

THE CAUSES OF MUTATION AND SUBSTITUTION RATE  
VARIATION IN PRIMATES

Gregg William Cline Thomas

Submitted to the faculty of the University Graduate School  
in partial fulfillment of the requirements  
for the degree  
Doctor of Philosophy  
in the School of Informatics, Computing, and Engineering and  
in the Department of Biology,  
Indiana University  
July, 2019

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the  
Requirements for the degree of Doctor of Philosophy.

Doctoral Committee

---

Matthew W. Hahn, Ph.D.  
(Chairperson, Biology)

---

Haixu Tang, Ph.D.  
(Chairperson, Informatics)

---

Erik Ragsdale, Ph.D.

---

Yuzhen Ye, Ph.D.

---

Predrag Radivojac, Ph.D.

July 1, 2019

Copyright © 2019

Gregg William Cline Thomas

For Clara and our future together

## Acknowledgements

Science is a collaborative effort, and the making of a scientist even more-so. As an individual learning and working towards this degree there have been many people that I've interacted with through the years, each of which has helped distribute the load of stresses intertwined with it, either professionally or personally. I'm grateful to all these people. First and foremost is my advisor, Dr. Matthew Hahn. His support throughout this process has helped mold me from a curious student into a scientist with direction. His ideas are never-ending and his willingness to share and discuss them with his students fosters an environment in which knowledge is palpable. I feel lucky to have him as a mentor and will be forever grateful for all that he has taught and done for me these past years.

Next, I would like to thank my committee members. Due to the nature of my double major and other circumstances, I have had many people advise me at one point or another. I'd like to thank my dissertation committee members who have all been fully supportive of me: Dr. Haixu Tang, Dr. Yuzhen Ye, Dr. Erik Ragsdale, and Dr. Predrag (Pedja) Radivojac. Dr. Michael Lynch also served on my committee briefly before leaving IU, and I value his comments during my proposal defense. Finally, Dr. Elizabeth Housworth and Dr. Sriraam Natarajan both served on my advisory committee to administer my qualifying exams. I learned so much about phylogenetics and machine learning from them, now two of my favorite areas.

Interactions during a Ph.D. are often brief, as students graduate and move on and others join, leading to many important, yet fleeting, relationships. However, three colleagues of mine stand out during my time here: James Pease, Kymberleigh Pagel, and Richard Wang, have helped me a lot along the way. James was a member of the Hahn lab when I joined as a Master's student and he is one of the kindest and smartest people I know, both in and out of the lab, and

his influence on me in both areas can't be understated. When James moved during the last part of his dissertation, it was on my couch that he stayed on return trips to the lab, becoming my temporary roommate (or "Bloomington wife" as his wife would say!). These times of watching movies, ordering food, and sometimes even talking science are fond memories that I will always cherish. Kym was a student in Pedja's lab and started her Ph.D. at the same time I started my Master's, making her one of the only people that has been a constant during my time here. We had many classes together, during which we would spend many hours studying together. I will always be grateful for her helping me understand things that were way over my head. While she graduated last fall, we are now working on our first scientific collaboration and I look forward to continuing to work with her in the future as a friend and colleague. Richard is my desk-mate and over the past two years working together I've learned so much from him about how to be a good scientist. He is a trusted friend that I can literally turn to whenever I have a question. His impact can be directly seen in Chapters 2 and 4 of this dissertation, in which he played a big part, but his influence will be present in all my future work as well.

Of course, science is a team effort, and I wouldn't have completed many projects without the guidance and help from many professional collaborations. A substantial part of this dissertation is the result of work done with colleagues at Baylor University who are expert genome sequencers. Particularly, Jeff Rogers, Raveendran Muthuswamy, and Alan Harris provided the genome sequences of owl monkeys and macaques used in Chapters 2 and 4 and were also integral in the baboon assembly used in Chapter 3. I've also had many collaborators for projects outside of the scope of this dissertation. Elias Dohmen (University of Münster), Stephen Richards (UC Davis), Rob Waterhouse (University of Lausanne), Ariel Chipman (The Hebrew University of Jerusalem), Erich Bornberg-Bauer (University of Münster) all worked on

the i5K project with me. Each of these people continually took a big share of work and kept my interest sparked in such a long project. Andrew Foote (Bangor University) and Yoonsoo Hahn (Chung-Ang University) helped me get interested in studying convergent evolution and Wes Warren (University of Missouri) and Patrice Minx (Washington University in St. Louis, retired) involved me in many interesting genome projects. I'm also grateful to Alexandra Bentz (Indiana University), Stefan Probst (UC Berkeley), Jean-Luc De Lage (University of Paris-Saclay), Virginie Courtier-Orgogozo (Paris Diderot University), and Cheng Sun (Chinese Academy of Agricultural Sciences) for including me in interesting projects that have allowed me to expand my skillset and piqued my curiosity in new organisms.

Many others have impacted my studies, including current and former members of the Hahn lab including Dan Schrider, Jimmy Denton, Claudio Casola, Rodrigo Ramalho, Melissa Toups, Simo Zhang, Hussain Ather, Fabio Mendes, Arthi Puri, Geoffrey House, Jeff Adrion, Rafael Gurrero, Ben Rosenzweig, Ben Fulton, Jelena Nguyen, Dan Vanderpool, and Mark Hibbins. Each of these are people I've interacted with on a day-to-day basis for some extended period and have made the lab a great place to work. I've also had institutional support from many people in the Informatics and Biology departments. Linda Hostetter helped me navigate the paperwork and milestones of the early part of my Ph.D. and was always great to interact with. Her successor, Beverly Diekhoff, has likewise helped me in the latter part of my dissertation, guiding me towards my defense. Their biology counterparts, Gretchen Clearwater and Mary White, are also indispensable resources for help in that department. I also want to thank Spencer Hall, the director of the EEB graduate program. As director, he has tons of work and many people to coordinate, yet I always felt that he was invested and involved in any problems I went to him with. Finally, as a bioinformatician, it would be difficult to complete my work without the

support of all the IT people at IU, including Dave King and the rest of the biocomp crew, and all the people that work to maintain the high-performance systems at IU. My work wouldn't have been possible without them.

Outside the lab, I've had the fortune to have many friends during my time here, including Sneha Palliyil, Suyog Chandramouli, Swetha Murali, Vikas Pejavar, Jose Lugo-Martinez, Christie Debelius, Travis Sullivan, Wanda Savala, and Eric Rosenbaum. And of course, my longest friend, Anna Kopp, who has talked me through so many parts of my life.

Anyone who knows me knows I love animals. My pets, Jenny, Momo, and Pippin, bring me great joy and are three of my best friends. Matt's pets, Igloo and Wells, have also been great companions this past year as I've watched them while Matt is on sabbatical. I've also been lucky to work at the animal shelter on the weekends and interact with all the dogs there. This has been one of the biggest stress relievers throughout my dissertation, so I'd like to thank Jenny Gibson, the volunteer coordinator at the animal shelter, for always being welcoming, and all the other great volunteers that I could share my love of animals with. I hope I've helped the dogs as much as they've helped me.

I of course would be remiss if I didn't mention my family. My parents, Dona and Bill Thomas have supported me in the right ways for much of my life and I definitely wouldn't be here without them, so I'm ever grateful. My sister, LaRonika Thomas, and brother-in-law Nate Larson have also always had my back.

And finally, I want to thank my loving partner Clara Boothby. We met in 2015, halfway through my PhD. After our first date I knew we had something special and we got married almost exactly two years later in 2017. Her intelligence and wit keep me constantly on my toes and entertained, and our mutual love for science fiction and puzzles ensure we never get bored.

Importantly, she knows the amount of work involved in getting a Ph.D. and has always supported me when I needed to code for days on end or read papers for hours. One thing she has endured is listening to me give practice talks on repeat for weeks leading up to an actual talk, so much so that she could probably step in and give the talks for me. This routine has been one of the most supportive things anyone has ever done for me and has helped me improve my presentation skills greatly. Similarly, she is always willing to give feedback on my papers and her writing skills have helped improve my own. I have always felt sure of myself since I met her, and I will always be thankful to her for that and so much more.

Gregg William Cline Thomas

## THE CAUSES OF MUTATION AND SUBSTITUTION RATE VARIATION IN PRIMATES

All genetic variation originates as a mutation in the DNA sequence of a single individual. The rate at which mutations arise is a parameter of utmost importance both for human health and evolutionary studies. While it is known that mutation and substitution rates vary between species, whether this is due to natural selection or some other phenomena remains unclear. Recent studies have shown that in mammals the rate of new nucleotide mutations is dependent almost entirely on the age of the father. This is likely due to errors accruing during DNA replication during spermatogenesis in the male parent. Based on these observations, I have developed a model of the single nucleotide mutation rate that incorporates parental age into estimates of both the mutation rate and substitution rate. To test this model, I sequenced the genomes of several families of owl monkeys and macaques, primates closely related to humans. I found that, in primates, variation in nucleotide mutation rates can be explained almost entirely by variation in the generation time and puberty age of the species considered. I also show that, for larger structural variants, parental age likely plays no role in the rate of these mutations. This stands in contrast to the paternal age effect of single nucleotide mutations and is in accordance with the accepted mechanism of formation for structural variants. Finally, since genome sequencing is still error-prone, mutation and substitution rate estimates are likely conflated by false positives. To remedy this, I developed a method to assign an intuitive quality score to genome assemblies that takes into account underlying sequence and mapping quality. This method can be used to annotate a genome assembly and subsequently correct or filter out low quality positions, thus reducing the number of false positive variants found. This in turn will lead to more accurate estimates of the mutation rate and substitution rate in any species.

---

Matthew W. Hahn, Ph.D.  
(Chairperson, Biology)

---

Haixu Tang, Ph.D.  
(Chairperson, Informatics)

---

Erik Ragsdale, Ph.D.

---

Yuzhen Ye, Ph.D.

---

Predrag Radivojac, Ph.D.

## Table of Contents

CHAPTER 1: The human mutation rate is increasing, even as it slows .....	1
1.1 Introduction.....	1
1.2 Causes of mutation rate and substitution rate variation.....	2
1.3 Is there a generation-time effect in primates?.....	5
1.4 The human nucleotide mutation rate is decreasing, and increasing .....	6
1.5 Selection on somatic mutation rate as an explanation for the hominoid slowdown.....	7
1.6 Conclusions.....	8
1.7 Mutation rate model.....	8
CHAPTER 2: Reproductive longevity predicts mutation rates in primates.....	14
2.1 Results & Discussion.....	14
2.2 Experimental Model and Subject Details .....	22
2.3 Sequencing.....	22
2.4 Mapping and variant calling .....	23
2.5 Filtering of putative mutations.....	25
2.6 Phasing mutations.....	26
2.7 Estimating mutation rates .....	26
2.8 Modeling mutation rates .....	28
2.9 Estimating mutational parameters from humans .....	31
2.10 Quantification and Statistical Analysis.....	35
2.11 Data Availability.....	35
2.12 Additional Resources.....	35
CHAPTER 3: Referee: reference assembly quality scores .....	37
3.1 Introduction.....	37
3.2 Materials & Methods .....	37
3.3 Referee’s scoring system .....	39
3.4 Results.....	40
3.5 Conclusions.....	43
CHAPTER 4: Origins and long-term patterns of copy-number variation in rhesus macaques ...	44
4.1 Introduction.....	44
4.2 Patterns of copy-number variation in rhesus macaques .....	46
4.3 De novo copy-number variants.....	51
4.4 Gene duplications and losses within and between species .....	52
4.5 Discussion.....	55

4.6	Sequencing and read mapping .....	57
4.7	Calling copy-number variants (CNVs) in Rhesus macaques .....	57
4.8	Filtering putative macaque CNVs .....	58
4.9	Identifying de novo CNVs and calculating the mutation rate .....	59
4.10	Human CNV data .....	59
4.11	Counting fixed macaque gene duplications and losses.....	60
	References.....	61
	Curriculum vitae	

## **CHAPTER 1: The human mutation rate is increasing, even as it slows**

### **1.1 Introduction**

It is well documented that rates of nucleotide substitution vary between species [1-4]. By examining nucleotide changes at genomic positions that are not affected by natural selection, we can infer that this substitution rate variation is driven by differences in underlying mutation rates and not simply differences in the efficacy of selection between species (cf. [5]). Because the nucleotide mutation rate itself can be influenced by selection and drift [6-9], understanding the relative impact of different evolutionary forces in driving changes in the mutation rate is key to understanding variation in substitution rates.

Many traits have been found to co-vary with substitution rates, and based on these trait correlations many explanations for rate differences between species have been proposed. One of the most consistent relationships is between body size and substitution rate: larger organisms tend to have slower rates of molecular evolution [10, 11]. As there are many life history traits that are associated with body size, these are also often correlated with substitution rates. Some examples of such traits include metabolic rate [12, 13], longevity [14], population size [9], and generation time [15]. Although many of these life history traits are correlated with one another, analyses of large datasets have to some extent been able to disentangle the contribution of each to variation in substitution rates (e.g. [11, 16, 17]).

One of the most well-known examples of nucleotide substitution rate variation between species is known as the “Hominoid slowdown” [18]. The slowdown is based on the observation that the substitution rate in hominoids (Great Apes) is slower than that in Old World monkeys, which is again slower than that in New World monkeys (reviewed in [19]). Within hominoids, humans show the slowest rate of all [20], and this rate may be continuing to fall [21]. The

general trend of lower substitution rates associated with longer generation times in the hominoids led to the proposal that this slowdown was directly due to differences in the generation time [22]. Similar differences between rodents, artiodactyls, and primates have also been ascribed to the so-called “generation-time effect” [1, 23].

The generation-time effect hypothesis proposes that shorter generation times lead to higher substitution rates “because in any arbitrary unit of time short-generation organisms will go through more generations and therefore more rounds of germ-cell divisions” ([24], p. 229). The generation-time effect therefore assumes that there are a fixed number of germline cell divisions per generation (or at least a declining number with increased generation time), regardless of the length of a generation. However, germline cell divisions in male primates continue as an individual ages [25], and therefore older males have gametes with more mutations [26, 27]. Although it has long been known that the increased number of cell divisions with increased generation time will dampen any proposed generation-time effect (e.g. [1, 28], until recently no quantitative estimates of this relationship were available.

Here, we show how new data on the per-generation mutation rate in humans directly contradict the generation-time hypothesis as an explanation for the hominoid slowdown. In order to understand why such data are relevant to the generation-time effect, we first discuss different ways in which the mutation rate can evolve and the effects of each of these on the substitution rate.

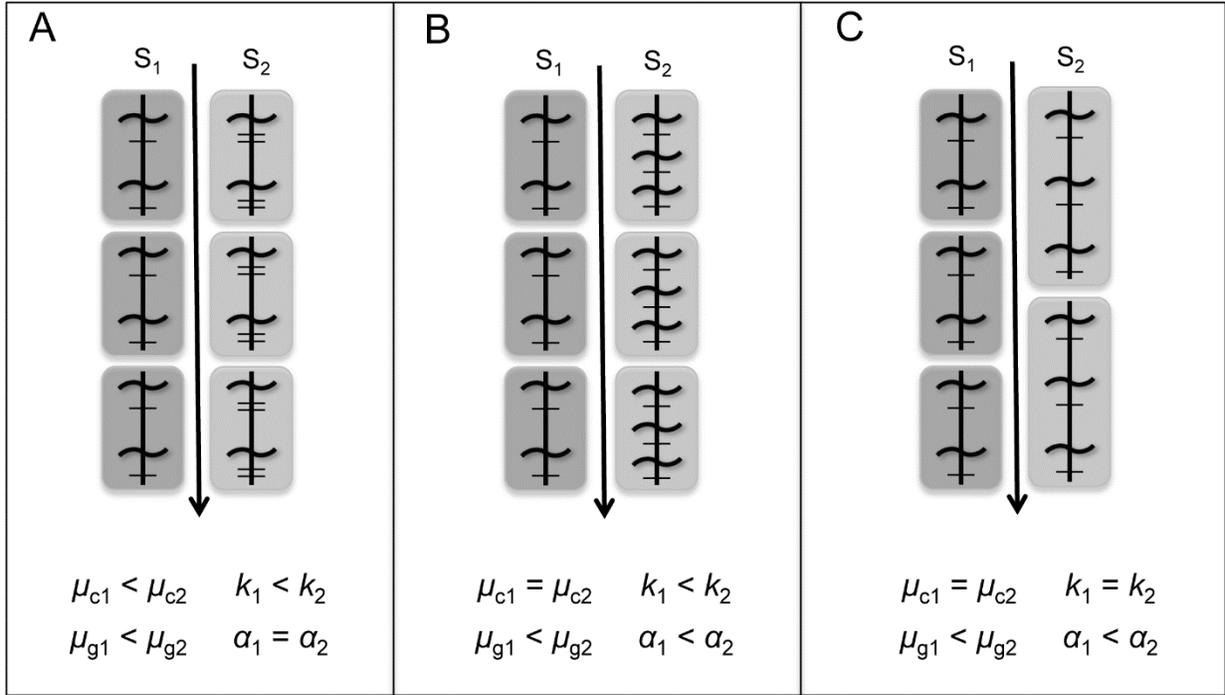
## **1.2 Causes of mutation rate and substitution rate variation**

The per-generation mutation rate ( $\mu_g$ ) is a fundamental parameter in evolutionary biology, relevant to almost every aspect of the genetics of populations. This key trait is determined by the combined effects of DNA damage, repair, replication, and associated

processes over the course of an individual's lifetime, and therefore can be affected by a change in any one of these underlying systems. Here we focus on three major mechanisms that can affect the per-generation mutation rate (Figure 1.1).

Because mutations arise via DNA replication error and/or failure to repair those errors, one possible mechanism for rate variation is for the DNA replication and repair machinery to be more or less efficient in a particular species (Figure 1.1a; [2]). Changes in either the amino acid sequence of replication-associated proteins or the number and identity of proteins involved in replication and repair can affect the per-cell-division mutation rate ( $\mu_c$ ). This rate is known to vary among species, with the human germline per-cell-division rate being more than ten times lower than the mouse rate [9]. Evolution of the per-cell-division rate affects both the per-generation mutation rate and the substitution rate between species ( $k$ ) as the number of mutations per unit time increases or decreases. Assuming that changes to the replication machinery affect males and females equally, evolution of  $\mu_c$  does not change the ratio of male-to-female mutations ( $\alpha$ ), which has a value greater than 1 in many species (e.g. [29, 30]).

A second way to change the per-generation mutation rate is to change the rate at which germline cells divide (Figure 1.1b). With more cell divisions come more replication events, which leads to more mutation. Evidence suggests that closely related species differing in the intensity of sperm competition differ in the number of male germline cell divisions, with more competition leading to higher per-generation mutation rates [31, 32]. Because the changing cell-division rate leads to more or fewer mutations per unit time, the substitution rate is changed as a consequence. For instance, mouse male stem cells divide every 8.6 days, while human male stem cells divide every 16 days [25]. If cell-division rates show equivalent change in males and females, then  $\alpha$  is not affected; however, changes biased to one sex will change the ratio of



**Figure 1.1:** Predictions about the per-cell-division ( $\mu_c$ ) and per-generation ( $\mu_g$ ) mutation rates, substitution rates ( $k$ ), and male-to-female mutation ratio ( $\alpha$ ) between two species ( $S_1$  and  $S_2$ ) by varying (A) the efficiency of the DNA repair machinery, (B) the number of replications per unit time, or (C) the generation time. Note that  $\alpha$  is only changed in panel B if the replication rate change occurs in males and not females. Each grey box represents one generation, while each wavy line indicates a germline replication event. Replications give rise to mutations, which are shown as notches. The arrow between the two species represents time.

male-to-female mutations, resulting in changes to  $\alpha$  [32].

