, E. Tannier, G. A. B. Marais, "Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes", *Genome Biol Evol* **2009**, *1*, 56–66.

[60] Hastings, P. J., J. R. Lupski, S. M. Rosenberg, G. Ira, "Mechanisms of change in gene copy number", *Nat Rev Genet* **2009**, *10*, 551–64.

[61] Zhao, J., A. Bacolla, G. Wang, K. M. Vasquez, "Non-B DNA structure-induced genetic instability and evolution", *Cell Mol Life Sci* **2010**, *67*, 43–62.

[62] Locke, D. P. et al., "Comparative and demographic analysis of orang-utan genomes", *Nature* **2011**, *469*, 529–33.

[63] Rogers, J., R. A. Gibbs, "Comparative primate genomics: emerging patterns of genome content and dynamics", *Nat Rev Genet* **2014**, *15*, 347–359.

[64] Marques-Bonet, T., J. M. Kidd, M. Ventura, T. A. Graves, Z. Cheng, L. W. Hillier, Z. Jiang, C. Baker, R. Malfavon-Borja, L. A. Fulton, C. Alkan, G. Aksay, S. Girirajan, P. Siswara, L. Chen, M. F. Cardone, A. Navarro, E. R. Mardis, R. K. Wilson, E. E. Eichler, "A burst of segmental duplications in the genome of the African great ape ancestor", *Nature* **2009**, *457*, 877–81.

[65] Ventura, M, C. R. Catacchio, C Alkan, T Marques-Bonet, S Sajjadian, T. A. Graves, F Hormozdiari, A Navarro, M Malig, C Baker, C Lee, E. H. Turner, L Chen, J. M. Kidd, N Archidiacono, J Shendure, R. K. Wilson, E. E. Eichler, "Gorilla genome

structural variation reveals evolutionary parallelisms with chimpanzee", *Genome Res* **2011**, *21*, 1640–9.

[66] Sudmant, P. H., J Huddleston, C. R. Catacchio, M Malig, L. W. Hillier, C Baker, K Mohajeri, I Kondova, R. E. Bontrop, S Persengiev, F Antonacci, M Ventura, J Prado-Martinez, Great Ape Genome Project, T Marques-Bonet, E. E. Eichler, "Evolution and diversity of copy number variation in the great ape lineage", *Genome Res* **2013**, *23*, 1373–82.

[67] Marques-Bonet, T., S. Girirajan, E. E. Eichler, "The origins and impact of primate segmental duplications", *Trends Genet* **2009**, *25*, 443–54.

[68] Gokcumen, O., V. Tischler, J. Tica, Q. Zhu, R. C. Iskow, E. Lee, M. H.-Y. Fritz, A. Langdon, A. M. Stütz, P. Pavlidis, V. Benes, R. E. Mills, P. J. Park, C. Lee, J. O. Korbel, "Primate genome architecture influences structural variation mechanisms and functional consequences", *PNAS* **2013**, *110*, 15764–9.

[69] Fang, X., Y. Zhang, R. Zhang, L. Yang, M. Li, K. Ye, X. Guo, J. Wang, B. Su, "Genome sequence and global sequence variation map with 5.5 million SNPs in Chinese rhesus macaque", *Genome Biol* **2011**, *12*.

[70] Iwase, M., Y. Satta, H. Hirai, Y. Hirai, N. Takahata, "Frequent gene conversion events between the X and Y homologous chromosomal regions in primates", *BMC Evol Biol* **2010**, *10*, 225.

[71] Trombetta, B., D. Sellitto, R. Scozzari, F. Cruciani, "Inter-and intra-species phylogenetic analyses reveal extensive XY gene conversion in the evolution of gametologous sequences of human sex chromosomes.", *Mol Biol Evol* **2014**, *31*, 2108–23.

[72] Hastings, P. J., "Mechanisms of ectopic gene conversion", *Genes* **2010**, *1*, 427–439.

[73] García-Moreno, J., D. P. Mindell, "Rooting a phylogeny with homologous genes on opposite sex chromosomes (gametologs): a case study using avian CHD", *Mol Biol Evol* **2000**, *17*, 1826–32.

[74] Rosser, Z. H., P. Balaresque, M. A. Jobling, "Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot", *Am J Hum Genet* **2009**, *85*, 130–4.

[75] Trombetta, B., F. Cruciani, P. A. Underhill, D. Sellitto, R. Scozzari, "Footprints of X-to-Y gene conversion in recent human evolution", *Mol Biol Evol* **2010**, *27*, 714–25.