Finally, the generation time itself can directly affect the per-generation mutation rate (Figure 1.1c). Assuming that germline cell divisions continue throughout an individual's lifetime, increasing the generation time increases the number of mutations that accumulate. Indeed, the large-scale association between per-generation mutation rates and generation time may be a consequence of the greater opportunity for errors given longer generations ([33], p. 86), as long as most mutations are derived from replication (which they seem to be; [34]). For species in which the number of mutations in offspring increases linearly with parental age, change in generation time should not affect the substitution rate (Figure 1.1c). This claim

assumes that the variation in generation time is occurring post-spermatogenesis; changes in the time to spermatogenesis between species could change the substitution rate if there are a fixed number of cell divisions that have to occur in this time. However, there does not appear to be a fixed number of cell divisions before spermatogenesis among mammals [25]. Any difference in male and female mutation rates due to differences in germline cell differentiation will be further magnified by longer generations, and  $\alpha$  is predicted to increase as a result.

### **1.3 Is there a generation-time effect in primates?**

Given the above considerations, it is worthwhile considering whether the conditions necessary for the generation-time effect hold in primates. The generation-time effect hypothesis states that substitution rates slow when there is both an increase in the generation time and a decrease in the germline replication rate [1, 20]. If a fixed number of germline cell divisions occur in each generation—as in female primates—then longer generations result in a lower average rate of cell division per unit time (Figure 1.2); as a consequence, substitution rates would indeed go down.

However, recent whole-genome data from humans show that the number of offspring mutations is a linear function of paternal age, and is not correlated with maternal age [35, 36]. The children of fathers age 20 have approximately 40 *de novo* nucleotide mutations, of fathers age 30 have 60 mutations, and so on [35]. Under these conditions longer generation times have no effect on either the cell division rate or the per-cell-division mutation rate, and therefore there is no effect on substitution rates (Figure 1.1c). This does not mean that there should be no correlation between increased generation-time and decreased substitution rates, only that an associated factor is the cause of such correlations (see below). In addition, the absence of a direct effect of generation time on substitution rates can help to explain why the rate of DNA

duplication can be *increasing* in hominoids [37, 38]. In this case it is the repair machinery itself that is evolving—possibly in different ways for nucleotide and duplication mutations—not a common life-history trait.

Given the predictions laid out in the previous section, in primates there should also be a positive correlation between generation time and  $\alpha$  because increased numbers of male germline cell divisions amplify differences between male and female mutation rates. Based on data from human, chimpanzee, gorilla, and orangutan,  $\alpha$  does in fact scale positively with generation time [32].

#### **1.4 The human nucleotide mutation rate is decreasing, and increasing**

In the absence of a generation-time effect, the observed decrease in hominoid substitution rates must be due to either a decrease in the per-cell-division mutation rate or a decrease in the germline cell division rate. The predictions laid out in Figure 1.1 show that decreased rates of cell division would lead to lower values of  $\alpha$ , which is contrary to the observed trends. Therefore, the data imply that there has been a decrease in the per-cell-division mutation rate ( $\mu_c$ ) in hominoids, and that this rate is further decreasing in humans.

On the other hand, because the per-generation mutation rate ( $\mu_g$ ) is determined by the accumulation of mutations across many germline cell divisions, consideration of recent demographic shifts in human populations suggests that  $\mu_g$  is actually increasing. In essence, the increased rate is simply a result of increases in the average human generation time, which is much longer now than it was in archaic humans [39, 40]. Even within the last 40 years, data from developed countries show an increasing average generation time for both females [41] males [42]. Taken together with the fact that mutation rates increase with paternal age, these increases in generation time result in higher per-generation mutation rates.

Experimental manipulation of the age at reproduction in mutation-accumulation experiments has shown that increased generation times result in increased  $\mu_g$  [43]. In particular, increased generation times lead to increased per-generation deleterious mutation rates, and increased variance in fitness among individuals. If similar increases in the variance in fitness among humans occur as a result of increases in  $\mu_g$ , such changes may have important consequences for understanding the ongoing evolution of human health (cf. [44, 45]).

### **1.5 Selection on somatic mutation rate as an explanation for the hominoid slowdown**

Without the generation-time effect as an explanation for the observed slowdown in nucleotide substitution rates, it behooves us to ask whether there are other viable hypotheses for this pattern. Non-adaptive hypotheses would seem to predict a higher rate of mutation in humans, as they have the lowest effective population size [9, 46]. These predictions run counter to the observed patterns, at least within primates.

Multiple adaptive hypotheses have been proposed for the negative association between body size and substitution rate, many of which are concerned with the increased somatic mutation load experienced by long-lived, large-bodied organisms [10, 14, 47, 48]. We hypothesize that the hominoid DNA repair machinery has evolved to be more efficient in response to selection on the somatic mutation rate, which has in turn led to a lower germline mutation rate; this hypothesis assumes that the same repair proteins are used in the germline and soma [49, 50]. Although mutations in somatic cells do not affect offspring fitness, they do affect the fitness of the individual in which they occur and can therefore be a target of selection [46, 51]. Because the number of somatic cell divisions experienced by an organism is affected by both longevity and body size—and is generally correlated with generation time—all of these measures may to some degree be associated with substitution rates.

## 1.6 Conclusions

Until recently, measuring substitution rates between species was the only way to assess mutation rates on a large scale. Next-generation sequencing technologies now allow for whole-genome sequencing of parent-offspring trios [35, 52] and mutation-accumulation lines [53-57]. These methods have enabled the collection of per-generation mutation rates for various organisms, and the hope is that we will be able to explain the observed differences in substitution rates in terms of these mutation rates. However, as shown here, understanding differences in substitution rates first requires that we understand what aspect of the mutational process to measure. The implications of  $\mu_g$  and  $\mu_c$  for long-term evolutionary rates can be distinct, and radically different conclusions may be reached (e.g. increasing or decreasing mutation rates) depending on the measure used.

## 1.7 Mutation rate model

The mutation rate per generation ( $\mu_g$ ) of any organism can be simply calculated by multiplying the mutation rate per germline cell division ( $\mu_c$ ) by the number of germline cell divisions per generation ( $d_g$ ):

$$\mu_g = \mu_c \cdot d_g \quad (\text{E1.1})$$

However, in dealing with organisms whose germlines go through different stages during the life cycle (such as mammals), a different mutation rate must be calculated for each stage. The mutation rates from each stage are then be averaged to determine the overall  $\mu_g$ .

For mammals, there are three life stages in which germ cells can potentially experience different mutation rates, based on either a unique  $d_g$  or  $\mu_c$  from that period. These stages are females per generation, males before puberty, and males after puberty. In both females and males

before puberty, a fixed number of germ cell divisions occur (although this fixed number is likely not the same between the genders) and these cells use mitosis to replicate their DNA. In males after puberty, the number of germ cell divisions is continuous and is thought to relate linearly with generation time. Male germ cells after puberty also replicate their DNA with meiosis. These stages lead us to consider three separate mutation rates when determining the overall  $\mu_g$  for any organism of interest: the mutation rate of females per generation ( $\mu_{gF}$ ), the mutation rate in males before puberty per generation ( $\mu_{gMBP}$ ), and the mutation rate in males after puberty per generation ( $\mu_{gMAP}$ ). Additionally, separate mutation rates per cell division may be considered for mitosis ( $\mu_{cMIT}$ ) and meiosis ( $\mu_{cMEI}$ ). We assume these to be equal, but include both terms in our model.

The calculation of  $\mu_{gF}$  and  $\mu_{gMBP}$  give constant terms based on the number of cell divisions in each stage:

$$\mu_{gF} = \mu_{cMIT} \cdot d_{gF} \quad (\text{E1.2})$$

$$\mu_{gMBP} = \mu_{cMIT} \cdot d_{gMBP} \quad (\text{E1.3})$$

Then, given that the number of cell divisions in males per generation after puberty ( $d_{gMAP}$ ) is a linear relationship between the number of male cell divisions per year after puberty ( $d_{yMAP}$ ) and generation time ( $GT$ ) after the age of puberty ( $AP$ ):

$$d_{gMAP} = d_{yMAP} \cdot (GT - AP) \quad (\text{E1.4})$$

$\mu_{gMAP}$  is calculated as:

$$\mu_{gMAP} = \mu_{cMEI} \cdot d_{gMAP} \quad (\text{E1.5})$$

The two terms for before and after puberty mutation rates in males can be averaged to give the overall male mutation rate per generation ( $\mu_{gM}$ ):

$$\mu_{gM} = \frac{(\mu_{gMBP} + \mu_{gMAP})}{2} \quad (\text{E1.6})$$

It then follows that the overall per-generation mutation rate ( $\mu_g$ ) of an organism is the average between the male and female contributions:

$$\mu_g = \frac{(\mu_{gM} + \mu_{gF})}{2} \quad (\text{E1.7})$$

Care must be taken when converting from  $\mu_g$  to mutation rate per year ( $\mu_y$ ). Because each of the terms that contribute to  $\mu_g$  occurs over a different period of absolute time, they must each be converted to mutation rates per year based on the amount of time they encompass, with the total male mutation rate per year ( $\mu_{yM}$ ) being the average of the per generation rates in the two male life stages:

$$\mu_{yF} = \frac{\mu_{gF}}{GT} \quad (\text{E1.8})$$

$$\mu_{yMBP} = \frac{\mu_{gMBP}}{AP} \quad (\text{E1.9})$$

$$\mu_{yMAP} = \frac{\mu_{gMAP}}{(GT - AP)} \quad (\text{E1.10})$$

$$\mu_{yM} = \frac{(\mu_{yMBP} + \mu_{yMAP})}{2} \quad (\text{E1.11})$$

Now the per-generation mutation rate can easily be converted to the per-year mutation rate by again averaging the male and female contributions:

$$\mu_y = \frac{(\mu_{yM} + \mu_{yF})}{2} \quad (\text{E1.12})$$

The substitution rate ( $k$ ) is assumed to be equal to  $\mu_y$ :

$$k = \mu_y \quad (\text{E1.13})$$

Finally, to calculate the male-to-female mutation ratio ( $\alpha$ ):

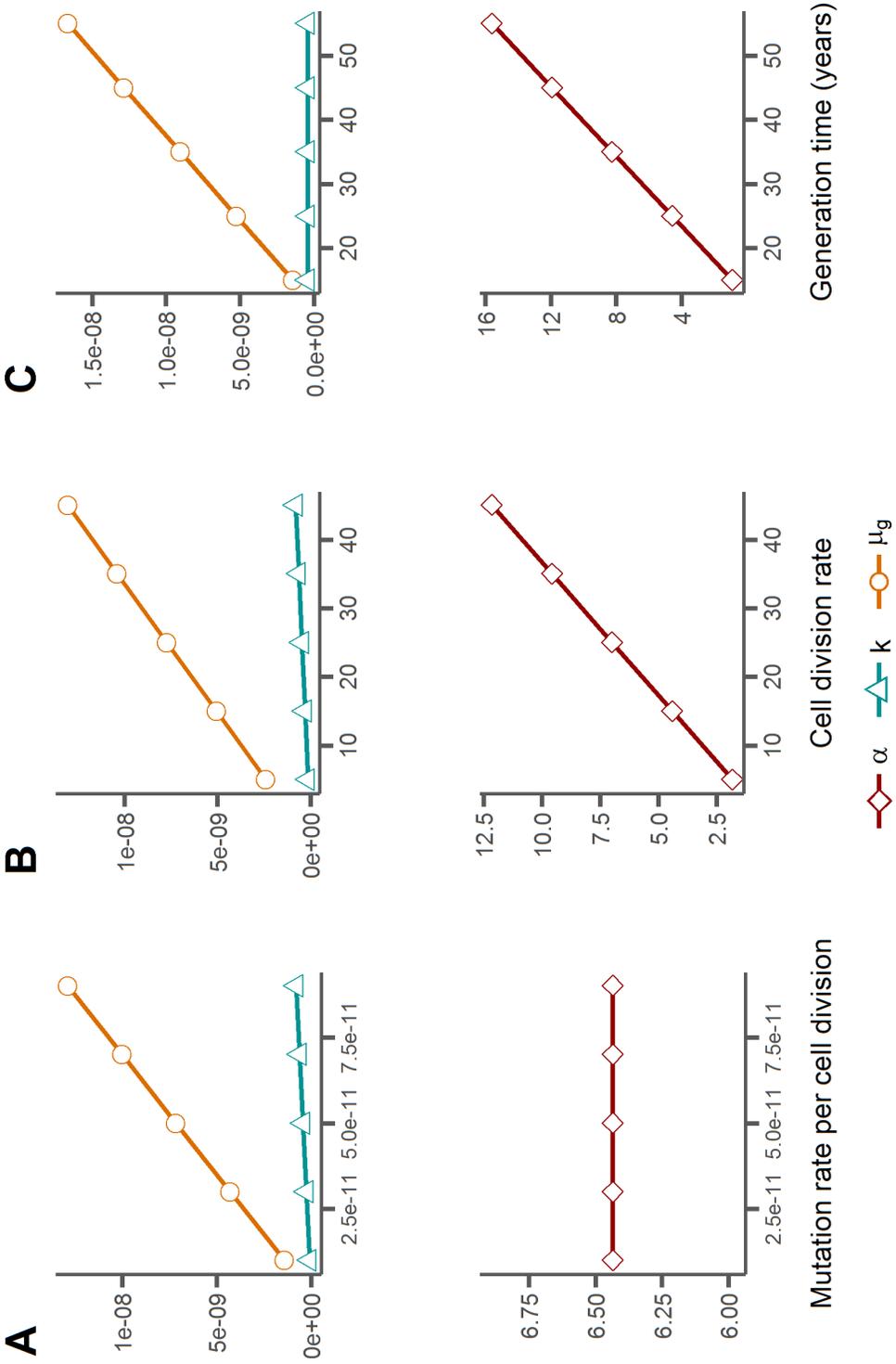
$$\alpha = \frac{\mu_{gM}}{\mu_{gF}} \quad (\text{E1.14})$$

**Table 1.1:** Life history and mutation rate parameters taken from Drost & Lee 1995 and used in conjunction with equations E1.7, E1.13, and E1.14 in Figures 1.2 and 1.3.

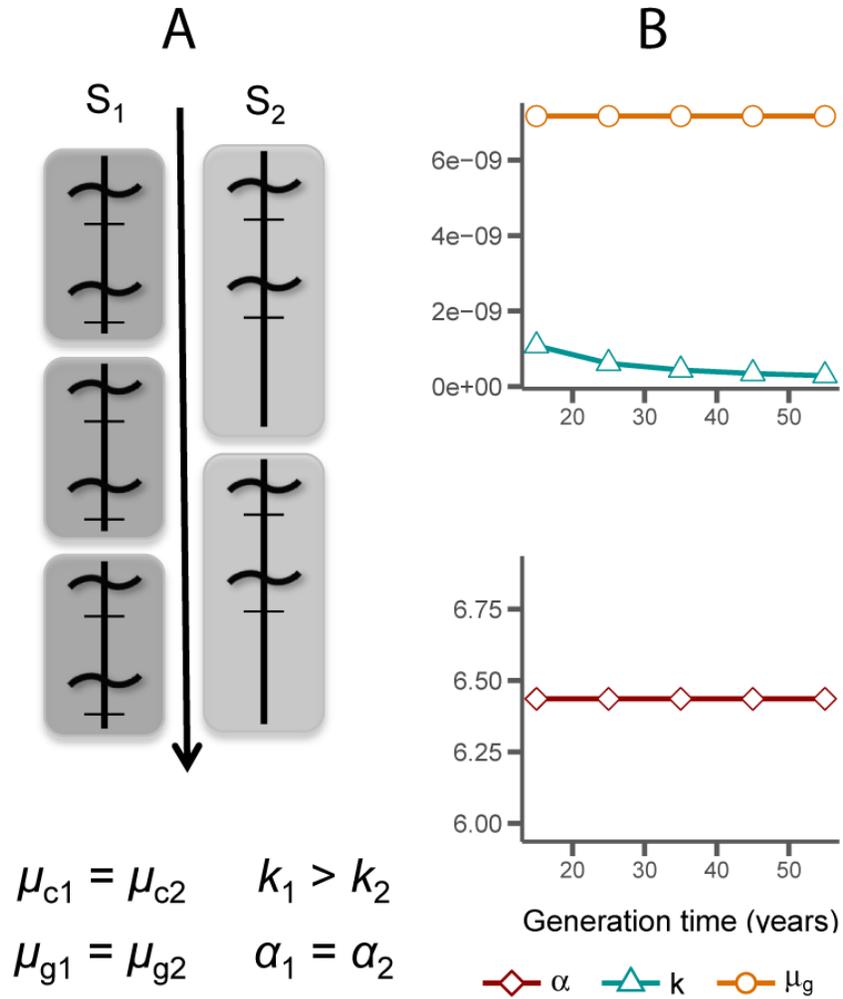
Generation Time	$\mu_c$	Age of puberty	$d_{gF}$	$d_{gMBP}$	$d_{yMAP}$
30 years	$2.3 \times 10^{-8}$	14 years	31	34	23
(20, 25, 30, 35, 40) <sup>a</sup>	(0.3, 1.2, 2.3, 4.3, 5.3) <sup>b</sup>				(5, 15, 25, 35, 45)

<sup>a</sup> The range of values used when a particular parameter was variable in Figure 1.2 are shown in parentheses

<sup>b</sup> All values for  $\mu_c$  are  $\times 10^{-8}$



**Figure 1.2:** The effect of changing the per cell division mutation rate (A), cell division rate per generation in males after puberty (B), or generation time (C) on the per generation mutation rate (Equation 1.7), substitution rate (Equation 1.13), and male-to-female mutation ratio (Equation 1.14). All parameters were taken from Drost & Lee 1995 (Table 1.1).



**Figure 1.3:** A demonstration of a “generation-time effect” in which generation time increases, but the rate of cell division does not. The left panel (A) demonstrates graphically how this occurs, while the right panel (B) uses values taken from Table 1.1 to calculate our model under this scenario.

## CHAPTER 2: Reproductive longevity predicts mutation rates in primates

### 2.1 Results & Discussion

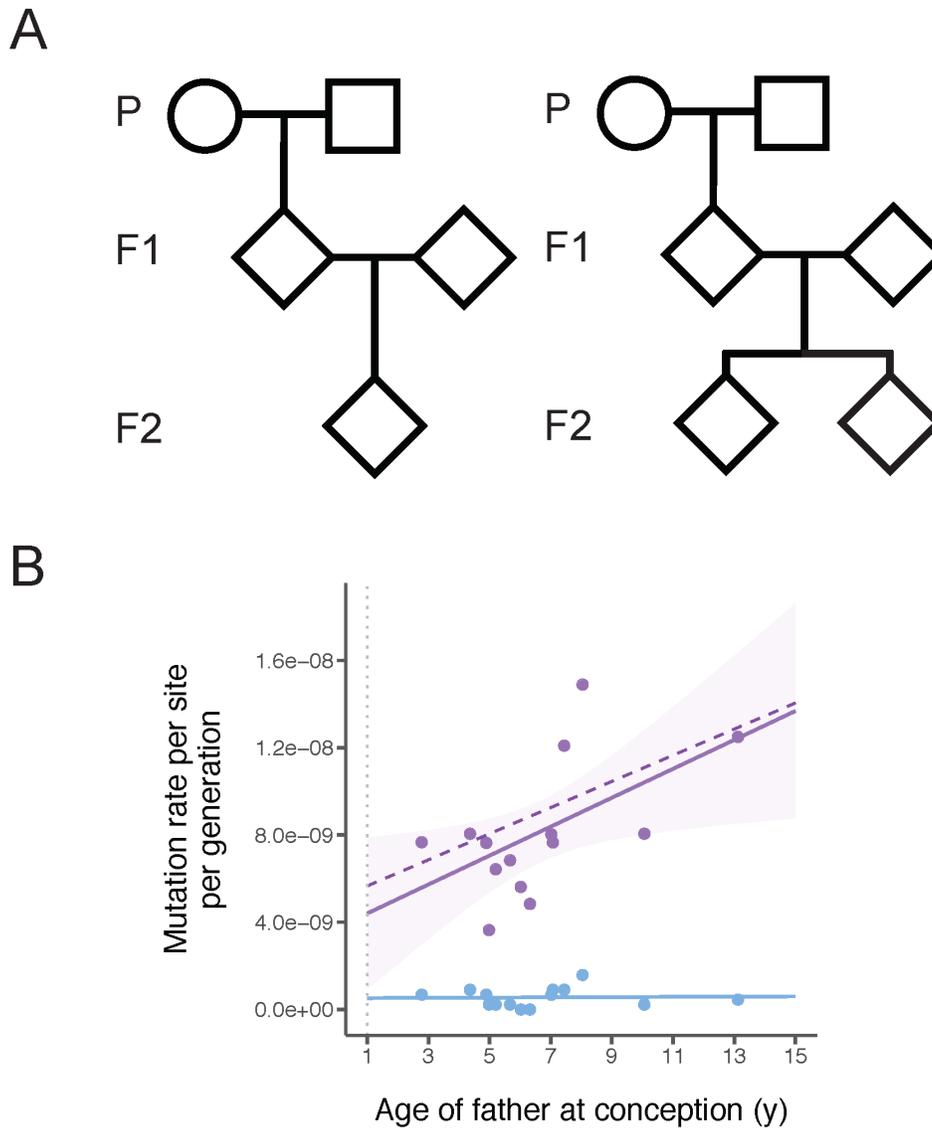
The rate at which new mutations arise is a key parameter of life on Earth, contributing to both individual disease risk and the evolution of novel traits. The mutation rate per generation varies among taxa, from as low as  $1 \times 10^{-10}$  per base in Archaea to more than  $1 \times 10^{-8}$  in mammals [58]. Two classes of models have been proposed to explain this variation. In one, the physiological and biochemical costs of increased fidelity during DNA replication limit the minimum mutation rate achievable [59, 60]. Selection for faster replication in smaller organisms constrains the accuracy with which the cellular machinery can copy DNA, resulting in an inverse relationship between body size and mutation rate. Alternatively, a population-genetic model invokes the limits to natural selection in organisms with smaller population sizes [33, 46, 61]. This model posits a higher rate of mutation in larger organisms because of their generally smaller population size [62].