[76] Lynch, M., J. S. Conery, "The evolutionary fate and consequences of duplicate genes", *Science* **2000**, *290*, 1151–55.

[77] Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, E. E. Eichler, "Recent segmental duplications in the human genome", *Science* **2002**, *297*, 1003–7.

[78] Gao, L.-z., H. Innan, "Very low gene duplication rate in the yeast genome", *Science* **2004**, *306*, 1367–70.

[79] Hahn, M. W., T. De Bie, J. E. Stajich, C. Nguyen, N. Cristianini, "Estimating the tempo and mode of gene family evolution from comparative genomic data", *Genome Res* **2005**, *15*, 1153–60.

[80] Liu, P. et al., "Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements", *Cell* **2011**, *146*, 889–903.

[81] Ames, R. M., D. Money, V. P. Ghatge, S. Whelan, S. C. Lovell, "Determining the evolutionary history of gene families", *Bioinformatics* **2012**, *28*, 48–55.

[82] Librado, P, F. G. Vieira, J Rozas, "BadiRate: estimating family turnover rates by likelihood-based methods", *Bioinformatics* **2012**, *28*, 279–81.

[83] Felsenstein, J., "Evolutionary trees from DNA sequences: a maximum likelihood approach", *J Mol Evol* **1981**, *17*, 368–76.

[84] Bailey, N. T. J., *The Elements of Stochastic Processes with Application to the Natural Sciences*, John Wiley & Sons, New York, **1964**.

[85] Hahn, M. W., J. P. Demuth, S. G. Han, "Accelerated rate of gene gain and loss in primates", *Genetics* **2007**, *177*, 1941–9.

[86] Han, M. V., G. W. C. Thomas, J Lugo-Martinez, M. W. Hahn, "Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3", *Mol Biol Evol* **2013**, *30*, 1987–97.

[87] Csűrös, M, I Miklós, "Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model", *Mol Biol Evol* **2009**, *26*, 2087–95.

[88] Carvalho, A. B., A. G. Clark, "Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes", *Genome Res* **2013**, *23*, 1894–1907.

[89] Cortez, D., R. Marin, D. Toledo-Flores, L. Froidevaux, A. Liechti, P. D. Waters, F. Grützner, H. Kaessmann, "Origins and functional evolution of Y chromosomes across mammals", *Nature* **2014**, *508*, 488–93.

[90] Veyrunes, F., P. D. Waters, P. Miethke, W. Rens, D. McMillan, A. E. Alsop, F. Grützner, J. E. Deakin, C. M. Whittington, K. Schatzkamer, et al., "Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes", *Genome Res* **2008**, *18*, 965–73.

[91] Wallis, M., P. Waters, M. Delbridge, P. Kirby, A. Pask, F Grützner, W Rens, M. Ferguson-Smith, J. A. Graves, "Sex determination in platypus and echidna: autosomal location of *SOX3* confirms the absence of *SRY* from monotremes", *Chromosome Res* **2007**, *15*, 949–59.

[92] Foster, J. W., J. A. Graves, "An *SRY*-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene", *PNAS* **1994**, *91*, 1927–31.

[93] Page, D. C., M. E. Harper, J Love, D Botstein, "Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution", *Nature* **1984**, *311*, 119–23.

[94] Skaletsky, H. et al., "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes", *Nature* **2003**, *423*, 825–37.

[95] Alföldi, J. E., PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, **2008**.

[96] Hughes, J. F., H. Skaletsky, T. Pyntikova, T. A. Graves, S. K. M. van Daalen, P. J. Minx, R. S. Fulton, S. D. McGrath, D. P. Locke, C. Friedman, B. J. Trask, E. R. Mardis, W. C. Warren, S. Repping, S. Rozen, R. K. Wilson, D. C. Page, "Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content", *Nature* **2010**, *463*, 536–9.

[97] Paria, N., T. Raudsepp, A. Pearks Wilkerson, P. O'Brien, M. A. Ferguson-Smith, C. C. Love, C. Arnold, P. Rakestraw, W. J. Murphy, B. P. Chowdhary, "A gene catalogue of the euchromatic male-specific region of the horse Y chromosome: comparison with human and other mammals", *PLOS One* **2011**, *6*, e21374.

[98] Li, G, B. W. Davis, T Raudsepp, A. J. Pearks Wilkerson, V. C. Mason, M Ferguson-Smith, P. C. O'Brien, P. D. Waters, W. J. Murphy, "Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution", *Genome Res* **2013**, *23*, 1486–95.