One difficulty in teasing apart the forces driving the evolution of the mutation rate among multicellular organisms is the fact that lifespan varies as much as the per-generation mutation rate. In multicellular organisms, the number of mutations passed on to offspring in a single generation is a combination of the errors made in each round of germline replication and the accumulation of unrepaired DNA damage. One hundred years after the first observation of increased disease incidence in the children of older parents [63, 64], whole-genome sequencing in humans revealed the precise contribution of parental age to the number of *de novo* mutations in their offspring [35, 65-71]. In particular, the number of mutations passed on to the next generation is largely dependent on the age of the father [35], though there is a non-negligible contribution from the age of the mother [66-69, 71]. This is a consequence of the fact that after a

set number of germline mitoses during development in both males and females, the male germline resumes cell division at puberty [72, 73]. A similar effect of paternal age has been found in chimpanzees [74], suggesting that the age of reproduction may generally be an important determinant of the per-generation mutation rate.

Studying closely related primates offers a unique opportunity to examine the role that life history traits—such as age of puberty and average generation time—may play in determining mutation rates. We sequenced the genomes of 30 owl monkeys (*Aotus nancymaae*) within 6 multi-generation pedigrees (Figure 2.1A; Data S1A) in order to estimate the effect of parental age on the mutation rate. Owl monkeys reach sexual maturity at ~1 year of age [75] and can live up to 20 years in captivity [76]. Our sample includes individuals conceived by sires ranging from 3-13 years old and dams ranging from 3-12 years old, with an average age of 6.64 and 6.53 for sires and dams, respectively (Data S1A). These ages are comparable to those observed in the wild, as owl monkeys are solitary for some time before joining a mating group at around age four [77]. The genomes of all parents and offspring were sequenced to an average of 37X coverage (range: 35X-38X) using paired-end Illumina reads. Sequencing multi-generation pedigrees allows us to determine whether *de novo* mutations arose in either sires or dams, as well as to validate mutations transmitted to the next generation.

We observe 283 *de novo* mutations across 14 trios (Data S1B) and estimate an average mutation rate for owl monkeys of  $0.81 \times 10^{-8}$  per site per generation (Data S1C). In addition to stringent quality filters (see Methods), the average transmission frequency of *de novo* mutations passed from F1 individuals to F2 individuals across families was 0.502, giving us high confidence in our final set of mutations. As in humans, we find a strong association between paternal age and the number of *de novo* mutations (Figure 2.1B), with 2.92 additional mutations

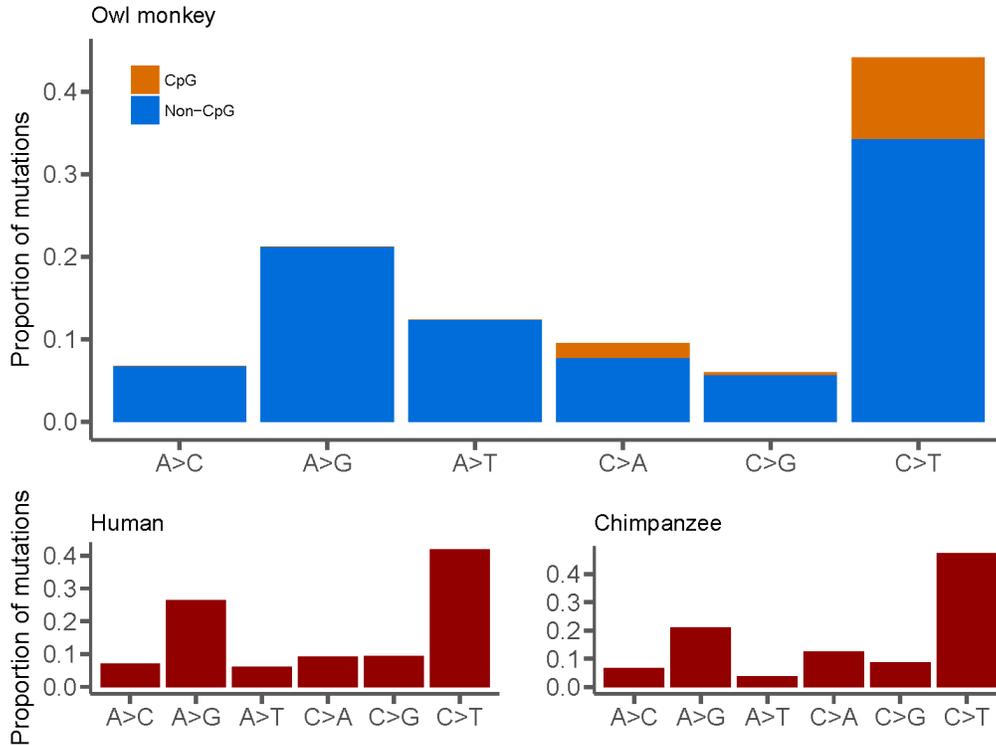


**Figure 2.1:** We used six multi-generation pedigrees in these two formats. Four families have a single F2 offspring (left) while two families have two F2 offspring (right). In total, 14 independent trios can be constructed from these pedigrees. B, Mutation rate estimates from the 14 owl monkey trios (purple points). A simple linear regression has been fit to these points (solid purple line) to show that the number of mutations increases with the father's age. Our model of reproductive longevity (dashed purple line) is not significantly different from the fit of the linear regression. The rate of non-replicative mutations, such as those that occur at CpG sites (blue dots), are not correlated with reproductive longevity (blue line). The dotted vertical grey line indicates expected age of puberty. See also Data S1, Figure 2.4.

accumulating per year ( $R^2=0.25$ , d.f.=12,  $P=0.040$ ). Also as expected, we find no effect of age on CpG mutations (Figure 2.1B, blue points and line), as these are not associated with replication errors. We were able to assign phase to 105 of the 283 *de novo* mutations via transmission to the third generation in our pedigrees (Data S1B). We find that 71 of these 105 phased mutations are paternal, with the number of mutations passed on increasing with the age of the father ( $R^2=0.58$ , d.f.=4,  $P=0.048$ ). We did not find an increasing number of mutations with maternal age ( $R^2=0.07$ , d.f.=4,  $P=0.307$ ) or age of the offspring ( $R^2=-0.02$ , d.f.=12,  $P=0.388$ ). This is the first direct observation of the paternal age effect outside of apes.

Inspection of the types of mutations found in the genomes of owl monkeys shows a transition:transversion (Ts:Tv) ratio of 1.97. This is in close agreement with the observed human Ts:Tv ratio of 2.10 [35]. In fact, the overall mutational spectrum between humans, chimpanzees, and owl monkeys appears almost identical, with the only difference being a slightly higher proportion of A→T mutations in owl monkeys (Figure 2.2). We also observe that 12.0% of mutations in owl monkeys occur at CpG sites, with CpG sites having a much higher Ts:Tv ratio (4.67), similar to observations in humans [35, 68]. Multinucleotide mutations (MNMs) are mutations that occur in close proximity to one another (<20 bp apart), likely caused by a single mutational event [78]. Here, we find 6 MNMs consisting of two mutations each, indicating that 2.1% of *de novo* mutations in owl monkeys are the result of MNMs (Data S1B). This fraction is also in agreement with that observed within humans [68, 78].

The mutation rate we observe in owl monkeys is 32.5% lower than the average human estimate of  $1.2 \times 10^{-8}$  mutations per site per generation [35, 79]. While traditional models of mutation rate evolution invoke changes to the underlying replication machinery as the main cause of such differences, we asked whether a shift in reproductive timing could explain the



**Figure 2.2: Comparison of mutational spectra from owl monkeys, humans, and chimpanzees.** There is a slight but significant difference in the frequency of A→T mutations between owl monkeys and humans ( $\chi^2 = 25.7$ , d.f. = 4,  $P < 0.05$ ), but otherwise no difference between mutational spectra for these three species. Human data were averaged across four studies (see Data S1D for references) and chimpanzee data was extrapolated from Figure 3A in Venn et al. Mutation categories include their reverse complement. See also Data S1B.

lower rate in owl monkeys. The effects of paternal age on per-generation mutation rates have previously been modeled by combining estimates of the rate of mutation from different life stages [73, 79, 80]. The germline in males and females undergo a fixed number of divisions before birth, but the male germline continues dividing upon reaching sexual maturity. This phenomenon suggests that the length of time between puberty and the conception of offspring in an individual—which we define here as the *reproductive longevity* of males—plays a key role in determining the number of mutations passed on to the next generation. While paternal age is sufficient for predicting mutation rates within a species [35, 65-71], the concept of reproductive longevity makes it possible to predict mutation rates between species with varying ages of

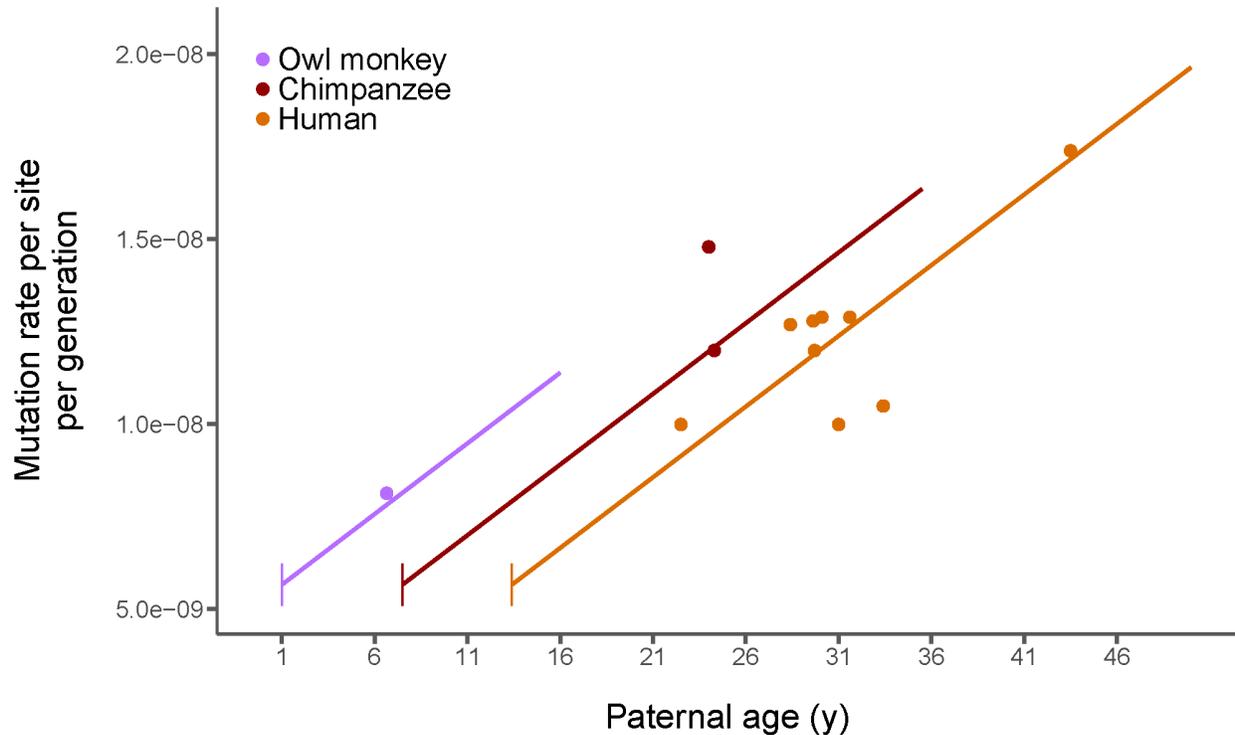
puberty. We modeled the owl monkey mutation rate as a linear combination of the mutations accumulated as a result of a constant number of germline divisions *in utero* and those accumulated during continued germline divisions post-puberty. The rate of mutation in these two stages were estimated from human studies, while sexual maturity was set at 1 year of age (see Section 2.7).

Our minimal model provides an excellent fit to the observed owl monkey data (Figure 2.1B, dashed line). In fact, a linear regression of the observed number of mutations with paternal age at conception is not significantly better than the predictions provided by our model ( $F=0.996$ , d.f.=13,  $P=0.994$ ). The main determinant of the mutation rate is reproductive longevity in sires, which determines the number of mitotic germline divisions before spermatogenesis. For instance, a 13-year-old owl monkey male (who reached sexual maturity at 1) will have the same reproductive longevity as a 25-year-old human male (who reached sexual maturity at 13). Our model therefore predicts the same estimated mutation rate if *de novo* mutations are sampled from offspring of these individuals, and this is what is observed (Figure 2.1B). Because reproductive longevity reflects replicative mutations, we observe no effect of father's age on non-replicative mutations, such as those found at CpG sites (Figure 2.1B, blue).

Given the fit of our model to owl monkey data, we calculated the expected mutation rates as a function of age for other primates, accounting for changes in the time to sexual maturity in each species. A model of reproductive longevity provides a good fit to the data from primate species for which direct mutation rate estimates are available (Figure 2.3; Data S1D). Our model explains why chimpanzees and humans have very similar per-generation mutation rates despite differences in average generation time: the earlier time to sexual maturity in chimpanzees causes reproductive longevity to be the same in both species. The model also accurately predicts

estimated mutation rates reported from various studies in humans where sampled parents were of different average age (Figure 2.3). Much of the variation in reported mutation rates in human studies is due to differences in the average reproductive longevity of sampled individuals ( $R^2=0.54$ , d.f.=7,  $P=0.01$ ). Variation in the age of reproduction across pedigrees will affect inferences regarding genetic variation in the mutation rate, as consistent differences in these ages may incorrectly be interpreted as heritable differences in this trait.

The association between mutation rates and reproductive longevity implies that changes in life history traits rather than changes to the mutational machinery are responsible for the evolution of these rates. Species that have evolved greater reproductive longevity will have a higher mutation rate per generation without any underlying change to the replication, repair, or proofreading proteins. The similarities between the mutational spectra of humans, chimpanzees, and owl monkeys (Figure 2.2) are further evidence that the molecular mechanisms responsible for mutation have not changed between these species. Many differences in the details of germline cell division may exist between these primates, but these differences do not appear to affect either pre-birth or post-puberty mutation accumulation. For instance, varying levels of sexual selection between species in the form of sperm competition leads to variation testis and ejaculate size [81]. This sort of variation likely also affects mutation rates through changing the germline replication rate [82], which can be accommodated in our model (see Section 2.7). The underlying consistency of mutation rates must also be reconciled with variation in the long-term substitution rate among primates [19, 79, 80, 83], as mutation rates are mechanistically tied to substitution rates (see Section 2.7 and Figure 2.5). Nevertheless, the close fit between the observed and expected mutation rates suggests that reproductive longevity is the major determinant of variation in mutation rates.



**Figure 2.3: A model of reproductive longevity fits estimated primate mutation rates.** Humans, chimpanzees, and owl monkeys are the only primates that currently have high-quality estimates of mutation rates via pedigree sequencing. Here we plot the average rate from each published study (points; see Data S1D for references). Predictions from our model of reproductive longevity (equations 3-8 in Section 2.7) using human mutational parameters—varying only life history traits—are also shown (lines). Vertical line segments represent the age of puberty for each species. See also Data S1C, DataS1D, and Figure 2.5.

Studies of mutation rate evolution will continue to accumulate across the tree of life as sequencing costs continue to plummet. In order to understand the forces affecting this important evolutionary parameter, future studies must recognize that the mutation rate is a function-valued trait: it is a function of reproductive longevity and other life history traits. Evidence from other species—for instance, arthropods [43] and long-lived plants [84]—suggests that reproductive longevity affects the mutation rate in many taxa, though the details of germline cell division will differ among lineages. If such a pattern holds widely in multicellular organisms, the effect of variation in life history traits should provoke a reexamination of the causes underlying the

correlation between body size and the per-generation mutation rate. At the very least, the null model for changes in the per-generation mutation rate must include reproductive longevity.

## **2.2 Experimental Model and Subject Details**

Thirty owl monkeys (*Aotus nancymae*) were selected for genome sequencing from the Owl Monkey Breeding and Research Resource at the Keeling Center based on available pedigrees, aiming for a spread of parental ages (Data S1A). Blood samples were taken from the femoral vein of unanesthetized animals. The animals were manually restrained in a supine position with one care staff holding the animal while another takes the sample, under approved IACUC protocols.