[99] Bellott, D. W. et al., "Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators", *Nature* **2014**, *508*, 494–9.

[100] Bhowmick, B. K., Y Satta, N Takahata, "The origin and evolution of human ampliconic gene families and ampliconic structure", *Genome Res* **2007**, *17*, 441–50.

[101] Larracuente, A. M., A. G. Clark, "Surprising differences in the variability of Y chromosomes in African and cosmopolitan populations of *Drosophila melanogaster*", *Genetics* **2013**, *193*, 201–14.

[102] Marais, G. A. B., P. R. A. Campos, I Gordo, "Can intra-Y gene conversion oppose the degeneration of the human Y chromosome?: a simulation study", *Genome Biol Evol* **2010**, *2*, 347–57.

[103] Fawcett, J. A., H. Innan, "Neutral and non-neutral evolution of duplicated genes with gene conversion", *Genes* **2011**, *2*, 191–209.

[104] Connallon, T, A. G. Clark, "Gene duplication, gene conversion and the evolution of the Y chromosome", *Genetics* **2010**, *186*, 277–86.

[105] Betrán, E., J. P. Demuth, A. Williford, "Why chromosome palindromes?", *Int J Evol Biol* **2012**, *2012*.

[106] Xue, Y., C. Tyler-Smith, "An exceptional gene: evolution of the *TSPY* gene family in humans and other great apes", *Genes* **2011**, *2*, 36–47.

[107] Mukherjee, A., G. Dass, J. M. G, M. Gohain, B. Brahma, T. K. Datta, S. De, "Absolute copy number differences of Y chromosomal genes between crossbred (*Bos taurus* x *Bos indicus*) and Indicine bulls", *J Anim Sci Biotechnol* **2013**, *4*.

[108] Yue, X.-P., C. Dechow, T.-C. Chang, J. M. Dejarnette, C. E. Marshall, C.-Z. Lei, W.-S. Liu, "Copy number variations of the extensively amplified Y-linked genes, *HSFY* and *ZNF280BY*, in cattle and their association with male reproductive traits in Holstein bulls", *BMC Genomics* **2014**, *15*, 113.

[109] Ellis, P. J. I., J. Bacon, N. A. Affara, “Association of *Sly* with sex-linked gene amplification during mouse evolution: a side effect of genomic conflict in spermatids?”, *Hum Mol Gen* **2011**, *20*, 3010–21.

[110] Schaller, F., A. M. Fernandes, C. Hodler, C. Münch, J. J. Pasantes, W. Rietschel, W. Schempp, “Y chromosomal variation tracks the evolution of mating systems in chimpanzee and bonobo”, *PLOS ONE* **2010**, *5*, e12482.

[111] Greve, G., E. Alechine, J. J. Pasantes, C. Hodler, W. Rietschel, T. J. Robinson, W. Schempp, “Y-chromosome variation in Hominids: intraspecific variation is limited to the polygamous chimpanzee”, *PLOS ONE* **2011**, *6*, e29311.

[112] Mano, S, H Innan, “The evolutionary rate of duplicated genes under concerted evolution”, *Genetics* **2008**, *180*, 493–505.

[113] Rozen, S., H. Skaletsky, J. D. Marszalek, P. J. Minx, H. S. Cordum, R. H. Waterston, R. K. Wilson, D. C. Page, “Abundant gene conversion between arms of palindromes in human and ape Y chromosomes”, *Nature* **2003**, *423*, 873–6.

[114] Charlesworth, B., “The organization and evolution of the human Y chromosome”, *Genome Biol* **2003**, *4*, 226.

[115] Lange, J., H. Skaletsky, S. K. M. van Daalen, S. L. Embry, C. M. Korver, L. G. Brown, R. D. Oates, S. Silber, S. Repping, D. C. Page, “Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes”, *Cell* **2009**, *138*, 855–69.

[116] Lange, J., M. J. Noordam, S. K. M. van Daalen, H. Skaletsky, B. A. Clark, M. V. Macville, D. C. Page, S. Repping, “Intrachromosomal homologous recombination between inverted amplicons on opposing Y-chromosome arms”, *Genomics* **2013**, 257–64.

[117] Foster, J. W., F. E. Brennan, G. K. Hampikian, P. N. Goodfellow, A. H. Sinclair, R Lovell-Badge, L Selwood, M. B. Renfree, D. W. Cooper, J. A. Graves, “Evolution of sex determination and the Y chromosome: *SRY*-related sequences in marsupials”, *Nature* **1992**, *359*, 531–3.