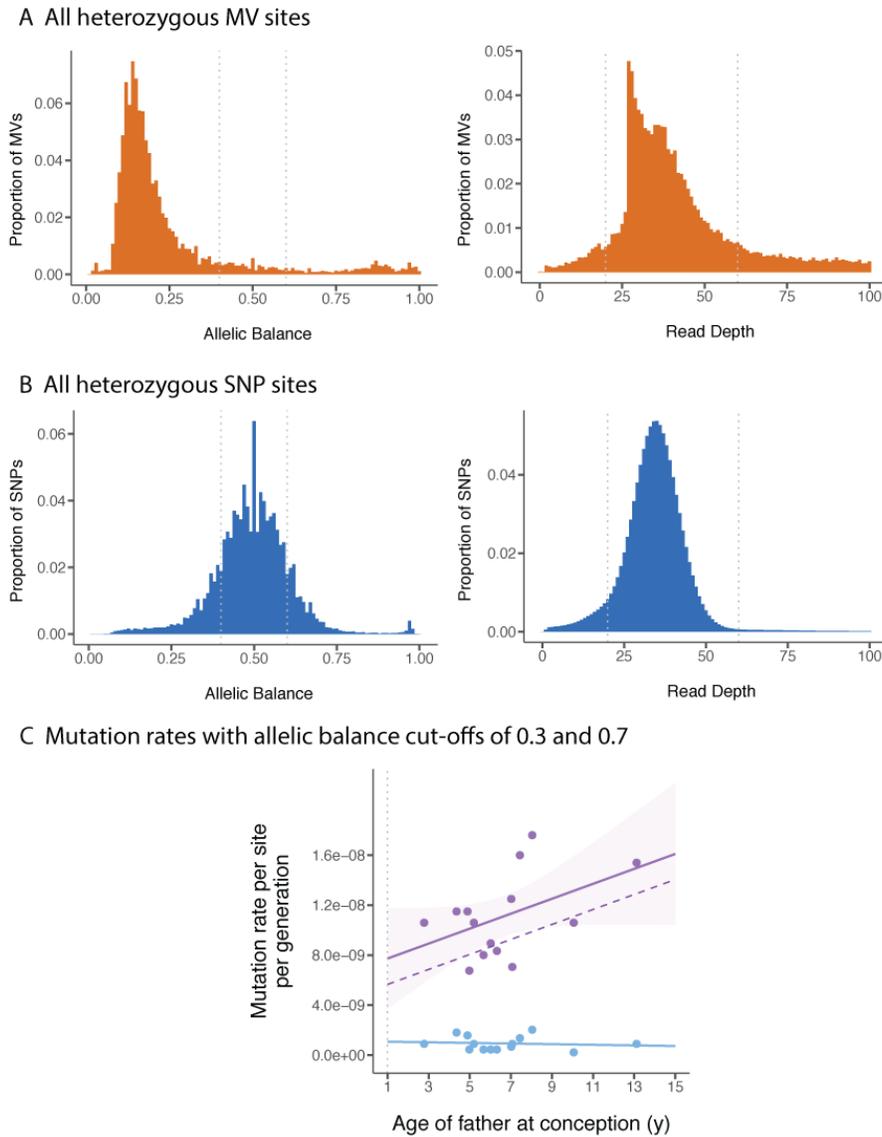
## **2.3 Sequencing**

Genomic DNA isolated from the blood samples was used to perform whole genome sequencing. We generated standard PCR-free Illumina paired-end sequencing libraries. Libraries were prepared using KAPA Hyper PCR-free library reagents (KK8505, KAPA Biosystems Inc.) in Beckman robotic workstations (Biomek FX and FXp models). We sheared total genomic DNA (500 ng) into fragments of approximately 200-600 bp in a Covaris E220 system (96 well format) followed by purification of the fragmented DNA using AMPure XP beads. A double size selection step was employed, with different ratios of AMPure XP beads, to select a narrow size band of sheared DNA molecules for library preparation. DNA end-repair and 3'-adenylation were then performed in the same reaction followed by ligation of the barcoded adaptors to create PCR-Free libraries, and the library run on the Fragment Analyzer (Advanced Analytical Technologies, Inc., Ames, Iowa) to assess library size and presence of remaining adapter dimers. This was followed by qPCR assay using KAPA Library Quantification Kit using their SYBR® FAST qPCR Master Mix to estimate the size and quantification. These WGS libraries were

sequenced on the Illumina HiSeq-X instrument to generate 150 bp paired-end reads. All flow cell data (BCL files) are converted to barcoded FASTQ files.

## **2.4 Mapping and variant calling**

BWA-MEM version 0.7.12-r1039 [85] was used to align Illumina reads to the owl monkey reference assembly Anan\_2.0 (GenBank assembly accession GCA\_000952055.2) and to generate BAM files for each of the 30 individuals. Picard MarkDuplicates version 1.105 (<http://broadinstitute.github.io/picard/>) was used to identify and mark duplicate reads. Single



**Figure 2.4: Read depth and allelic balance distributions and the effect of varying the allelic balance cut-off on rate estimates. Related to Figure 2.1.** **A**, read depth and allelic balance distributions for all unfiltered Mendelian violations (MVs) in the 30 owl monkey individuals. The cut-offs used to filter MVs are indicated by the vertical dotted grey lines. **B**, read depth and allelic balance distributions for all SNP sites in the 30 owl monkey individuals. Filtering cut-offs are again indicated by the vertical dashed grey lines for comparison. **C**, mutation rate estimates for the 14 owl monkey trios when using a less stringent allelic balance cut-off to 0.3 and 0.7 (purple dots). A linear regression still shows a correlation with father's age (solid purple line;  $R^2=0.15$ , d.f.=12  $P=0.10$ ); shaded area indicates 95% confidence interval) that is not significantly different from our model's prediction (dashed purple line;  $F=1.0$ , d.f.=13,  $P=1.0$ ). Mutations at CpG sites (blue dots) are not correlated with father's age (blue line). The grey dotted line indicates age of puberty for owl monkeys.

nucleotide variants (SNVs) and small indels (up to 60bp) were called using GATK version 3.3-0 following best practices [86, 87]. HaplotypeCaller was used to generate gVCFs for each sample. Joint genotype calling was performed on all samples using GenotypeGVCFs to generate a VCF file. GATK hard filters (SNPs: "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"; Indels: "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0") (<https://software.broadinstitute.org/gatk/documentation/article?id=2806>) were applied and calls that failed the filters were removed.

GATK's PhaseByTransmission was used to identify Mendelian violations that represent possible *de novo* variants. After removing Mendelian violations (MVs) that resulted from missing genotypes or had other anomalies (i.e. 5 MVs with read depth of 0 and 1,984 MVs with allelic depth of 0,0), we obtained 45,432 putative Mendelian violations. We also identified 62 scaffolds as deriving from the X chromosome. These scaffolds had significantly higher homozygosity and lower mean read depth among males (one-tailed t-test,  $q < 0.05$  for both mean homozygosity and read depth). MVs on these scaffolds and scaffolds shorter than 10 kb were removed. This resulted in an initial set of 34,189 putative MVs.

## 2.5 Filtering of putative mutations

Stringent filters are necessary to avoid potential false positive calls of *de novo* mutations [35, 88, 89]. To address this issue we applied the following filters to our initial set of MVs:

1. Removed 32,638 MVs with allelic balance less than 0.4 or greater than 0.6 in the child.
2. Removed 112 MVs that are not homozygous reference in both parents.
3. Removed 636 MVs with read depth below 20 or above 60 in any individual in the trio.

4. Removed 520 MVs where the alternate allele is present in an unrelated individual in the sample.

We define allelic balance as the fraction of reads that are a non-reference allele at a given site, meaning that a true heterozygous site should have allelic balance of roughly 0.5.

Importantly, we observed that 95% of all initial MVs have allelic balance less than 0.4 (Figure 2.4A). This indicates that many of these initial calls are false positives. After these four filtering steps we find a total of 283 *de novo* mutations across our 14 trios (Data S1B).

## 2.6 Phasing mutations

Genotypes from three generations allow us to trace the parent of origin for *de novo* mutations transmitted to the third generation. We accomplished this by phasing chromosomal segments with respect to the grandparents (P generation in Figure 2.1A). Phase informative sites were identified in each family and assembled into haplotype blocks. We selected bi-allelic informative sites where: the grandparents had different genotypes, their offspring was heterozygous, and this individual's partner and offspring were not both heterozygous. The transmission of alleles at these sites can be unambiguously traced to one of the grandparents. We assembled these sites into blocks under the assumption that no more than one recombination occurred per 0.5 Mb interval [74, 90]. The phases of haplotype blocks supported by fewer than 100 informative sites were left unassigned, as were the phases of short scaffolds (less than 0.5 Mb). The parent of origin for *de novo* mutations transmitted to the third generation can then be established from the phase of their corresponding haplotype block.

## 2.7 Estimating mutation rates

To estimate mutation rates per generation per site ( $\mu_g$ ) we must consider rates of error. Our stringent filters ensure that we have few to no false positives; however, we expect that these filters removed a number of true *de novo* variants, leading to a substantial false negative rate ( $\alpha$ ). To estimate  $\alpha$  resulting from the allelic balance filter, we used the distribution of allelic balance from the total set of 471,532,403 heterozygous autosomal sites in our sample. Unlike the initial set of MVs, the distribution of allelic balance for these sites conforms to the expected distribution for true heterozygous sites, with a single peak at about 0.5 (Figure 2.4B). We find that the number of heterozygous sites with allelic balance below 0.4 or above 0.6 is 206,358,774 resulting in an estimate of  $\alpha = 0.44$ . With a less stringent allelic balance filter of 0.3-0.7 the false negative rate falls to 0.29, but changing this filter does not greatly impact the number of mutations called (Figure 2.4C). These numbers represent false negative estimates from the allelic balance filter alone and in that sense only represent the upper-bound from that filter. False negatives may occur during other filtering steps, or due to mis-calls from the variant identification process, however these numbers are difficult to estimate. Therefore, we correct the observed number of mutations ( $m_g$ ) in each trio using  $\alpha = 0.44$  and an assumed false positive rate of 0. After correction we estimate that there are about 36 *de novo* mutations passed on in a single owl monkey generation.

To calculate the mutation rate per site, we counted the number of callable sites in each trio ( $C$ ). A site was determined to be callable if it passed filters (1) and (4) in the child, filter (2) in the parents, and filter (3) in all individuals in the trio. We find an average number of callable sites of 2,207,614,768 in our 14 trios (range: 2,198,415,883-2,214,425,687). Mutation rates were then calculated by dividing the number of observed mutations (corrected for  $\alpha$ ) in a trio by 2 times the number of callable sites:

$$\mu_g = \frac{m_g}{((1 - \alpha) \cdot (2 \cdot C))} \quad (\text{E2.15})$$

This results in mutation rates ranging from  $0.63 \times 10^{-8}$  to  $1.5 \times 10^{-8}$  with an average mutation rate of  $0.81 \times 10^{-8}$  among the 14 trios (Data S1C). Mutation rate was then regressed on father's age ( $A_M$ ) (Figure 1B, solid line) with the resulting formula for a best fit line:

$$\mu_g = 3.74 \times 10^{-9} + (A_M \cdot 6.62 \times 10^{-10}) \quad (\text{E2.16})$$

With an average haploid genome size of 2.21 billion base pairs, this means that 16.53 mutations accumulate in males and females before puberty at age 1 and that there are 2.92 additional mutations for every year of the father's life after puberty in owl monkeys.

## 2.8 Modeling mutation rates

Large-scale pedigree sequencing projects in humans have shown the importance of different life-stages in the determination of mutation rates [35, 65-71, 89, 91]. Models for predicting mutation rates generally account for the three important life stages in the mammalian germline [73, 80]. These life stages are (1) female ( $F$ ), (2) male before puberty ( $M0$ ), (3) and male after puberty ( $M1$ ). The relative contribution of each of these stages must be accounted for when estimating mutation rates per generation [80] or per year [73, 80, 92]. Here, we re-frame this model in terms of reproductive longevity. Reproductive longevity depends on both the age of puberty in males ( $P_M$ ) and the age of the father at conception of his offspring ( $A_M$ ) and we find that it is the main determinant of mutation rate variation in primates. We define the value of reproductive longevity ( $RL$ ) as:

$$RL = A_M - P_M \quad (\text{E2.17})$$

$RL$  therefore measures the amount of time mutations have accumulated post-puberty in a male, which only occurs during stage  $M1$ .

To see how reproductive longevity affects the per-generation mutation rate,  $\mu_g$ , we must model the combined contribution from all life stages. In any given period of time  $t$ , the mutation rate due to errors in DNA replication,  $\mu_t$ , is simply a product of the mutation rate per cell division,  $\mu_c$ , and the number of cell divisions that occur,  $d_t$ :

$$\mu_t = \mu_c \cdot d_t \quad (\text{E2.18})$$

Since females (stage  $F$ ) and pre-puberty males (stage  $M0$ ) have a fixed number of cell divisions, their contribution to the mutation rate per-generation is constant and requires only the substitution of appropriate terms into equation 4:

$$\text{Female contribution to } \mu_g: \quad \mu_{gF} = \mu_c \cdot d_F \quad (\text{E2.19})$$

$$\text{Pre-puberty male contribution to } \mu_g: \quad \mu_{gM0} = \mu_c \cdot d_{M0} \quad (\text{E2.20})$$

However, in males after puberty (stage  $M1$ ) the number of cell divisions is a linear function of time, and the mutation rate per-generation in this life stage therefore depends on the yearly rate of cell division ( $d_{yM1}$ ) and reproductive longevity ( $RL$ ):

$$\text{Post-puberty male contribution to } \mu_g: \quad \mu_{gM1} = \mu_c \cdot d_{yM1} \cdot RL \quad (\text{E2.21})$$

Finally, since an autosome will spend roughly half of its time in females and half in males, the mutation rate per generation ( $\mu_g$ ) for a given species is the average of the male and female contributions:

$$\mu_g = \frac{\mu_{gF} + (\mu_{gM0} + \mu_{gM1})}{2} \quad (\text{E2.22})$$

Given estimates of the underlying mutational parameters, this model allows us to predict the mutation rate as a function of reproductive longevity. In order to assess reproductive longevity in species that reach puberty at different times, we used published values for  $P_M$ . For owl monkeys, we set  $P_M$  at 1 year [75] (purple line in Figure 3), for humans, we used a value of  $P_M$  of 13.4 years [93] (orange line in Figure 3), for chimpanzees we used 7.5 years [94] (red line in Figure 3). The ages at conception for all parents in all studies of the mutation rate (points in Figure 3; Data S1D) were taken from the original papers [35, 65, 66, 68-71, 74, 89, 95].

This mutational model can easily be extended to calculate mutation rates per year ( $\mu_y$ ) [80, 92] by averaging the mutational contribution from each life stage per generation and weighting by the amount of time that passes:

$$\mu_y = \frac{(\mu_{gF} + (\mu_{gM0} + \mu_{gM1}))}{(A_F + P_M + RL)} \quad (\text{E2.23})$$

Considering yearly rates is useful when comparing long term evolutionary rates ( $k$ ) between species since the neutral mutation rate ( $\mu$ ) is inextricably linked to the neutral substitution rate ( $\mu_y = k_y$ ).

Unlike  $\mu_g$ , which is only dependent on the age of puberty and age at conception in males,  $\mu_y$  is also dependent on the age of conception in the female ( $A_F$ ; Figure S3). This means that increasing  $A_F$  will most likely decrease the yearly mutation rate because it increases the absolute

amount of time without increasing the number of germline cell divisions. However, variation in either  $P_M$  or  $RL$  will have more complicated effects as they appear in both the numerator (as  $RL$  in  $\mu_{g_{M1}}$ ) and the denominator. Increasing  $RL$  at some points in parameter space will increase  $\mu_y$ , while decreasing it at others. Increasing  $P_M$  tends to increase  $\mu_y$  (Figure S3).

## 2.9 Estimating mutational parameters from humans

Empirical observations from developmental studies and large-scale pedigree data from humans inform us about some of the underlying mutational parameters of our model (equations 5, 6, and 7). For example, we use 31 and 34 as estimates for the number of cell divisions in human females ( $d_F$ ) and males before puberty ( $d_{M0}$ ) [25]. We use 16 days as the length of a single spermatogenic cycle ( $t_{sc}$ ) [96], which means we expect  $d_{yM1} = 23$  spermatogenic cycles to occur in a year if all spermatogonial cells are constantly dividing (but see next paragraph).

The remaining parameter of the model,  $\mu_c$ , can be estimated from human pedigrees. We confirm the estimate of  $\mu_c$  made by Amster and Sella [92] by using the  $\mu_{gF}$  observed in Kong et al. [35] of 14.2, the number of female germline divisions, and rearranging equation 5:

$$\mu_c = \frac{14.2}{31} = 0.458 \quad (\text{E.224})$$

or  $1.74 \times 10^{-10}$  given a haploid genome size of 2.63 billion base pairs [35]. We assume this rate is the same between females and males before puberty. However, the observation that 2.01 mutations are passed on per year from the father after puberty [35] (the mutation rate per year in this lifestage,  $\mu_{yM1}$ ) could imply two things about  $\mu_c$  in this life-stage: either the mutation rate per cell division has been reduced by an order of magnitude in males after puberty to  $0.33 \times 10^{-11}$  [92] or there are fewer than the expected 23 cell divisions per year [97]. There is no evidence

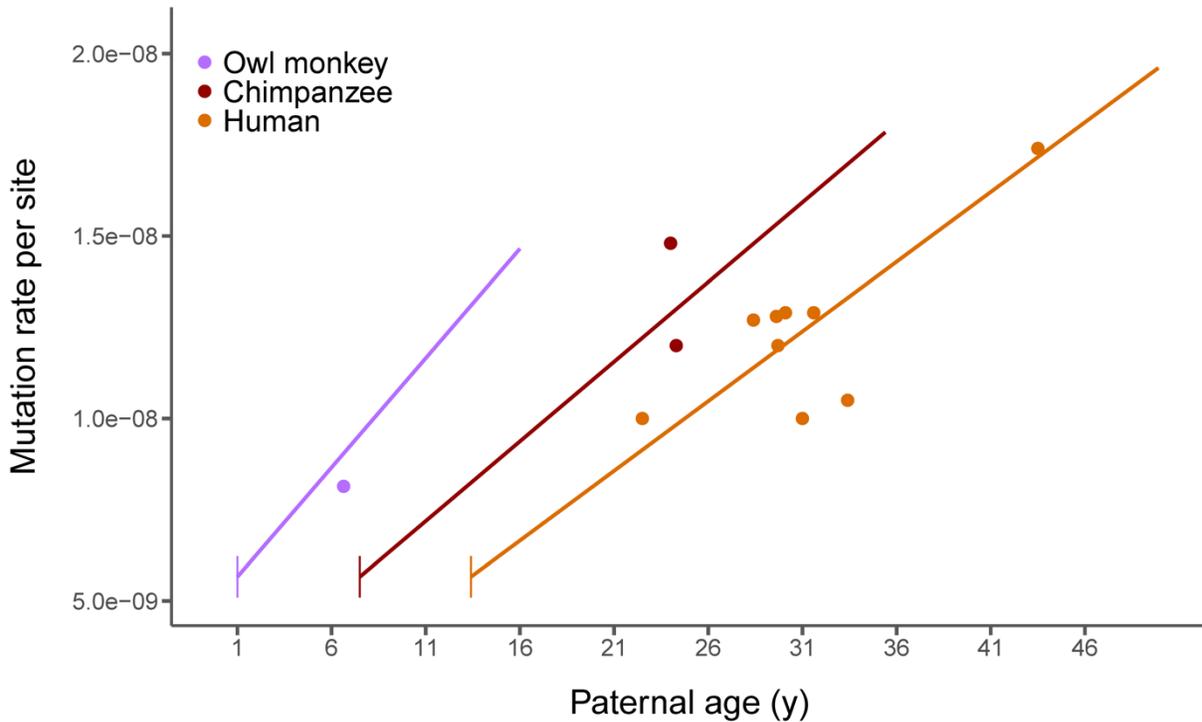
to support such a dramatic reduction in the mutation rate per cell division, especially since there does not appear to be a large shift in mutational mechanisms between life stages [71]. The hypothesis that fewer cell divisions have taken place is also more likely based on observations that, of the two types of spermatogonial cells observed in humans, pale and dark, only pale cells actively divide [97, 98]. If dividing pale cells transition into non-dividing dark cells and vice versa, then not all spermatogonial cells necessarily undergo 23 spermatogenic cycles in a year and we must re-estimate  $d_{yM1}$ . If we assume the mutation rate per cell division in humans is constant before and after puberty, we can estimate the expected number of spermatogenic cycles per year ( $d_{yM1}$ ):

$$d_{yM1}^{Human} = \frac{\mu_{yM1}}{\mu_c} = \frac{2.01}{0.458} = 4.39 \quad (\text{E2.25})$$

This implies that roughly only 19% of spermatogonial cells are in the pale dividing state at any given time.

Though either a decreased  $\mu_c$  in males after puberty or a decreased proportion of dividing spermatogonial cells can be fit equally well to the model, we make predictions with the latter assumption. When predicting a mutation rate function for owl monkeys (Figure 2.1B) we also decrease the length of the spermatogenic cycle to  $t_{sc}^{Owl\ monkey} = 10.2$  days [99] and adjust the expected  $d_{yM1}$  assuming 19% of spermatogonial cells are undergoing spermatogenesis at one time:

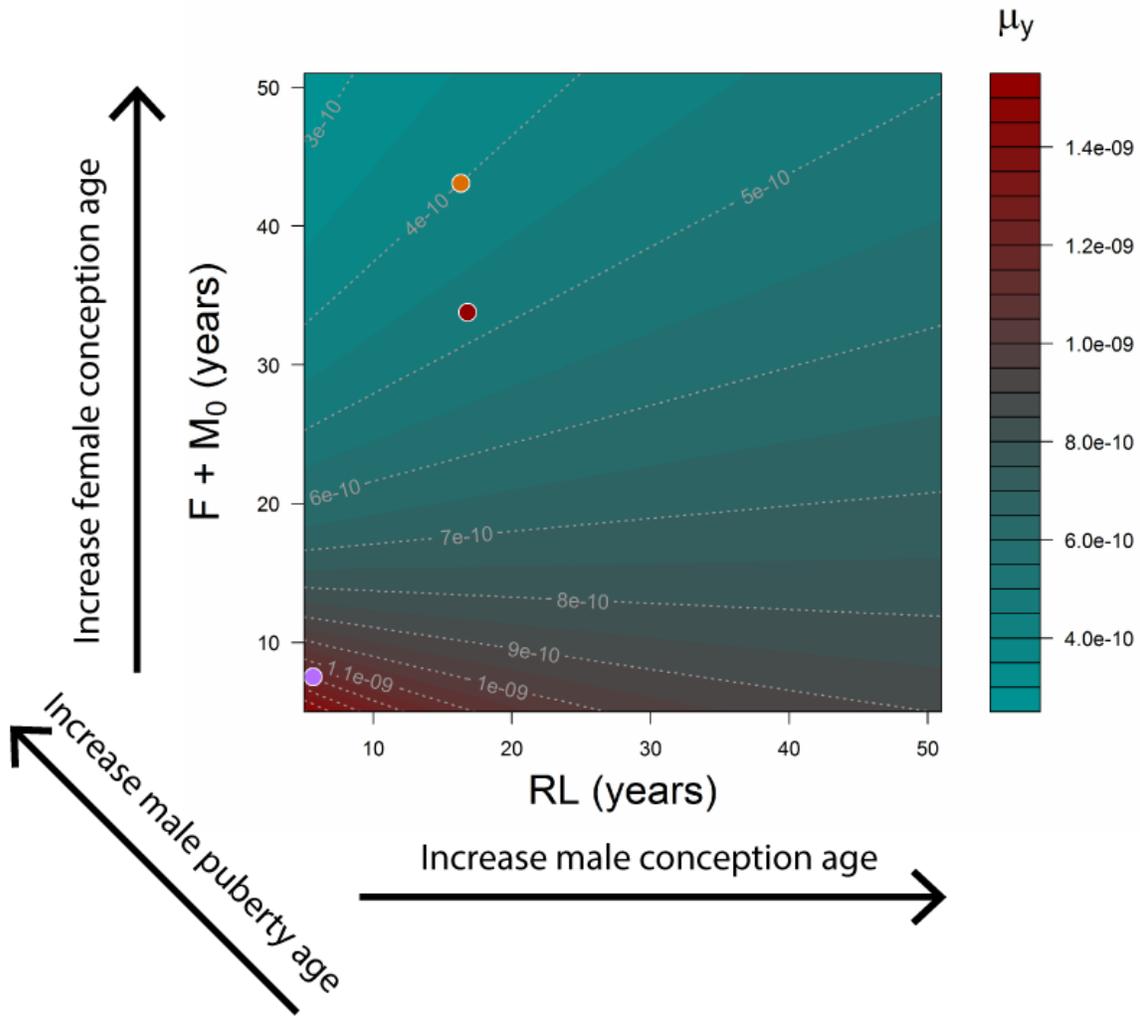
$$d_{yM1}^{Owl\ monkey} = \left( \frac{365}{t_{sc}^{Owl\ monkey}} \right) \cdot 0.19 = 6.88 \quad (\text{E2.26})$$



**Figure 2.5:** Functions of mutational accumulation predicted using species specific rates of spermatogenesis. Related to Figure 3. Using species specific rates for the three species with high-quality mutation estimates when making predictions from our model of reproductive longevity (lines; equations 3-8 in Methods) does not greatly affect the fit of the model to published estimates of mutation rates (points; see Data SID for references). Vertical line segments represent the age of puberty for each species. In this figure we used  $t_{sc}^{Owl\ monkey} = 10.2$ ,  $t_{sc}^{Chimp} = 14$ ,  $t_{sc}^{Human} = 16$ .

However, when comparing mutation rate functions between species (Figure 2.3) all underlying mutational parameters are those estimated above from observations in humans, in order to demonstrate that minimal changes to the model can still make accurate predictions of mutation rate functions. Using species-specific parameters of spermatogenesis does not change our results (Figure 2.5).

Using equation 9, we are also able to predict  $\mu_y$  for an assumed age of puberty and average age of conception for humans and owl monkeys. For humans, with an age of puberty of roughly 13.4 years and average age of conception for both males and females of 30 years, we estimate a



**Figure 2.6:** Modeling mutation rate per year over various parental ages at reproduction and puberty. Mutation rates per year ( $\mu_y$ ) are a function of the length of all three life stages of the mammalian germline: Female ( $F$ ), males before puberty ( $M_0$ ), and males after puberty ( $RL$ ). Increasing  $F$  consistently decreases  $\mu_y$ , while increasing  $M_0$  by increasing the age of puberty in males tends to increase  $\mu_y$  (note that the y-axis plots the sum of these two parameters). Increasing the length of  $RL$  either by decreasing the age of puberty or increasing the average age of conception in males has varying effects based on the length of the other two stages. Points are the predicted mutation rates per year for humans (orange), chimpanzees (red), and owl monkeys (purple).

yearly mutation rate of  $0.4 \times 10^{-9}$  mutations per site per year (orange point in Figure 2.6). This is remarkably close to the calculated average yearly rate from several studies of human mutation rates: from  $0.43 \times 10^{-9}$  mutations per site per year [71] to  $0.5 \times 10^{-9}$  mutations per site per year [79]. With an average age of puberty of 7.5 years and average age at reproduction of 24.3, we predict the yearly chimp mutation rate is to be  $0.48 \times 10^{-9}$  per site per year (red point in Figure S3), on par with the previous estimate of  $0.46 \times 10^{-9}$  (Venn et al. 2014). For owl monkeys we assumed a puberty age of 1 year and average ages of conception of 6.64 and 6.53 years for males and females, respectively. Using these values we estimate a yearly mutation rate of  $1.2 \times 10^{-9}$  mutations per site per year, three times higher than the yearly human rate (purple point in Figure S3).

## 2.10 Quantification and Statistical Analysis

All data were analyzed using Python v2.7 and R v3.4.1. Linear regression was performed on the observed mutation rate per trio and paternal age to obtain the solid lines in Figures 2.1B and 2.4. To assess how well our model predicts this relationship, we performed an *F*-test on the residuals of the observed relationship (solid lines in Figures 2.1B and 2.4) and the predicted relationship (dashed lines in Figures 2.1B and 2.4). Comparing variance in the residuals between the two lines captures variation in both the slope and intercept of the predicted and observed lines. A similar *F*-test was performed on the human study points in Figures 2.3 and 2.5.

## 2.11 Data Availability

Raw sequence reads for the 30 owl monkey individuals have been deposited as an NCBI BioProject (Accession: PRJNA451475; <https://www.ncbi.nlm.nih.gov/bioproject/451475>).

## 2.12 Additional Resources

All code used to analyze data and generate figures is available as an R Markdown document at the following link: <https://github.com/gwct/owl-monkey>.

## **CHAPTER 3: Referee: reference assembly quality scores**

### **3.1 Introduction**

Reference assemblies are haploid representations of the genome sequence of a species. Their use is ubiquitous in modern genetic and evolutionary research, especially in comparative genomics studies. Such studies range from questions about phylogenetic relationships to analyses searching for targets of adaptive natural selection. The conclusions of all analyses depend on the accuracy of the reference sequence; however, both genome assembly methods and the underlying sequencing technologies are error-prone [100]. This inevitably leads to errors in downstream analyses and conclusions [e.g. 101, 102, 103].

Many technologies provide a measure of base accuracy for every position in a sequencing read in the form of the quality score. This score represents the log-scaled value of the probability that the called base is incorrect. However, when assembling reads from genomes, transcriptomes, or other reduced-representation sequencing approaches [e.g. 104] this quality information is lost. Here we present Referee, a program that provides a measure of the underlying quality for an assembled reference sequence. Referee uses genotype quality likelihoods, which are standard in resequencing studies [e.g. 105], to calculate a haploid reference quality score. The quality score,  $Q_R$ , ranges between 0 and 90 and represents the confidence we have that the called base at that position is correct. For positions where we have no confidence in the called base, Referee can suggest an alternate, better-scoring base. While tools do exist that examine assembly quality at a per-base level [106], Referee aims to produce an easily interpretable quality score for any type of assembly, using any sequencing technology. These scores can then be used to inform any downstream analysis.

### **3.2 Materials & Methods**

Referee uses the genotype likelihoods of all 10 possible diploid genotypes at a site to calculate a the quality score,  $Q_R$ , of the single base in the reference sequence. Referee summarizes the diploid genotype likelihoods for the haploid representation of the assembly by taking the sum of the likelihoods of the genotypes that contain the called base ( $L_{match}$ ) and the sum of those that do not contain the called base ( $L_{mismatch}$ ). For instance, if the called base is A, then  $L_{match} = L(AA) + L(AT) + L(AC) + L(AG)$  and  $L_{mismatch} = L(TT) + L(TC) + L(TG) + L(CC) + L(CG) + L(GG)$ .

Taking the log-scaled ratio of these two sums gives us a quality score:

$$Q_R = \log\left(\frac{L_{match}}{L_{mismatch}}\right) \quad (\text{E3.1})$$

This scoring has the desirable behavior of being positive when we think the called reference base is correct and negative when we think it is incorrect due to lack of support; scores close to 0 indicate uncertainty in the called base. For sites that show more support for an alternate base call (i.e. sites with  $Q_R \leq 0$ ), Referee can calculate  $Q_R$  for each of the three alternate bases and suggest the highest scoring base for that position.

### *Genotype likelihoods*

Referee's quality score requires genotype likelihoods from the reference individual. Such likelihoods are calculated by mapping the reads used in generating the assembly back to the reference assembly. Referee can calculate genotype likelihoods for each site if given a pileup file as input. For this calculation we have implemented the Bayesian model of genotype likelihood developed in McKenna, et al. [107], with the additional consideration of mapping quality:

$$P(R | g = \{A_1, A_2\}) = \prod_r^R \left( \frac{1}{2} P(b_r | A_1) + \frac{1}{2} P(b_r | A_2) \right) \quad (\text{E3.2})$$

Where  $R$  is the full set of reads that have mapped to a site and  $A_1$  and  $A_2$  are the two alleles in the genotype. To calculate the probability of a base  $b_r$  given an allele, we use the standard Phred conversion between quality scores  $Q$  and error probabilities  $e$ :

$$e = 10^{-\frac{Q}{10}} \quad (\text{E3.3})$$

This conversion is used for both base calling and mapping qualities, resulting in error probabilities for both base calling ( $e_b$ ) and read mapping ( $e_m$ ). Then:

$$P(b_r | A_i) = \begin{cases} \frac{e_b \cdot e_m}{3} & : b \neq A_i \\ 1 - (e_b \cdot e_m) & : b = A_i \end{cases} \quad (\text{E3.4})$$

To avoid underflows, sums of logs of probabilities are taken in E3.2 rather than products of raw probabilities.

Referee also accepts genotype log-likelihoods as input from any method provided that they are formatted correctly. For example, the program ANGSD [108] has the capability to output all 10 genotype log-likelihoods in a format readily acceptable by Referee. Note that although ANGSD scales log-likelihoods by subtracting the highest score from all scores, this has no effect on Referee's calculations.

### 3.3 Referee's scoring system

Since the quality score calculated by Referee is a ratio of probabilities, theoretically any score from negative to positive infinity is possible. In practice, scores tend to be limited to a range of -300 to +300 and have a strong correlation with read-depth. For practical reasons, Referee's standard output limits the scores to a range of 0 to 90. This means that any negative score is converted to a score of 0, and any score above 90 is converted to 90. This makes the

**Table 3.1:** Referee score special cases

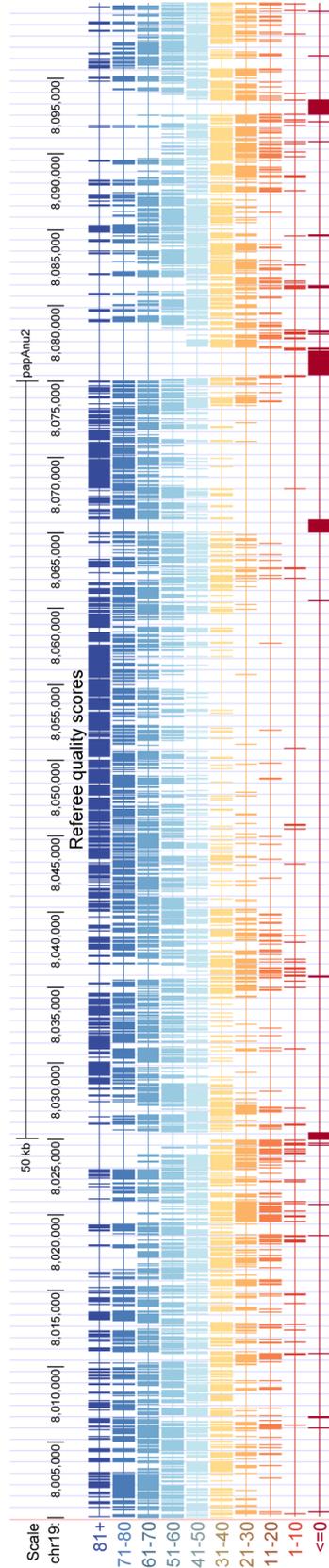
Scenario	$Q_R$ score
$L_{mismatch} = 0$	91
Reference base called as N	-1
No reads mapped to site	-2

scores easily interpretable on a Phred-like scale and allows for conversion to ASCII characters for condensed FASTQ output.

There are several scenarios in which it is not possible to calculate a quality score (Table 3.1): In cases of very high read-depth, with all or most reads supporting the called base, it is possible that the sum of likelihoods for genotypes that do not contain the reference base ( $L_{mismatch}$ ) will be 0. In these cases we are confident that the reference base is correct and assign a score of 91. If the reference base is an N or if no reads have mapped to the site we have no way of calculating  $Q_R$ , so we assign scores of -1 and -2, respectively, to indicate our uncertainty. In order to accommodate the -1 and -2 scores, quality scores are output as ASCII characters corresponding to  $Q_R + 35$  (note that this scaling differs slightly from the standard Phred conversion).

### 3.4 Results

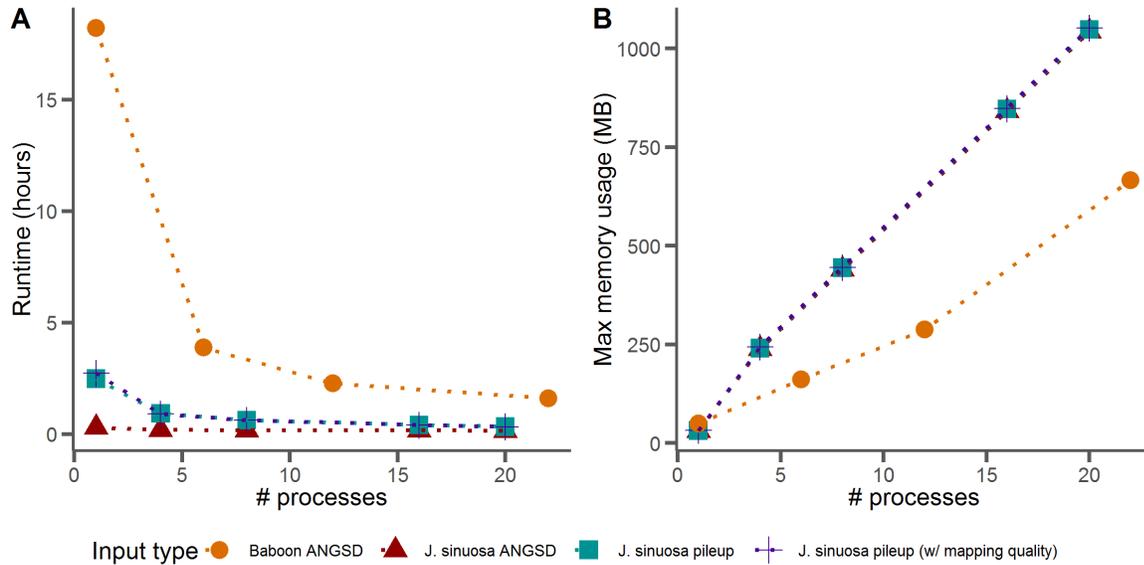
Referee is implemented entirely in Python, compatible with versions 2.7 and above and is freely available (<https://gwct.github.io/referee/>). Referee takes as input a single reference FASTA file representing the reference assembly and either pre-calculated genotype log-likelihoods or a pileup file from which it can calculate genotype likelihoods. Referee will output quality scores for every position in the input FASTA in either a simple tab delimited format (akin to the pileup)



**Figure 3.1:** Reference quality scores visualized for a 100,000 bp stretch of chromosome 19 in the baboon genome (*Papio anubis* v2.0) on the UCSC Genome Browser (<http://genome.ucsc.edu>)

or in FASTQ format, with quality scores being converted to ASCII characters. Referee can also output quality scores in BED format, which can be used for visualizing tracks of scores in most genome browsers. Figure 3.1 shows a 100 kb stretch of Referee quality scores on chromosome 19 of the baboon genome (papAnu v2.0) in the UCSC Genome Browser [109].

Referee is intended for use on assemblies of any size, and from any technology that provides reads with base quality scores (e.g. Illumina or Oxford Nanopore). To make it scalable with even the largest of today's sequenced genomes, Referee is designed to use multiple processes without a large memory footprint. We tested the performance of Referee on two datasets: a transcriptome assembly from *Jaltomata sinuosa* [110] using Illumina RNA-seq reads (SRA accession SRX2676125) and a genome assembly from the baboon, *Papio Anubis* (GCF\_000264685.2) using only the Illumina paired-end reads that were used in the assembly process (SRA accessions: SRR927653, SRR927654, SRR927655, SRR927656, SRR927657, SRR927658, SRR927659). Test runs were done on Indiana University's Carbonate computer cluster (Red Hat Enterprise 7.x with 256 GB of RAM and two 12-core Intel Xeon E5-2680 v3 CPUs). For *J. sinuosa* the reads were assembled with Trinity [111] and for both species reads were mapped back to their respective assemblies with BWA [112]. We find that for the *J. sinuosa* transcriptome, even when utilizing only one process, Referee completes in 20 minutes with pre-calculated genotype likelihoods. Unsurprisingly, calculating the likelihoods is detrimental to run-time, raising it to 2.73 hours, but allocating additional processes more than makes up for this time loss (Figure 3.2a). For the much larger baboon genome dataset we observe a run time of 18 hours when using pre-calculated genotype likelihoods. Again this is drastically reduced to 1.6 hours when using multiple processes (Figure 3.2a). Memory usage never exceeds 1 GB (Figure 3.2b). This makes Referee widely usable regardless of operating



**Figure 3.2:** Referee’s run time (A) and max memory usage (B) on the *J. sinuosa* transcriptome and baboon genome. Note the memory improvement in the baboon genome data compared to the *J. sinuosa* transcriptome data as a result of splitting the input files by chromosome. ANGSD: genotype log-likelihoods were pre-calculated with the ANGSD software package and given as input to Referee; pileup: A pileup was created from the mapped reads which Referee used to calculate genotype likelihoods using only the base quality scores from the reads; pileup (w/ mapping quality): The mapping qualities for the reads were included in the pileup and incorporated into Referee’s genotype likelihood calculations.

system.