[118] Meredith, R. W. et al., “Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification”, *Science* **2011**, *334*, 521–4.

[119] Ross, M. T. et al., “The DNA sequence of the human X chromosome”, *Nature* **2005**, *434*, 325–37.

[120] Vallender, E. J., B. T. Lahn, “How mammalian sex chromosomes acquired their peculiar gene content”, *BioEssays* **2004**, *26*, 159–69.

[121] Charlesworth, B, D Charlesworth, “The degeneration of Y chromosomes”, *Phil Trans R Soc B* **2000**, *355*, 1563–72.

[122] Wilson, M. A., K. D. Makova, “Evolution and survival on eutherian sex chromosomes”, *PLOS Genetics* **2009**, *5*, e1000568.

[123] Hughes, J. F., H. Skaletsky, N. Koutseva, T. Pyntikova, D. C. Page, “Sex chromosome-to-autosome transposition events counter Y-chromosome gene loss in mammals”, *Genome Biol* **2015**, *16*, 104.

[124] Charlesworth, B., “Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation”, *Nat Rev Genet* **2009**, *10*, 195–205.

[125] Evans, B. J., B. Charlesworth, “The effect of nonindependent mate pairing on the effective population size”, *Genetics* **2013**, *193*, 545–56.

[126] Haldane, J. B., “The mutation rate of the gene for haemophilia, and its segregation ratios in males and females”, *Ann Eugen* **1947**, *13*, 262–71.

[127] Goetting-Minesky, M. P., K. D. Makova, “Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates”, *J Mol Evol* **2006**, *63*, 537–44.

[128] Makova, K. D., W. H. Li, “Strong male-driven evolution of DNA sequences in humans and apes”, *Nature* **2002**, *416*, 624–6.

[129] Backström, N, H Ceplitis, S Berlin, H Ellegren, “Gene conversion drives the evolution of *HINTW*, an ampliconic gene on the female-specific avian W chromosome”, *Mol Biol Evol* **2005**, *22*, 1992–9.

[130] Davis, J. K., P. J. Thomas, J. W. Thomas, “A W-linked palindrome and gene conversion in New World sparrows and blackbirds”, *Chromosome Res* **2010**, *18*, 543–53.

[131] Klein, H. L., T. D. Petes, “Intrachromosomal gene conversion in yeast”, *Nature* **1981**, *289*, 144–8.

[132] Jackson, J. A., G. R. Fink, “Gene conversion between duplicated genetic elements in yeast”, *Nature* **1981**, *292*, 306–11.

[133] Bosch, E, “Dynamics of a human interparalog gene conversion hotspot”, *Genome Res* **2004**, *14*, 835–44.

[134] Roach, J. C., G. Glusman, A. F. A. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, D. J. Galas, “Analysis of genetic inheritance in a family quartet by whole-genome sequencing”, *Science* **2010**, *328*, 636–9.

[135] Hallast, P., P. Balaresque, G. R. Bowden, S. Ballereau, M. A. Jobling, “Recombination dynamics of a human Y-chromosomal palindrome: rapid GC-biased gene conversion, multi-kilobase conversion tracts, and rare inversions”, *PLOS Genetics* **2013**, *9*, e1003666.

[136] Mann, A., M. Weiss, “Hominoid phylogeny and taxonomy: a consideration of the molecular and fossil evidence in an historical perspective”, *Mol Phylogenet Evol* **1996**, *5*, 169–81.

[137] Yu, Y.-H., Y.-W. Lin, J.-F. Yu, W. Schempp, P. H. Yen, “Evolution of the *DAZ* gene and the AZFc region on primate Y chromosomes”, *BMC Evol Biol* **2008**, *8*, 96.

[138] Jobling, M. A., “Copy number variation on the human Y chromosome”, *Cytogenet Genome Res* **2008**, *123*, 253–62.

[139] Vodicka, R., R. Vrtel, L. Dusek, A. R. Singh, K. Krizova, V. Svacinova, V. Horinova, J. Dostal, I. Oborna, J. Brezinova, A. Sobek, J. Santavy, “*TSPY* gene copy number as a potential new risk factor for male infertility”, *Reprod Biomed Online* **2007**, *14*, 579–87.

[140] Giachini, C, F Nuti, D. Turner, I Laface, Y Xue, F Daguin, G Forti, C Tyler-Smith, C Krausz, "*TSPY* 1 copy number variation influences spermatogenesis and shows differences among Y lineages", *J Clin Endocrinol Metab* **2009**, *94*, 4016–22.

[141] Krausz, C., C. Giachini, G. Forti, "*TSPY* and male fertility", *Genes* **2010**, *1*, 308–16.

[142] Nickkholgh, B., M. J. Noordam, S. E. Hovingh, A. M. M. van Pelt, F. van der Veen, S. Repping, "Y chromosome *TSPY* copy numbers and semen quality", *Fertil Steril* **2010**, *94*, 1744–7.

[143] Evans, B. J., L Pin, D. J. Melnick, S. I. Wright, "Sex-Linked inheritance in macaque monkeys: implications for effective population size and dispersal to Sulawesi", *Genetics* **2010**, *185*, 923–37.

[144] Tosi, A. J., J. C. Morales, D. J. Melnick, "Comparison of Y chromosome and mtDNA phylogenies leads to unique inferences of macaque evolutionary history", *Mol Phylogenet Evol* **2000**, *17*, 133–44.

[145] Tosi, A. J., T. R. Disotell, J. C. Morales, D. J. Melnick, "Cercopithecine Y-chromosome data provide a test of competing morphological evolutionary hypotheses", *Mol Phylogenet Evol* **2003**, *27*, 510–21.

[146] Hughes, J. F., H. Skaletsky, D. C. Page, "Sequencing of rhesus macaque Y chromosome clarifies origins and evolution of the *DAZ* (Deleted in AZoospermia) genes", *BioEssays* **2012**, *34*, 1035–44.

[147] Bhowmick, B. K., N Takahata, M Watanabe, Y Satta, "Comparative analysis of human masculinity", *Gen Mol Res* **2006**, *5*, 696–712.

[148] Goto, H., L. Peng, K. D. Makova, "Evolution of X-degenerate Y chromosome genes in greater apes: conservation of gene content in human and gorilla, but not chimpanzee", *J Mol Evol* **2009**, *68*, 134–44.

[149] Park, S. G., S. S. Choi, "Expression breadth and expression abundance behave differently in correlations with evolutionary rates", *BMC Evol Biol* **2010**, *10*, 241.

[150] Yang, J., A. I. Su, W.-H. Li, "Gene expression evolves faster in narrowly than in broadly expressed mammalian genes", *Mol Biol Evol* **2005**, *22*, 2113–8.

[151] Zhang, L., W.-H. Li, "Mammalian housekeeping genes evolve more slowly than tissue-specific genes", *Mol Biol Evol* **2004**, *21*, 236–9.

[152] Geraldes, A., T. Rambo, R. A. Wing, N. Ferrand, M. W. Nachman, "Extensive gene conversion drives the concerted evolution of paralogous copies of the *SRY* gene in European rabbits", *Mol Biol Evol* **2010**, *27*, 2437–40.

[153] Soh, Y. Q. S. et al., "Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes", *Cell* **2014**, *159*, 800–13.

[154] Murphy, W. J., A. J. P. Wilkerson, T. Raudsepp, R. Agarwala, A. A. Schäffer, R. Stanyon, B. P. Chowdhary, "Novel gene acquisition on carnivore Y chromosomes", *PLOS Genetics* **2006**, *2*, e43.

[155] Dorus, S., S. L. Gilbert, M. L. Forster, R. J. Barndt, B. T. Lahn, "The *CDY*-related gene family: coordinated evolution in copy number, expression profile and protein sequence", *Hum Mol Genet* **2003**, *12*, 1643–50.

[156] Katsura, Y., M. Iwase, Y. Satta, “Evolution of genomic structures on mammalian sex chromosomes”, *Curr Genomics* **2012**, *13*, 115–23.

[157] Andrés, O., T. Kellermann, F. López-Giráldez, J. Rozas, X. Domingo-Roura, M. Bosch, “*RPS4Y* gene family evolution in primates”, *BMC Evol Biol* **2008**, *8*, 142.

[158] Lopes, A. M., R. N. Miguel, C. A. Sargent, P. J. Ellis, A. Amorim, N. A. Affara, “The human *RPS4* paralogue on Yq11.223 encodes a structurally conserved ribosomal protein and is preferentially expressed during spermatogenesis”, *BMC Mol Biol* **2010**, *11*, 33.