### 3.5 Conclusions

The wide-ranging applicability of genome assemblies in modern biological research means their accuracy is of utmost importance in order to reach unambiguous conclusions. Evolutionary inferences into species relationships and the targets of positive selection depend on this accuracy. Referee adds a simple step between the assembly and analysis of a genome to improve the assembly for all purposes. By accounting for the underlying base quality in the reads and the diploid nature of most genome assemblies, Referee’s scores can be used to inform researchers of sites to filter from their analyses or of better scoring alternate bases. This is accomplished through a fast and easy to use software package: <https://gwct.github.io/referee/>.

## CHAPTER 4: Origins and long-term patterns of copy-number variation in rhesus macaques

### 4.1 Introduction

Mutations are the source of all genetic variation and can have both immediate and lasting impacts for the evolution of a species. Understanding how mutations arise and spread through a population in the short-term can therefore aid our understanding of disease while understanding its effects in the long-term aid our understanding of evolution in populations and species. Recent work in humans and other primates have unveiled patterns of mutation for single nucleotide variants using pedigrees of related individuals. For instance, studies in primates have found a strong paternal age effect on the number of *de novo* single nucleotide mutations: older fathers tend to pass on more mutations [35, 71, 74, 113]. This is likely due to a combination of errors accruing from both ongoing spermatogenesis and unrepaired DNA damage. However, no such paternal age effect has been found among *de novo* deletions and duplications (also known as copy-number variants, or CNVs) in humans [69, 114-116] though the origin of CNVs are studied more rarely than single nucleotide mutations [57, 69, 114-119].

The frequency and locations of CNVs have been found to be highly variable among primates [120-123], though several CNV hotspots in multiple species have been described [124-126]. Duplications in genic regions have been found to outnumber deletions in many lineages when comparing closely related species [122, 127, 128], possibly indicating a selective difference between gene duplications and deletions. However, recent whole-genome studies within humans point to different a pattern in non-genic regions, with deletions far outnumbering duplications [115]. In order to determine whether such patterns are specific to humans, or are

representative of the joint effects of mutational input and selection on the long-term survival of duplicates and deletions, we require fine-scale studies in additional species.

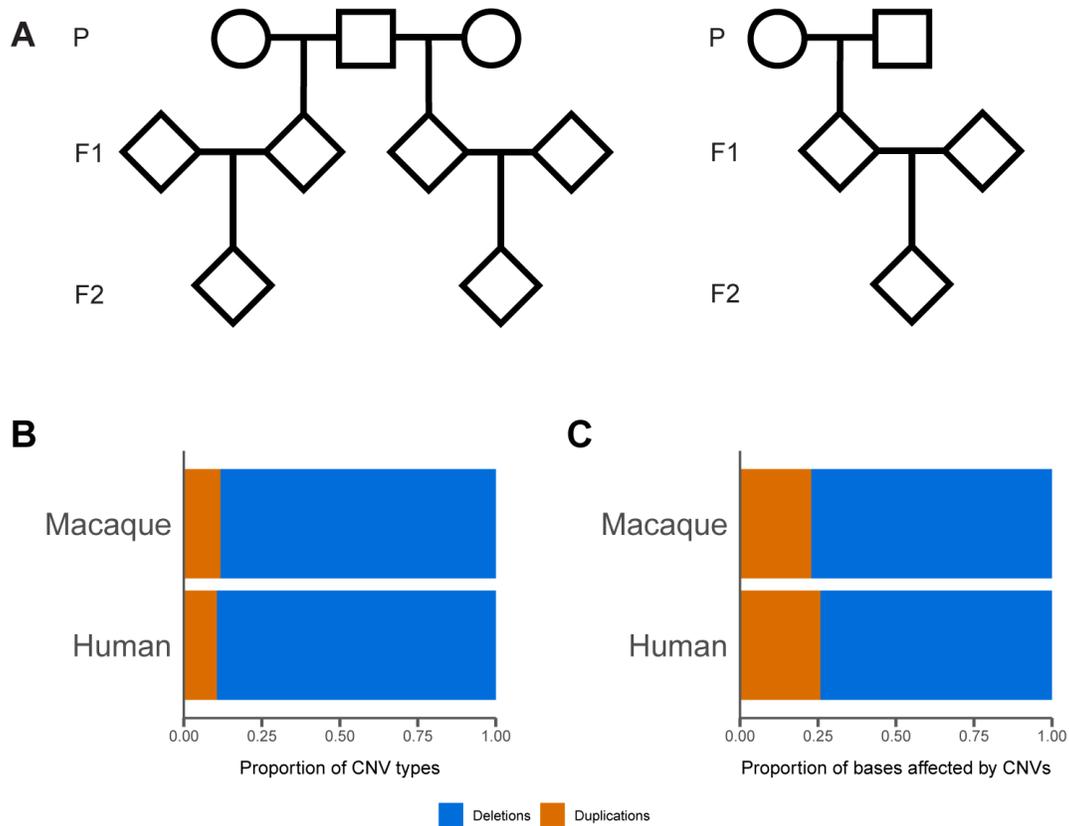
Rhesus macaques are a widely used model organism, especially for diseases in humans. Understanding the underpinnings of genetic variation in this species may help to enhance disease models, in addition to aiding our understanding of the genetic basis of evolutionary change. Previous studies of rhesus macaque CNVs have used array-based comparative genomic hybridization (aCGH) to detect events and have found that the frequency of duplications either matches or exceeds that of deletions [126, 129]. However, aCGH methods are limited in their detection of short deletions and duplications [130, 131]. Patterns of variation CNVs shorter than the detectable limit by aCGH remain uncharacterized. Read-based methods—which use read depth, read orientation, discordance of paired-end reads from a reference genome, or a combination of these signals (reviewed in [131, 132])—may help to clarify patterns of duplication and loss and will allow us to learn how CNVs arise in primates.

Here, we use deep sequencing of 32 rhesus macaques in 14 trios to uncover patterns of copy-number variation in this species. We find that, contrary to aCGH studies, deletions make up the vast majority of polymorphic CNVs within rhesus macaques. Using unrelated individuals, we find that patterns of segregating CNVs are similar between macaques and humans. By sequencing parent-offspring trios we are also able to investigate the occurrence of *de novo* CNVs. We find that the number of *de novo* CNVs per generation is less than one per genome in both macaques and humans, and that parental age has no effect on the rate of these types of mutations in either species. Finally, we compare patterns of deletions and duplications in our sample to those of long-term gene gains and losses along the lineage leading to macaques. Interestingly, while deletions make up the vast majority of polymorphisms in our sample, the

number of genes gained and lost along the macaque lineage are roughly equal. These patterns give us a first look at structural variation using whole-genome sequencing in a non-human primate and will help model these types of mutations better in disease prediction and evolutionary analyses.

## 4.2 Patterns of copy-number variation in rhesus macaques

We identified CNVs by sequencing the whole genomes of 32 rhesus macaque (*Macaca mulatta*) individuals within 14 trios (Figure 4.1A; [133]). We mapped the reads from these samples to the reference macaque genome ([134]; rheMac8.0.1 downloaded April 12, 2018) and identified CNVs based on split and discordant read patterns using Lumpy [135], SVtyper [136], and SVtools [137] and filtered these calls by read-depth using Duphold [138]. In total we found 3,313 deletions and 441 duplications relative to the reference genome, meaning that roughly 88% of variants segregating in our sample are deletions (Figure 4.1B). This is in stark contrast to previous CNV studies in rhesus macaques, which found roughly half of events to be deletions and half to be duplications [126]. One possibility for this difference is that the previous study could not resolve events shorter than a few kilobases (minimum length 3518 bases), while the read-based methods employed here can. This contrast from an increased level of resolution is consistent with studies in *Drosophila melanogaster* that found a bias toward deletions for short events [57]. We find that macaque CNVs are distributed across all macaque chromosomes, but unevenly, with some stretches completely void of events and others where CNVs seem to be enriched (Figure 4.2). Contrary to previous studies in rhesus macaques [129], we find that the number of CNVs on a chromosome is strongly correlated with the length of the chromosome (Figure 4.3). This may again be the result of the increased resolution in our study. We also observe some clustering in the telomeric regions (Figure 4.2). This telomeric clustering is



**Figure 4.1:** (A) Pedigrees of sequenced macaques. The 14 trios were contained within 3 families similar to the one on the left, and 1 family similar to the one on the right. (B) The proportion of CNV types (deletions or duplications) and (C) bases affected by CNVs for rhesus macaques compared to humans.

consistent with the duplication maps of the macaque genome [134] and the human genome [122, 130, 139], and is likely driven by the higher concentration of transposable elements in these regions, which mediates higher levels of non-allelic homologous recombination (i.e. unequal crossing-over).

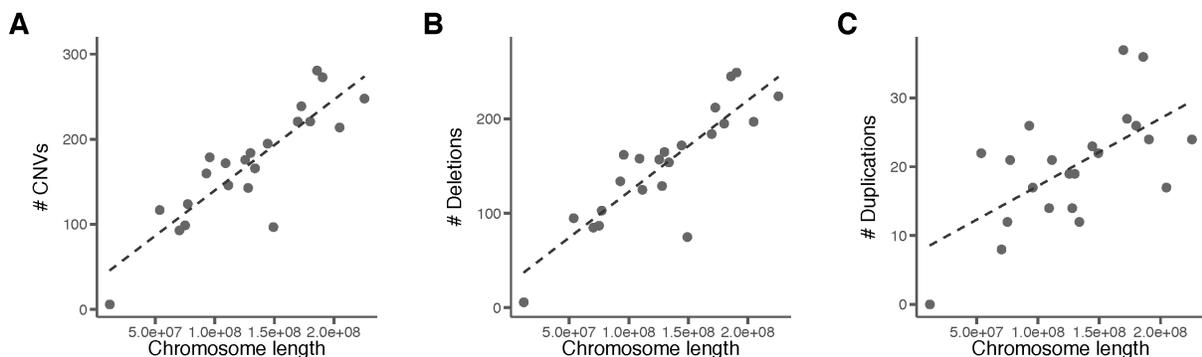
We used published CNVs from a sample of 235 humans [115] to study the similarities and differences between primate species. We find that the proportions of segregating deletions and duplications are not significantly different between the two species (Figure 1B;  $\chi^2 = 3.77$ , d.f. = 1,  $p > 0.05$ ). Given the observed bias toward deletions, it is unsurprising that both species have a higher proportion of bases deleted than duplicated (Figure 4.1C). The average individual



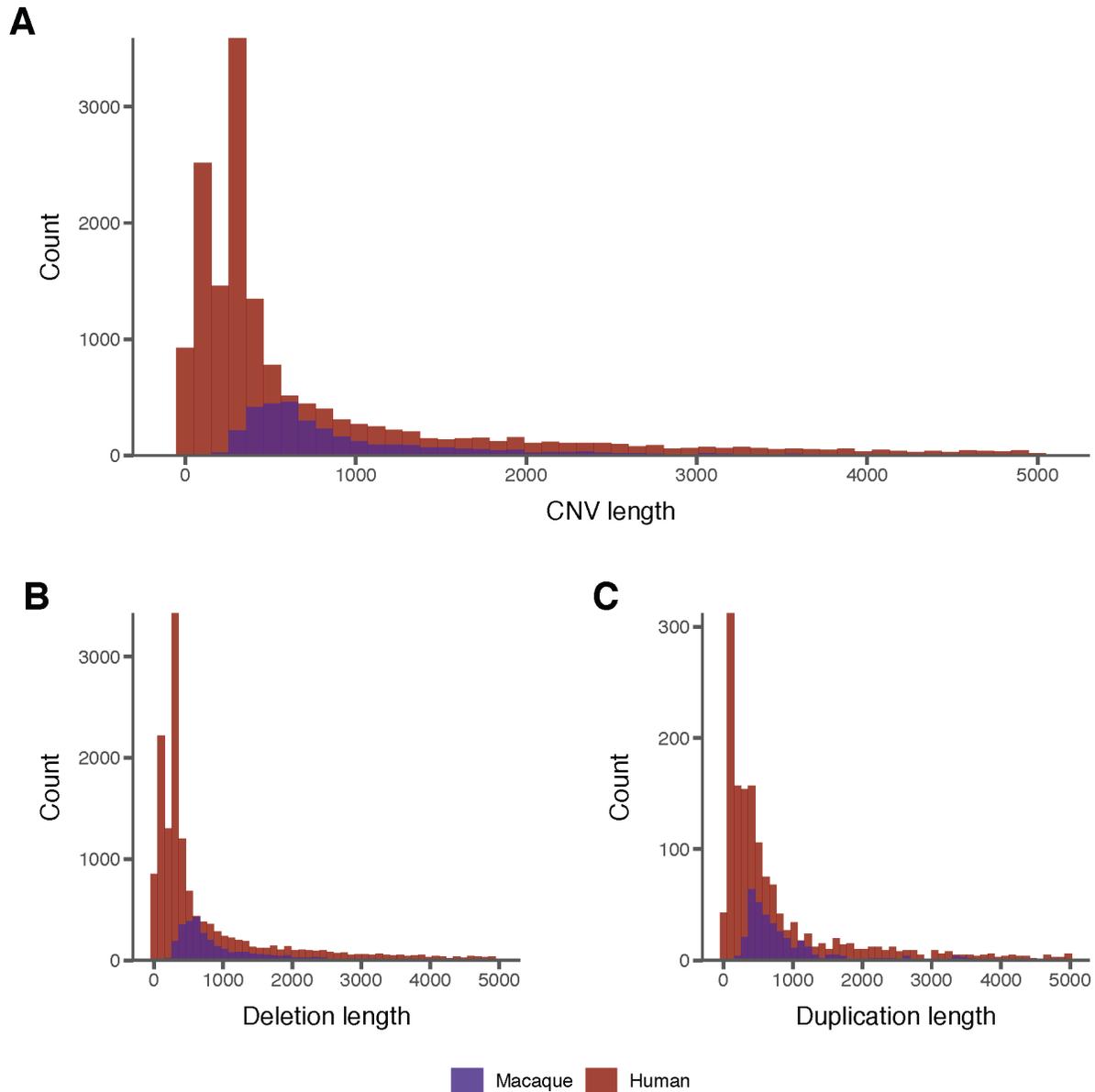
**Figure 4.2:** Locations of identified CNVs on the 22 rhesus macaque chromosomes.

in our sample is heterozygous for 1,317 CNVs that delete 2,804,650 base pairs (bp) and duplicate 471,100 bp.

CNVs in macaques have an average length of 3,519 bases, with duplications (mean length 6,853 bp; min length 138 bp; max length 97,301 bp) being longer than deletions (mean length 3,076 bp; min length 40 bp; max length 98,035 bp). Compared to humans, macaques have longer CNVs on average (Figure 4.4A; Kolmogorov-Smirnov  $D = 0.46$ ,  $p \ll 0.01$ ) and this pattern holds for both deletions (Figure 4.4B; Kolmogorov-Smirnov  $D = 0.48$ ,  $p \ll 0.01$ ) and duplications (Figure 4.4C; Kolmogorov-Smirnov  $D = 0.38$ ,  $p \ll 0.01$ ). It is unclear whether this shift in CNV length distributions between macaques and humans is a true biological phenomenon, which would point to some change in the underlying CNV mechanism, or simply reflects our inability to detect very small variants in macaques. We took every effort to eliminate methodological bias between the macaque CNV calls and the human CNV data set. In their paper, Brandler et al. (2016) [115] use several different CNV calling and genotyping methods. We have restricted our comparisons to CNVs called with the same methods we have employed for the macaque data, namely CNVs called with Lumpy [135] and genotyped with SVtyper [136]. To test the effects of different CNV calling methods and filtering steps, we made



**Figure 4.3:** The number of CNVs is strongly correlated with chromosome length in macaques for (A) all CNVs, (B) deletions, (C) and duplications.



**Figure 4.4:** Length distributions of CNVs shorter than 5000 bases using the full human CNV dataset. Macaque CNVs are longer on average than humans for (A) all CNVs (Kolmogorov-Smirnov  $D = 0.43$ ,  $p \ll 0.01$ ), (B) deletions only (Kolmogorov-Smirnov  $D = 0.46$ ,  $p \ll 0.01$ ), and (C) duplications only (Kolmogorov-Smirnov  $D = 0.32$ ,  $p \ll 0.01$ ).

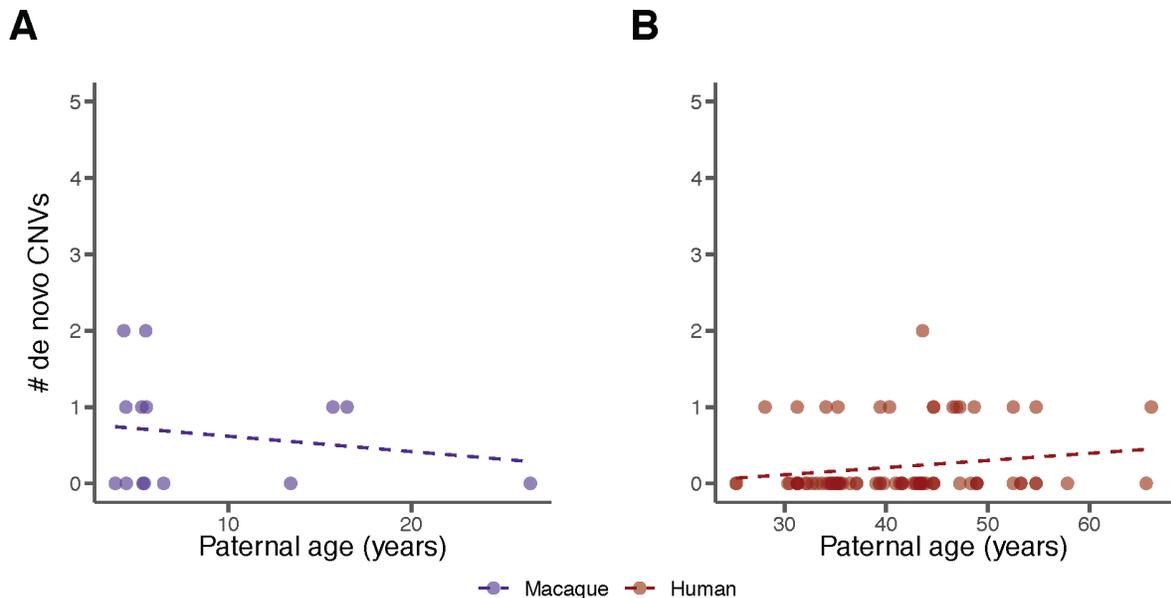
the same comparison between macaque and human CNV lengths while using the full human

dataset and without filtering the macaque CNV calls. Regardless of the partitioning method used, we still observe that macaques have, on average, longer CNVs than humans. Another possible technical explanation for this observation may be the sequencing libraries used in the two datasets: while Brandler et al. sequenced most samples with a read length of 100 bp and insert sizes of 113 bp, the read length of the macaque sequences was larger at 150 bp and an insert size of 128 bp. Although we would expect that this difference in read length would allow the macaque calls to be more sensitive to smaller events, it may play a role in the resulting length of CNV calls. It is also possible that the difference in length distributions is due to a still unidentified technical difference between the two studies.

### 4.3 De novo copy-number variants

In a companion study we have described the rate and pattern of *de novo* single nucleotide variants in rhesus macaques [133]. Here, we identify *de novo* CNVs in the same individuals by looking for CNVs that are unique to the offspring in a trio, as well as being in a heterozygous state. We find only 9 total *de novo* CNVs among our 14 macaque trios, consisting of 7 deletions and 2 duplications. This number of mutations makes the expected number of *de novo* CNVs 0.32 (95% CI 0.13-0.52) per generation per haploid genome. This rate of mutation is similar to that reported in humans [115], which is consistent with the similar genome size between the two species. In contrast, the mutation rate of CNVs in *D. melanogaster* was found to be much lower (0.025 per genome; [57]), though correcting for the ~30-fold smaller size of the fly genome puts the mutation rates on the same order of magnitude per nucleotide.

By considering the age of sires when the offspring of each trio was conceived, we can ask whether the number of *de novo* CNVs increases in older fathers. We find no paternal age effect



in macaques (Figure 4.5A;  $R^2 = 0.033$ , d.f. = 12,  $p = 0.53$ ), though with only nine events our study has low statistical power to detect an increase. However, we also performed the same

**Figure 4.5:** There is no correlation between de novo structural variants in (A) 14 macaque trios or (B) 97 human trios (13 validated + 6 unvalidated CNVs). Each point represents a single trio.

analysis using 19 *de novo* CNVs from human trios [115], and found no increase in the number of mutations in the offspring of older fathers (Figure 4.5B;  $R^2 = 0.032$ , d.f. = 77,  $p = 0.12$ ). Because the rate of new CNVs seems to be very low, increasing the sample size in macaques will increase confidence in our conclusion of a lack of paternal age effect in this species.

#### 4.4 Gene duplications and losses within and between species

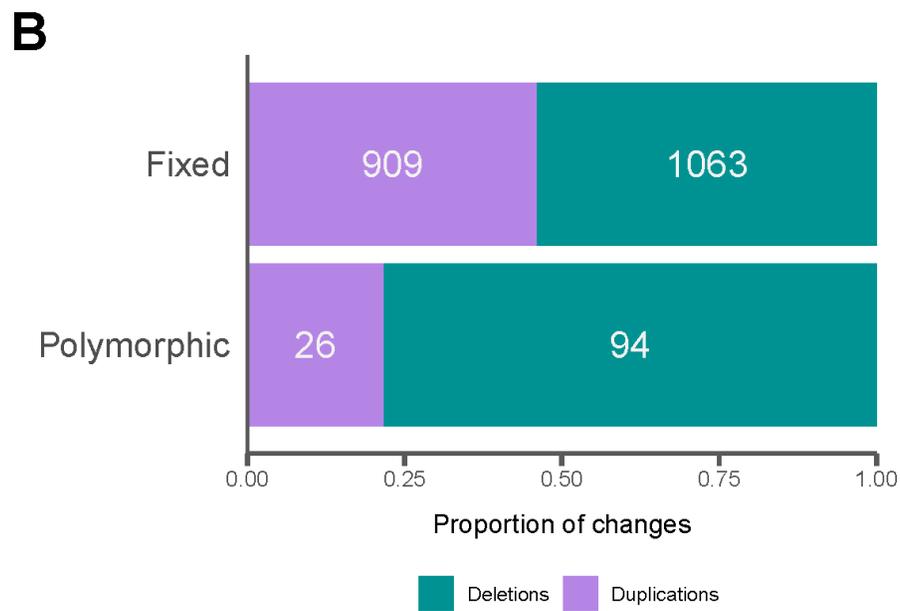
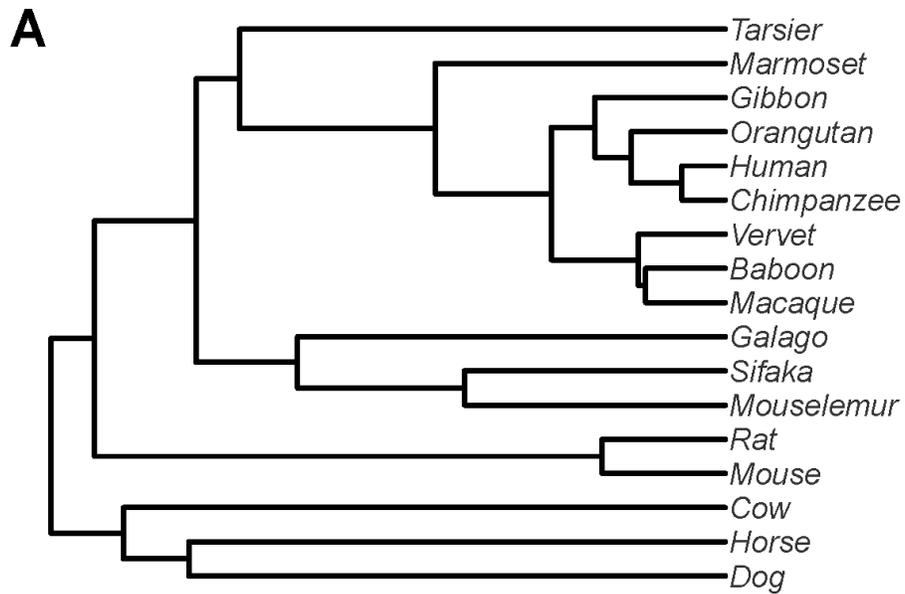
The ultimate fate of structural variants is to either become fixed in a population or to be lost. Genes overlapping CNVs can play a role in this process by conveying fitness benefits or costs depending on their copy-number. We investigated the long-term fate of genes involved in copy-number variation in macaques using gene gains and losses among 17 mammal species (Figure 4.6A). By comparing the number of genes gained and lost between species to the number

of genes overlapping segregating CNVs within macaques, we uncover patterns in short- and long-term evolution of gene copy number.

We find that among the 3,754 CNVs in the macaque samples, 203 of them overlap 338 genes (out of 32,382 total annotated genes); the vast majority occur in intergenic regions. Of the CNVs that overlap a genic region, most span more than one gene (average 1.67 genes per event). However, this is driven by a few CNVs larger than 25 kb that overlap 2-3 genes. CNVs shorter than 25 kb overlap on average only 1.18 genes. This is similar to the pattern observed in humans, where, of CNVs overlapping genes, they include an average of 1.18 genes.

Among the genes within CNVs in macaques, 244 (72%) have been wholly or partially deleted, while 94 (28%) have been wholly or partially duplicated. The ratio of deleted to duplicated genes in macaques is 2.60, which is much lower than the overall ratio of deleted to duplicated regions (7.51). The over-representation of duplicated genic regions compared to non-genic regions has been observed previously in primates [122, 127, 128] and suggests that gene duplication is less costly in the short-term than deletion. The ratio of deleted-to-duplicated genes in macaques is also significantly higher than the ratio in humans of 1.46 ( $\chi^2 = 13.17$ , d.f. = 1,  $p \ll 0.01$ ), possibly because of an increased rate of duplication in the Great apes [37, 38].

To examine the long-term fate of gene duplications and losses, we analyzed copy-number variation in 10,798 gene families across 17 species (Figure 4.6A). Along the branch leading to macaques since their common ancestor with baboons (~11 million years ago), we infer the loss of 1,063 genes and the gain of 909 genes, for a loss-to-gain ratio of 1.17 (Figure 4.6B). This ratio is half that observed among segregating CNVs (see above) but could be biased because different genes may be included in the different annotation sets used. Restricting our CNV analysis to the 19,496 genes present in the gene family analysis, we find a ratio of 3.62 deletions to



**Figure 4.6:** (A) Long-term patterns of gene gain and loss were inferred for macaques by comparing gene copy-numbers among 17 mammal species. (B) Among genes in both the gene family (Fixed) and CNV (Polymorphic) analyses, we find that genes are more likely to be part of polymorphic deletions, and conversely that there is a larger proportion of duplications among fixed differences.

duplications, still significantly higher than the long-term ratio of gene gain to loss (Figure 4.6B;  $\chi^2 = 26.33$ , d.f. = 1,  $p \ll 0.01$ ). Together, these results indicate that, while deletions dominate among *de novo* mutations and segregating CNVs in macaques, the number of genes gained and lost are balanced over time.

## 4.5 Discussion

Copy-number variation can play a key role in disease and evolution [140-142]. Here, we have shown that patterns of copy-number variation in rhesus macaques are largely similar to humans: segregating CNVs in both species are overwhelmingly made up of deletions. CNVs in macaques appear to be on average longer than in humans, though this may also be the result of an unidentified methodological bias. We found that *de novo* CNVs show no correlation with parental age in either species. This is in contrast to single nucleotide variants (SNVs), which have been found to increase with paternal age in both species [35, 71, 133]. The difference between SNVs and CNVs is likely due to the differences in how these mutations arise. SNVs are thought to arise as errors in the DNA replication process during mitosis, or more rarely as unrepaired damage to DNA caused by the environment [27]. For male mammals, both of these processes are ongoing throughout the lifetime, with recurring mitoses occurring during spermatogenesis. However, copy-number variation is thought to arise only during unequal cross-over events during meiosis [140, 143]. Since meiosis only occurs once per generation, we expect no age effects for mutations that arise from it. This expectation is consistent with our present observations in macaques and previous studies in humans [114, 116].

With no age effect for copy-number variation, we expect the rate of new copy-number variants per unit time (i.e. year) to be subject to a classic generation-time effect [23, 144]. The generation-time effect says that species with shorter generation times accumulate more mutations

over time because they experience more germline cell divisions per unit time. This generation-time effect has been found to be dampened for single nucleotide mutations, which are dependent on mitosis, because of ongoing spermatogenesis [145]. However, for structural variants that occur during meiosis this effect should hold. This would mean that we would expect rhesus macaques, with shorter generation times, to have a higher rate of long-term copy number variation than humans.

Contrary to these expectations, the reverse relationship has been observed between species, with humans and chimps having the highest rate of gene gain and loss among primates [37, 146]. One possible explanation for the discrepancy between the expected and observed rate patterns of genic copy-number variation is a difference in selection between the two species. In this scenario, the underlying mutation rates per unit time differ, but studies of genic copy-number variation reveal the combined effects of mutation and selection in shaping the accumulation of change. In support of this is our observation in macaques that deletions make up the majority of polymorphic copy-number events, but fixed gene gains and losses are balanced when comparing gene copy-number evolution between species. This is also further evidence in support of the claim that deletions are under stronger purifying selection than duplications [57, 147, 148].

Taken together, the patterns of copy-number variation we have uncovered will help model this type of mutation and its evolutionary consequences. While the patterns discovered here provide a good basis for this understanding, larger samples in future studies will provide higher confidence. Quantifying *de novo* CNVs may help us refine both disease models and the processes governing the evolution of the mutation rate. Being able to follow new variants from

their introduction as mutations, to variation within populations, and finally to their fixation between species will reveal the evolutionary forces acting at every stage.

#### **4.6 Sequencing and read mapping**

Genomic DNA was isolated from blood samples of 32 rhesus macaques for whole genome sequencing (Illumina Nova-Seq, average 40X average coverage). Reads were mapped to the reference macaque genome (rheMac8.0.1, GenBank assembly accession GCA\_000772875.3) using BWA-MEM version 0.7.12-r1039 [149] to generate a BAM file for each individual. Duplicate reads were identified with Picard MarkDuplicates version 1.105 (<http://broadinstitute.github.io/picard/>) and these reads were excluded from subsequent analyses. All BAM files were sorted and indexed with samtools version 1.9 [150].

Reads that map to the reference with unexpected distances given their insert size (split reads) or orientations (discordant reads) between mate pairs can be used as signals of genomic deletion and duplication. These split and discordant reads were identified in each individual with samtools version 1.9 (-F 1294 for discordant reads) and the extractSplitReads\_BwaMem script included in the Lumpy [135] software package. This resulted in three BAM files for each individual used as input for the CNV calling software listed below: all reads, discordant reads, and split reads.

#### **4.7 Calling copy-number variants (CNVs) in Rhesus macaques**

Copy-number variants were called only on contigs that map to assembled macaque chromosomes. We used a suite of methods in the SpeedSeq software [136] that use patterns of split and discordant read mappings to identify structural variant breakpoints throughout the genome to call CNVs. First, Lumpy [135] was used to find putative breakpoint sites in all 32 macaque individuals. Lumpy uses several pieces of evidence (such as split and discordant reads)

to probabilistically model where breakpoints occur in the genome. CNVs called by Lumpy were genotyped with SVtyper [136], which uses a Bayesian framework much like that used to genotype single nucleotide variants to determine whether CNVs are homozygous or heterozygous. For CNV calling with Lumpy, repetitive regions were masked using the rheMac8 RepeatMasker table from the UCSC table browser ([151]; <http://genome.ucsc.edu/>).

The software SVtools [137] was used to combine the calls from the 32 individuals into a single set. This set was then re-genotyped with SVtyper to obtain information for all CNVs in all samples (even if they were not present in that sample) for filtering. CNV calls were annotated with read depth information using Duphold [138] and finally CNVs were pruned with SVtools such that, among events found to occur within 100 bp, only the event with the highest quality score was retained. CNVs were then annotated as to their overlap with genes by using the UCSC table browser.

#### **4.8 Filtering putative macaque CNVs**

The process for calling CNVs resulted in 157,914 events at 8,515 sites. To reduce the number of false positives, we applied the following filters to our set of CNVs:

1. Removed 83,371 CNVs at 2,615 sites that are present in at least 31 of the 32 individuals. These are most likely events in the reference individual.
2. Removed 4,934 CNVs at 464 sites over 100,000 bp in length.
3. Removed 435 CNVs at 244 sites with a quality score less than 100.
4. Retained only deletions in which the fold-change of read depth for the variant is  $< 0.7$  of the flanking regions. This filter removed 12,763 CNVs at 870 sites.
5. Retained only duplications in which the fold-change of read depth for the variant is  $> 1.3$  of regions with similar GC content. This filter removed 9,954 CNVs at 568 sites.

These filters yield a reduced CNV call-set of 46,457 events at 3,754 sites which was used for all subsequent analyses.

#### **4.9 Identifying *de novo* CNVs and calculating the mutation rate**

From the full set of 3,754 CNVs, we identified *de novo* events as those that occur only in one of the probands of the 14 trios. We required both parents to be homozygous for the reference allele and the child to be heterozygous. For F<sub>1</sub> probands, the *de novo* CNV was allowed to be present in the proband's offspring, as new mutations would be expected to be transmitted roughly half the time. This occurred in 2 out of the 3 F<sub>1</sub> CNVs.

We calculated the CNV mutation rate per generation for a haploid genome by taking the mean number of transmissions in the 14 macaque trios and dividing by 2. Standard error for this rate was calculated by taking the standard deviation of the number of transmissions for the 14 trios divided by the square root of the number of trios times a critical value of 1.96 for the 95% confidence interval.

#### **4.10 Human CNV data**

Human CNVs were downloaded from the supplemental material of Brandler et al. (2016) [115]. This study used 235 individuals in 69 families to look for patterns of *de novo* structural variation among autism patients. Their validated *de novo* mutations along with parental ages were obtained from their supplemental spreadsheet S1 and used for Figure 3B. The entire CNV call-set from their supplemental data S1 was used for all other comparisons. These authors used two methods to call CNVs, Lumpy [135] and ForestSV [152], and two methods to genotype their CNV calls, SVtyper [136] and gtCNV (now known as SV<sup>2</sup>; [153]). We restrict our comparisons to those called with Lumpy and genotyped with SVtyper for consistency with our methods.

#### 4.11 Counting fixed macaque gene duplications and losses

In order to identify genes gained and lost on the macaque lineage we obtained peptides from human, chimpanzee, orangutan, gibbon, macaque, vervet, baboon, marmoset, tarsier, mouse lemur [154], sifaka, galago, rat, mouse, dog, horse, and cow from ENSEMBL 95 [155]. To ensure that each gene was counted only once, we used only the longest isoform of each protein in each species. We then performed an all-vs-all BLAST [156] search on these filtered sequences. The resulting e-values from the search were used as the main clustering criterion for the MCL program to group peptides into gene families [157]. This resulted in 15,662 clusters. We then removed all clusters only present in a single species, resulting in 10,798 gene families. We also obtained an ultrametric tree (Figure 4A) from a previous study [158] for 12 mammal species and added mouse lemur, tarsier, vervet, and galago based on their divergence times from timetree.org [159].

With the gene family data and ultrametric phylogeny as input, we estimated gene gain and loss rates with CAFE v4.2 [160] using a three-rate model, which has been shown to best fit mammalian data [37, 161, 162]. CAFE uses the estimated rates to infer ancestral gene counts and we subsequently counted the number of genes gained and lost in the macaque lineage relative to its ancestor.

## References

1. Wu, C.I., and Li, W.H. (1985). Evidence for Higher Rates of Nucleotide Substitution in Rodents Than in Man. *P Natl Acad Sci USA* 82, 1741-1745.
2. Britten, R.J. (1986). Rates of DNA-Sequence Evolution Differ between Taxonomic Groups. *Science* 231, 1393-1398.
3. Kumar, S., and Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* 99, 803-808.
4. Bromham, L. (2009). Why do species vary in their rate of molecular evolution? *Biol Letters* 5, 401-404.
5. Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, (Cambridge: Cambridge University Press).
6. Kimura, M. (1967). On the evolutionary adjustment of spontaneous mutation rates, Volume 9.
7. Kondrashov, A.S. (1995). Modifiers of mutation-selection balance: general approach and the evolution of mutation rates. *Genetics Research* 66, 53-69.
8. Sniegowski, P.D., Gerrish, P.J., Johnson, T., and Shaver, A. (2000). The evolution of mutation rates: separating causes from consequences. *Bioessays* 22, 1057-1066.
9. Lynch, M. (2010). Evolution of the mutation rate. *Trends Genet* 26, 345-352.
10. Bromham, L. (2011). The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philos T R Soc B* 366, 2503-2513.
11. Lanfear, R., Ho, S.Y.W., Davies, T.J., Moles, A.T., Aarssen, L., Swenson, N.G., Warman, L., Zanne, A.E., and Allen, A.P. (2013). Taller plants have lower rates of molecular evolution. *Nat Commun* 4.
12. Martin, A.P., and Palumbi, S.R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A* 90, 4087-4091.
13. Bleiweiss, R. (1998). Relative-rate tests aid biological causes of molecular evolution in hummingbirds. *Mol Biol Evol* 15, 481-491.
14. Nabholz, B., Glemin, S., and Galtier, N. (2008). Strong variations of mitochondrial mutation rate across mammals--the longevity hypothesis. *Mol Biol Evol* 25, 120-130.
15. Li, W.H., Ellsworth, D.L., Krushkal, J., Chang, B.H., and Hewett-Emmett, D. (1996). Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol* 5, 182-187.

16. Bromham, L., Rambaut, A., and Harvey, P.H. (1996). Determinants of rate variation in mammalian DNA sequence evolution. *Journal of Molecular Evolution* 43, 610-621.
17. Lourenco, J.M., Glemin, S., Chiari, Y., and Galtier, N. (2013). The determinants of the molecular substitution process in turtles. *J Evol Biol* 26, 38-50.
18. Goodman, M. (1985). Rates of molecular evolution: the hominoid slowdown. *Bioessays* 3, 9-14.
19. Yi, S.V. (2013). Morris Goodman's hominoid rate slowdown: The importance of being neutral. *Molecular Phylogenetics and Evolution* 66, 569-574.
20. Elango, N., Thomas, J.W., Yi, S.V., and Progra, N.C.S. (2006). Variable molecular clocks in hominoids. *P Natl Acad Sci USA* 103, 1370-1375.
21. Scally, A., and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution (vol 13, pg 745, 2012). *Nature Reviews Genetics* 13, 824-824.
22. Goodman, M. (1962). Evolution of the immunologic species specificity of human serum proteins. *Hum Biol* 34, 104-150.
23. Laird, C.D., McConaughy, B.L., and McCarthy, B.J. (1969). Rate of fixation of nucleotide substitutions in evolution. *Nature* 224, 149-154.
24. Li, W. (1997). *Molecular evolution*, (Sunderland: Sinauer Associates Incorporated).
25. Drost, J.B., and Lee, W.R. (1995). Biological basis of germline mutation: comparisons of spontaneous germline mutation-rates among *Drosophila*, mouse, and human. *Environmental and Molecular Mutagenesis* 25, 48-64.
26. Haldane, J.B. (1947). The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann Eugen* 13, 262-271.
27. Crow, J.F. (2000). The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* 1, 40-47.
28. Hasegawa, M., Kishino, H., and Yano, T. (1987). Man's place in Hominoidea as inferred from molecular clocks of DNA. *J Mol Evol* 26, 132-147.
29. Shimmin, L.C., Chang, B.H.J., and Li, W.H. (1993). Male-Driven Evolution of DNA-Sequences. *Nature* 362, 745-747.
30. Berlin, S., Brandstrom, M., Backstrom, N., Axelsson, E., Smith, N.G., and Ellegren, H. (2006). Substitution rate heterogeneity and the male mutation bias. *J Mol Evol* 62, 226-233.