[159] DeBie, T, N Cristianini, J. Demuth, M. Hahn, “CAFE: a computational tool for the study of gene family evolution”, *Bioinformatics* **2006**, *22*, 1269–71.

[160] Perry, G. H., R. Y. Tito, B. C. Verrelli, “The evolutionary history of human and chimpanzee Y-chromosome gene loss”, *Mol Biol Evol* **2006**, *24*, 853–9.

[161] Harcourt, A. H., P. H. Harvey, S. G. Larson, R. V. Short, “Testis weight, body weight and breeding system in primates”, *Nature* **1981**, *293*, 55–7.

[162] Matsumura, S, “Female reproductive cycles and the sexual behavior of Moor macaques (*Macaca maurus*) in their natural habitat, south Sulawesi, Indonesia”, *Primates* **1993**, *34*, 99–103.

[163] Enomoto, T, K Matsubayashi, M Nakano, Y Nagato, T. L. Yusuf, D Sajuthi, “A comparative study on histology of testes in *Macaca nemestrina*, *M. fascicularis* and *M. fuscata*”, *Anthropol Sci* **1997**, *105*, 99–116.

[164] Reed, C, T. G. OBrien, M. F. Kinnaird, “Male social behvaior and dominance hierarchy in the Sulawesi crested black macaque (*Macaca nigra*)”, *Int J Primatol* **1997**, *18*, 247–60.

[165] Schillaci, M. A., R. R. Stallmann, “Ontogeny and sexual dimorphism in booted macaques (*Macaca ochreata*)”, *J Zool* **2005**, *267*, 19–29.

[166] Yang, Y., T. C. Chang, H. Yasue, A. K. Bharti, E. F. Retzel, W. S. Liu, “*ZNF280BY* and *ZNF280AY*: autosome derived Y-chromosome gene families in *Bovidae*”, *BMC Genomics* **2011**, *12*, 13.

[167] Felsenstein, J, S Yokoyama, “The evolutionary advantage of recombination. II. Individual selection for recombination”, *Genetics* **1976**, *83*, 845–59.

[168] Eyre-Walker, A, “Evidence of selection on silent site base compositions in mammals: potential implications for the evolution of isochores and junk DNA”, *Genetics* **1999**, *152*, 675–83.

[169] Galtier, N, G Piganeau, D Mouchiroud, L Duret, “GC-content evolution in mammalian genomes: the biased gene conversion hypothesis”, *Genetics* **2001**, *159*, 907–11.

[170] Lartillot, N, “Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes”, *Mol Biol Evol* **2012**, *30*, 489–502.

[171] Romiguier, J, V Ranwez, E. J. P. Douzery, N Galtier, “Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome size”, *Genome Res* **2010**, *20*, 1001–9.

[172] Kirsch, S, C Munch, Z Jiang, Z Cheng, L Chen, C Batz, E. E. Eichler, W Schempp, "Evolutionary dynamics of segmental duplications from human Y-chromosomal euchromatin/heterochromatin transition regions", *Genome Res* **2008**, *18*, 1030–42.

[173] Maher, C. A., R. K. Wilson, "Chromothripsis and human disease: piecing together the shattering process", *Cell* **2012**, *148*, 29–32.

[174] Chippindale, A. K., W. R. Rice, "Y chromosome polymorphism is a strong determinant of male fitness in *Drosophila melanogaster*", *PNAS* **2001**, *98*, 5677–82.

[175] Lemos, B, L. O. Araripe, D. L. Hartl, "Polymorphic Y chromosomes harbor cryptic variation with manifold functional consequences", *Science* **2008**, *319*, 91–3.

[176] Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z Zhang, W Miller, D. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res* **1997**, *25*, 3389–402.

[177] Drummond, A. J., A. Rambaut, "BEAST: Bayesian evolutionary analysis by sampling trees", *BMC Evol Biol* **2007**, *7*, 214.

[178] McBrearty, S., N. G. Jablonski, "First fossil chimpanzee", *Nature* **2005**, *437*, 105–8.

[179] Perelman, P., W. E. Johnson, C. Roos, H. N. Seuánez, J. E. Horvath, M. A. M. Moreira, B. Kessing, J. Pontius, M. Roelke, Y. Rumpler, M. P. C. Schneider, A. Silva, S. J. O'brien, J. Pecon-Slattery, "A molecular phylogeny of living primates", *PLoS Genet* **2011**, *7*, e1001342.

[180] Scally, A. et al., "Insights into hominid evolution from the gorilla genome sequence", *Nature* **2012**, *483*, 169–75.

[181] Delson, E., "Evolutionary history of the Cercopithecidae", *Contrib Primatol* **1975**, *5*, 167–217.

[182] Delson, E., "The Macaques: Studies in Ecology, Behavior and Evolution" in, (Ed.: Lindburg, D. G.), Van Nostrand Reinhold, New York, NY, **1980**, Chapter Fossil macaques, phyletic relationships and a scenario of deployment, pp. 10–30.

[183] Bustin, S. A., V Benes, J. A. Garson, J Hellemans, J Huggett, M Kubista, R Mueller, T Nolan, M. W. Pfaffl, G. L. Shipley, J Vandesompele, C. T. Wittwer, "The MIQE guidelines: Minimum Information for publication of Quantitative real-time PCR Experiments", *Clin Chem* **2009**, *55*, 611–22.

[184] D'haene, B., J. Vandesompele, J. Hellemans, "Accurate and objective copy number profiling using real-time quantitative PCR", *Methods* **2010**, *50*, 262–70.

[185] Dittus, W. P. J., "Socioecology and Psychology of Primates" in, (Ed.: Tuttle, R. H.), Mouton & Co, Chicago, IL, **1975**, Chapter Population dynamics of the toque macaque, *Macaca sinica*, pp. 125–51.

[186] Lindburg, D. G., N. C. Harvey, "Evolution and Ecology of Macaque Societies" in, (Eds.: Fa, J. E., D. G. Lindburg), Cambridge University Press, New York, NY, **1996**, Chapter Reproductive biology of captive lion-tailed macaques, pp. 318–41.

[187] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 3.1.0, **2008-2015**.

[188] Paradis, E. et al., ape: Analyses of Phylogenetics and Evolution, 3.0-6, **2012**.

[189] Untergasser, A, I Cutcutache, T Koressaar, J Ye, B. C. Faircloth, M Remm, S. G. Rozen, "Primer3-new capabilities and interfaces", *Nucleic Acids Res* **2012**, *40*, e115.

[190] Evans, B. J., J. C. Morales, J. Supriatna, D. J. Melnick, "Origin of the Sulawesi macaques (Cercopithecidae: *Macaca*) as suggested by mitochondrial DNA phylogeny", *Biol J Linnean Soc* **1999**, *66*, 539–60.

[191] Evans, B. J., J Supriatna, D. J. Melnick, "Hybridization and population genetics of two macaque species in Sulawesi, Indonesia", *Evolution* **2001**, *55*, 1686–702.

[192] Evans, B. J., J. Supriatna, N. Andayani, D. J. Melnick, "Diversification of Sulawesi macaque monkeys: decoupled evolution of mitochondrial and autosomal DNA", *Evolution* **2003**, *57*, 1931–46.

[193] Maddison, D. R., W. P. Maddison, MacClade 4: Analysis of phylogeny and character evolution, Sinauer Associates Inc., Sunderland, MA, 4.08a, **2005**.

[194] Edgar, R. C., "MUSCLE: multiple sequence alignment with high accuracy and high throughput", *Nucleic Acids Res* **2004**, *32*, 1792–7.

[195] Darriba, D., G. L. Taboada, R. Doallo, D. Posada, "jModelTest 2: more models, new heuristics and parallel computing", *Nat Methods* **2012**, *9*, 772–2.

[196] Drummond, A. J., M. A. Suchard, D. Xie, A. Rambaut, "Bayesian phylogenetics with BEAUti and the BEAST 1.7", *Mol Biol Evol* **2012**, *29*, 1969–73.

[197] Hedges, S. B., J Dudley, S Kumar, "TimeTree: a public knowledge-base of divergence times among organisms", *Bioinformatics* **2006**, *22*, 2971–2.

[198] Rambaut, D, A. J. Drummond, Tracer: MCMC trace analysis tool, 1.5.0, **2003-2009**.

[199] Zuker, M, "Mfold web server for nucleic acid folding and hybridization prediction", *Nucleic Acids Res* **2003**, *31*, 3406–15.

[200] Ye, J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, T. L. Madden, "Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction", *BMC Bioinformatics* **2012**, *13*, 134.

[201] Fernandez-Jimenez, N., A. Castellanos-Rubio, L. Plaza-Izurieta, G. Gutierrez, I. Irastorza, L. Castaño, J. C. Vitoria, J. R. Bilbao, "Accuracy in copy number calling by qPCR and PRT: a matter of DNA", *PLOS ONE* **2011**, *6*, e28910.

[202] Bio-Rad Laboratories, CFX Manager Software, 3.0.

[203] Tuomi, J. M., F. Voorbraak, D. L. Jones, J. M. Ruijter, "Bias in the Cq value observed with hydrolysis probe based quantitative PCR can be corrected with the estimated PCR efficiency value", *Methods* **2010**, *50*, 313–22.

[204] Ruijter, J. M., M. W. Pfaffl, S. Zhao, A. N. Spiess, G. Boggy, J. Blom, R. G. Rutledge, D. Sisti, A. Lievens, K. De Preter, et al., "Evaluation of qPCR curve analysis methods for reliable biomarker discovery: bias, resolution, precision, and implications", *Methods* **2013**, *59*, 32–46.

[205] Wallis, M. C., P. D. Waters, J. A. M. Graves, "Sex determination in mammals–before and after the evolution of SRY", *Cell Mol Life Sci* **2008**, *65*, 3182–95.

[206] Lundrigan, B. L., P. K. Tucker, "Evidence for multiple functional copies of the male sex-determining locus, Sry, in African murine rodents", *J Mol Evol* **1997**, *45*, 60–5.

[207] Perne, A., X. Zhang, L. Lehmann, M. Groth, F. Stuber, M. Book, "Comparison of multiplex ligation-dependent probe amplification and real-time PCR accuracy for gene copy number quantification using the β-defensin locus", *BioTechniques* **2009**, *47*, 1023–8.

[208] Nuytten, H., I. Wlodarska, K. Nackaerts, S. Vermeire, J. Vermeesch, J.-J. Cassiman, H. Cuppens, "Accurate determination of copy number variations (CNVs): application to the α- and β-defensin CNVs", *J Immunol Methods* **2009**, *344*, 35–44.

[209] Aldhous, M. C., S Abu Bakar, N. J. Prescott, R Palla, K Soo, J. C. Mansfield, C. G. Mathew, J Satsangi, J. A. L. Armour, "Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease", *Hum Mol Gen* **2010**, *19*, 4930–8.

[210] Fode, P., C. Jespersgaard, R. J. Hardwick, H. Bogle, M. Theisen, D. Dodoo, M. Lenicek, L. Vitek, A. Vieira, J. Freitas, P. S. Andersen, E. J. Hollox, "Determination of beta-defensin genomic copy number in different populations: a comparison of three methods", *PLOS ONE* **2011**, *6*, e16768.

[211] Zhang, X., S. Müller, M. Möller, K. Huse, S. Taudien, M. Book, F. Stuber, M. Platzer, M. Groth, "8p23 beta-defensin copy number determination by single-locus pseudogene-based paralog ratio tests risk bias due to low-frequency sequence variations", *BMC Genomics* **2014**, *15*, 64.

[212] Goulet, V., C. Dutang, M. Maechler, D. Firth, M. Shapira, M. Stadelmann, expm: Matrix exponential, 0.99-1.1, **Feb. 2014**.

[213] Felsenstein, J., "Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters", *Syst Zool* **1973**, *22*, 240–9.

[214] Pagel, M., "Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters", *Proc R Soc B* **1994**, *255*, 37–45.

[215] Bolker, B., bbmle: Tools for general maximum likelihood estimation, 1.0.17, **May 2014**.

[216] Wagenmakers, E.-J., S. Farrell, "AIC model selection using Akaike weights", *Psychon Bull Rev* **2004**, *11*, 192–6.

[217] Kondrashov, F. A., A. S. Kondrashov, "Measurements of spontaneous rates of mutations in the recent past and the near future", *Phil Trans R Soc B* **2010**, *365*, 1169–76.

[218] Vitti, J. J., S. R. Grossman, P. C. Sabeti, "Detecting natural selection in genomic data", *Annu Rev Genet* **2013**, *47*, 97–120.

[219] Méndez-Lago, M., C. M. Bergman, B. de Pablos, A. Tracey, S. L. Whitehead, A. Villasante, "A large palindrome with interchromosomal gene duplications in the pericentromeric region of the *D. melanogaster* Y chromosome", *Mol Biol Evol* **2011**, *28*, 1967–71.

[220] Evans, B. J., J. Supriatna, N. Andayani, M. I. Setiadi, D. C. Cannatella, D. J. Melnick, “Monkeys and toads define areas of endemism on Sulawesi”, *Evolution* **2003**, *57*, 1436–43.