31. Bartosch-Harlid, A., Berlin, S., Smith, N.G., Moller, A.P., and Ellegren, H. (2003). Life history and the male mutation bias. *Evolution* 57, 2398-2406.
32. Presgraves, D.C., and Yi, S.V. (2009). Doubts about complex speciation between humans and chimpanzees. *Trends Ecol Evol* 24, 533-540.
33. Lynch, M. (2007). *The Origins of Genome Architecture*, (Sunderland, Mass.: Sinauer Associates).
34. Kim, S.H., Elango, N., Warden, C., Vigoda, E., and Yi, S.V. (2006). Heterogeneous genomic molecular clocks in primates. *PLoS Genet* 2, e163.
35. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488, 471-475.
36. Sun, J.X., Helgason, A., Masson, G., Ebenesersdottir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., et al. (2012). A direct characterization of human mutation based on microsatellites. *Nat Genet* 44, 1161-1165.
37. Hahn, M.W., Demuth, J.P., and Han, S.G. (2007). Accelerated rate of gene gain and loss in primates. *Genetics* 177, 1941-1949.
38. Marques-Bonet, T., Kidd, J.M., Ventura, M., Graves, T.A., Cheng, Z., Hillier, L.W., Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L.A., et al. (2009). A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* 457, 877-881.
39. Fenner, J.N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* 128, 415-423.
40. Hemmer, H. (2013). Estimation of Basic Life History Data of Fossil Hominoids. In *Handbook of Paleoanthropology: Vol I: Principles, Methods and Approaches Vol II: Primate Evolution and Human Origins Vol III: Phylogeny of Hominids*, W. Henke and I. Tattersall, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 1-28.
41. Bongaarts, J. (2001). Fertility and reproductive preferences in post-transitional societies. *Popul Dev Rev* 27, 260-281.
42. Svensson, A.C., Abel, K., Dalman, C., and Magnusson, C. (2011). Implications of Advancing Paternal Age: Does It Affect Offspring School Performance? *Plos One* 6.
43. Latta, L.C., Morgan, K.K., Weaver, C.S., Allen, D., Schaack, S., and Lynch, M. (2013). Genomic Background and Generation Time Influence Deleterious Mutation Rates in *Daphnia*. *Genetics* 193, 539-+.
44. Nesse, R.M., and Williams, G.C. (2012). *Why we get sick: The new science of Darwinian medicine*, (Vintage).

45. Stearns, S.C., and Koella, J.C. (2008). *Evolution in health and disease*, (Oxford University Press).
46. Lynch, M. (2008). The cellular, developmental and population-genetic determinants of mutation-rate evolution. *Genetics* *180*, 933-943.
47. Promislow, D.E.L. (1994). DNA-Repair and the Evolution of Longevity - a Critical Analysis. *J Theor Biol* *170*, 291-300.
48. Welch, J.J., Bininda-Emonds, O.R.P., and Bromham, L. (2008). Correlates of substitution rate variation in mammalian protein-coding sequences. *Bmc Evol Biol* *8*.
49. Marcon, E., and Moens, P.B. (2005). The evolution of meiosis: recruitment and modification of somatic DNA-repair proteins. *Bioessays* *27*, 795-808.
50. Galetzka, D., Weis, E., Kohlschmidt, N., Bitz, O., Stein, R., and Haaf, T. (2007). Expression of somatic DNA repair genes in human testes. *J Cell Biochem* *100*, 1232-1239.
51. Crow, J.F. (1986). Population consequences of mutagenesis and antimutagenesis. *Basic Life Sci* *39*, 519-530.
52. Conrad, D.F., Keebler, J.E.M., DePristo, M.A., Lindsay, S.J., Zhang, Y.J., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., et al. (2011). Variation in genome-wide mutation rates within and between human families. *Nat Genet* *43*, 712-U137.
53. Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C.R., Dopman, E.B., Dickinson, W.J., Okamoto, K., Kulkarni, S., and Hartl, D.L. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences* *105*, 9272-9277.
54. Denver, D.R., Dolan, P.C., Wilhelm, L.J., Sung, W., Lucas-Lledo, J.I., Howe, D.K., Lewis, S.C., Okamoto, K., Thomas, W.K., Lynch, M., et al. (2009). A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc Natl Acad Sci U S A* *106*, 16310-16314.
55. Keightley, P.D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M.L. (2009). Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* *19*, 1195-1201.
56. Ossowski, S., Schneeberger, K., Lucas-Lledo, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* *327*, 92-94.
57. Schrider, D.R., Houle, D., Lynch, M., and Hahn, M.W. (2013). Rates and Genomic Consequences of Spontaneous Mutational Events in *Drosophila melanogaster*. *Genetics* *194*, 937-+.

58. Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics* 26, 345-352.
59. Drake, J.W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences of the United States of America* 88, 7160-7164.
60. Kondrashov, A.S. (1995). Modifiers of mutation-selection balance: general approach and the evolution of mutation rates. *Genetical Research* 66, 53-69.
61. Lynch, M. (2006). The origins of eukaryotic gene structure. *Molecular Biology and Evolution* 23, 450-468.
62. Damuth, J. (1981). Population-density and body size in mammals. *Nature* 290, 699-700.
63. Weinberg, W. (1912). Zur vererbung des zwergwuchses. *Arch Rassen-u Gesell Biolog* 9, 710-718.
64. Crow, J.F. (2000). The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics* 1, 40-47.
65. Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., et al. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* 151, 1431-1442.
66. Wong, W.S., Solomon, B.D., Bodian, D.L., Kothiyal, P., Eley, G., Huddleston, K.C., Baker, R., Thach, D.C., Iyer, R.K., Vockley, J.G., et al. (2016). New observations on maternal age effect on germline de novo mutations. *Nature Communications* 7, 10486.
67. Goldmann, J.M., Wong, W.S., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A.B., Glusman, G., Vissers, L.E., Hoischen, A., Roach, J.C., et al. (2016). Parent-of-origin-specific signatures of *de novo* mutations. *Nature Genetics* 48, 935-939.
68. Besenbacher, S., Sulem, P., Helgason, A., Helgason, H., Kristjansson, H., Jonasdottir, A., Jonasdottir, A., Magnusson, O.T., Thorsteinsdottir, U., Masson, G., et al. (2016). Multi-nucleotide *de novo* mutations in humans. *PLoS Genetics* 12, e1006315.
69. Girard, S.L., Bourassa, C.V., Lemieux Perreault, L.P., Legault, M.A., Barhdadi, A., Ambalavanan, A., Brendgen, M., Vitaro, F., Noreau, A., Dionne, G., et al. (2016). Paternal Age Explains a Major Portion of De Novo Germline Mutation Rate Variability in Healthy Individuals. *Plos One* 11, e0164212.
70. Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Turki, S.A., Dominiczak, A., Morris, A., Porteous, D., Smith, B., et al. (2016). Timing, rates and spectra of human germline mutation. *Nature Genetics* 48, 126-133.
71. Jonsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A., et al. (2017). Parental

- influence on human germline *de novo* mutations in 1,548 trios from Iceland. *Nature* *549*, 519-522.
72. Goriely, A. (2016). Decoding germline *de novo* point mutations. *Nature Genetics* *48*, 823-824.
  73. Segurel, L., Wyman, M.J., and Przeworski, M. (2014). Determinants of mutation rate variation in the human germline. *Annual Reviews in Genomics and Human Genetics* *15*, 47-70.
  74. Venn, O., Turner, I., Mathieson, I., de Groot, N., Bontrop, R., and McVean, G. (2014). Strong male bias drives germline mutation in chimpanzees. *Science* *344*, 1272-1275.
  75. Dixson, A.F., Gardner, J.S., and Bonney, R.C. (1980). Puberty in the male owl monkey (*Aotus trivirgatus griseimembra*): A study of physical and hormonal development. *International Journal of Primatology* *1*, 129-139.
  76. Rowe, N. (1996). *The Pictorial Guide to the Living Primates*, (East Hampton, N.Y.: Pogonias Press).
  77. Huck, M., Rotundo, M., and Fernandez-Duque, E. (2011). Growth and development in wild owl monkeys (*Aotus azarai*) of Argentina. *International Journal of Primatology* *32*, 1133-1152.
  78. Schrider, D.R., Hourmozdi, J.N., and Hahn, M.W. (2011). Pervasive multinucleotide mutational events in eukaryotes. *Current Biology* *21*, 1051-1054.
  79. Scally, A. (2016). The mutation rate in human evolution and demographic inference. *Current Opinions in Genetics and Development* *41*, 36-43.
  80. Thomas, G.W.C., and Hahn, M.W. (2014). The human mutation rate is increasing, even as it slows. *Molecular Biology and Evolution* *31*, 253-257.
  81. Dixson, A., and Anderson, M. (2001). Sexual selection and the comparative anatomy of reproduction in monkeys, apes, and human beings. *Annu Rev Sex Res* *12*, 121-144.
  82. Presgraves, D.C., and Yi, S.V. (2009). Doubts about complex speciation between humans and chimpanzees. *Trends Ecol Evol* *24*, 533-540.
  83. Moorjani, P., Amorim, C.E.G., Arndt, P.F., and Przeworski, M. (2016). Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences of the United States of America* *113*, 10607-10612.
  84. Schmid-Siegert, E., Sarkar, N., Iseli, C., Calderon, S., Gouhier-Darimont, C., Chrast, J., Cattaneo, P., Schutz, F., Farinelli, L., Pagni, M., et al. (2017). Low number of fixed somatic mutations in a long-lived oak tree. *Nature Plants* *3*, 926-929.

85. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997.
86. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297-1303.
87. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43, 491-498.
88. Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30, 2843-2851.
89. Besenbacher, S., Liu, S., Izarzugaza, J.M., Grove, J., Belling, K., Bork-Jensen, J., Huang, S., Als, T.D., Li, S., Yadav, R., et al. (2015). Novel variation and *de novo* mutation rates in population-wide *de novo* assembled Danish trios. *Nature Communications* 6, 5969.
90. Smeds, L., Mugal, C.F., Qvarnstrom, A., and Ellegren, H. (2016). High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genetics* 12, e1006044.
91. Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Genome of the Netherlands, C., van Duijn, C.M., Swertz, M., Wijmenga, C., et al. (2015). Genome-wide patterns and properties of *de novo* mutations in humans. *Nature Genetics* 47, 822-826.
92. Amster, G., and Sella, G. (2016). Life history effects on the molecular clock of autosomes and sex chromosomes. *Proceedings of the National Academy of Sciences of the United States of America* 113, 1588-1593.
93. Nielsen, C.T., Skakkebaek, N.E., Richardson, D.W., Darling, J.A.B., Hunter, W.M., Jorgensen, M., Nielsen, A., Ingerslev, O., Keiding, N., and Muller, J. (1986). Onset of the release of spermatozoa (spermarche) in boys in relation to age, testicular growth, pubic hair, and height. *Journal of Clinical Endocrinology & Metabolism* 62, 532-535.
94. Marson, J., Meuris, S., Cooper, R.W., and Jouannet, P. (1991). Puberty in the male chimpanzee: progressive maturation of semen characteristics. *Biology of Reproduction* 44, 448-455.
95. Tatsumoto, S., Go, Y., Fukuta, K., Noguchi, H., Hayakawa, T., Tomonaga, M., Hirai, H., Matsuzawa, T., Agata, K., and Fujiyama, A. (2017). Direct estimation of *de novo* mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing. *Scientific Reports* 7, 13561.
96. Heller, C.G., and Clermont, Y. (1963). Spermatogenesis in man: an estimate of its duration. *Science* 140, 184-186.

97. Scally, A. (2016). Mutation rates and the evolution of germline structure. *Philos T R Soc B* 371.
98. Plant, T.M. (2010). Undifferentiated primate spermatogonia and their endocrine control. *Trends in Endocrinology and Metabolism* 21, 488-495.
99. Derooij, D.G., Vanalphen, M.M.A., and Vandekant, H.J.G. (1986). Duration of the cycle of the seminiferous epithelium and its stages in the rhesus-monkey (*Macaca mulatta*). *Biology of Reproduction* 35, 587-591.
100. Hubisz, M.J., Lin, M.F., Kellis, M., and Siepel, A. (2011). Error and error mitigation in low-coverage genome assemblies. *Plos One* 6, e17034.
101. Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res* 19, 922-933.
102. Prosdocimi, F., Linard, B., Pontarotti, P., Poch, O., and Thompson, J.D. (2012). Controversies in modern evolutionary biology: the imperative for error detection and quality control. *BMC Genomics* 13, 5.
103. Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G.H., and Graur, D. (2009). Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol* 1, 114-118.
104. Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3, e3376.
105. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18, 1851-1858.
106. Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T.D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biol* 14, R47.
107. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-1303.
108. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15, 356.
109. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res* 12, 996-1006.
110. Wu, M., Kostyun, J.L., Hahn, M.W., and Moyle, L.C. (2018). Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. *Mol Ecol* 27, 3301-3316.

111. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29, 644-652.
112. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760.
113. Thomas, G.W.C., Wang, R.J., Puri, A., Harris, R.A., Raveendran, M., Hughes, D.S.T., Murali, S.C., Williams, L.E., Doddapaneni, H., Muzny, D.M., et al. (2018). Reproductive Longevity Predicts Mutation Rates in Primates. *Curr Biol* 28, 3193-3197 e3195.
114. Kloosterman, W.P., Francioli, L.C., Hormozdiari, F., Marschall, T., Hehir-Kwa, J.Y., Abdellaoui, A., Lameijer, E.W., Moed, M.H., Koval, V., Renkens, I., et al. (2015). Characteristics of de novo structural changes in the human genome. *Genome Res* 25, 792-801.
115. Brandler, W.M., Antaki, D., Gujral, M., Noor, A., Rosanio, G., Chapman, T.R., Barrera, D.J., Lin, G.N., Malhotra, D., Watts, A.C., et al. (2016). Frequency and Complexity of De Novo Structural Mutation in Autism. *Am J Hum Genet* 98, 667-679.
116. MacArthur, J.A., Spector, T.D., Lindsay, S.J., Mangino, M., Gill, R., Small, K.S., and Hurles, M.E. (2014). The rate of nonallelic homologous recombination in males is highly variable, correlated between monozygotic twins and independent of age. *PLoS Genet* 10, e1004195.
117. Werling, D.M., Brand, H., An, J.Y., Stone, M.R., Zhu, L., Glessner, J.T., Collins, R.L., Dong, S., Layer, R.M., Markenscoff-Papadimitriou, E., et al. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* 50, 727-736.
118. Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J., and Eichler, E.E. (2010). De novo rates and selection of large copy number variation. *Genome Res* 20, 1469-1481.
119. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. *Science* 316, 445-449.
120. Gazave, E., Darre, F., Morcillo-Suarez, C., Petit-Marty, N., Carreno, A., Marigorta, U.M., Ryder, O.A., Blancher, A., Rocchi, M., Bosch, E., et al. (2011). Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res* 21, 1626-1639.
121. Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., Genomes, P., et al. (2010). Diversity of human copy number variation and multicopy genes. *Science* 330, 641-646.

122. Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., et al. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* 2, E207.
123. Jiang, Z., Tang, H., Ventura, M., Cardone, M.F., Marques-Bonet, T., She, X., Pevzner, P.A., and Eichler, E.E. (2007). Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* 39, 1361-1368.
124. Perry, G.H., Tchinda, J., McGrath, S.D., Zhang, J., Picker, S.R., Caceres, A.M., Iafrate, A.J., Tyler-Smith, C., Scherer, S.W., Eichler, E.E., et al. (2006). Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A* 103, 8006-8011.
125. Perry, G.H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A.S., Hyland, C., Stone, A.C., Hurles, M.E., Tyler-Smith, C., et al. (2008). Copy number variation and evolution in humans and chimpanzees. *Genome Res* 18, 1698-1710.
126. Gokcumen, O., Babb, P.L., Iskow, R.C., Zhu, Q., Shi, X., Mills, R.E., Ionita-Laza, I., Vallender, E.J., Clark, A.G., Johnson, W.E., et al. (2011). Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biol* 12, R52.
127. Dumas, L., Kim, Y.H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J.R., and Sikela, J.M. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res* 17, 1266-1277.
128. Sudmant, P.H., Huddleston, J., Catacchio, C.R., Malig, M., Hillier, L.W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R.E., Persengiev, S., et al. (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* 23, 1373-1382.
129. Lee, A.S., Gutierrez-Arcelus, M., Perry, G.H., Vallender, E.J., Johnson, W.E., Miller, G.M., Korbel, J.O., and Lee, C. (2008). Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* 17, 1127-1136.
130. Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. *Nat Rev Genet* 16, 172-183.
131. Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 6, S13-20.
132. Zhang, L., Bai, W., Yuan, N., and Du, Z. (2019). Comprehensively benchmarking applications for detecting copy number variation. *PLoS Comput Biol* 15, e1007069.
133. Wang, R.J. ((in prep)). No indication that germline mutation accumulation is associated with lowered child sociability in rhesus macaque.

134. Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., Venter, J.C., Wilson, R.K., et al. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science* *316*, 222-234.
135. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* *15*, R84.
136. Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., and Hall, I.M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* *12*, 966-968.
137. Larson, D.E., Abel, H.J., Chiang, C., Badve, A., Das, I., Eldred, J.M., Layer, R.M., and Hall, I.M. (2018). svtools: population-scale analysis of structural variation. *bioRxiv*.
138. Pedersen, B.S., and Quinlan, A.R. (2019). Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *GigaScience* *8*.
139. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* *11*, 1005-1017.
140. Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* *10*, 451-481.
141. Girirajan, S., Campbell, C.D., and Eichler, E.E. (2011). Human copy number variation and complex genetic disease. *Annu Rev Genet* *45*, 203-226.
142. Eichler, E.E., Nickerson, D.A., Altshuler, D., Bowcock, A.M., Brooks, L.D., Carter, N.P., Church, D.M., Felsenfeld, A., Guyer, M., Lee, C., et al. (2007). Completing the map of human genetic variation. *Nature* *447*, 161-165.
143. Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nat Rev Genet* *10*, 551-564.
144. Wu, C.I., and Li, W.H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci U S A* *82*, 1741-1745.
145. Thomas, G.W., and Hahn, M.W. (2014). The human mutation rate is increasing, even as it slows. *Mol Biol Evol* *31*, 253-257.
146. Demuth, J.P., De Bie, T., Stajich, J.E., Cristianini, N., and Hahn, M.W. (2006). The evolution of mammalian gene families. *PLoS One* *1*, e85.
147. Schrider, D.R., and Hahn, M.W. (2010). Gene copy-number polymorphism in nature. *Proc Biol Sci* *277*, 3213-3221.

148. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* *38*, 75-81.
149. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv 1303*, 3997.
150. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* *25*, 2078-2079.
151. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* *32*, D493-496.
152. Michaelson, J.J., and Sebat, J. (2012). forestSV: structural variant discovery through statistical learning. *Nat Methods* *9*, 819-821.
153. Antaki, D., Brandler, W.M., and Sebat, J. (2018). SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics* *34*, 1774-1777.
154. Larsen, P.A., Harris, R.A., Liu, Y., Murali, S.C., Campbell, C.R., Brown, A.D., Sullivan, B.A., Shelton, J., Brown, S.J., Raveendran, M., et al. (2017). Hybrid de novo genome assembly and centromere characterization of the gray mouse lemur (*Microcebus murinus*). *BMC Biol* *15*, 110.
155. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C.G., et al. (2018). Ensembl 2018. *Nucleic Acids Res* *46*, D754-D761.
156. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* *215*, 403-410.
157. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* *30*, 1575-1584.
158. Rogers, J., Raveendran, M., Harris, R.A., Mailund, T., Leppala, K., Athanasiadis, G., Schierup, M.H., Cheng, J., Munch, K., Walker, J.A., et al. (2019). The comparative genomics and complex population history of *Papio* baboons. *Sci Adv* *5*, eaau6947.
159. Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* *34*, 1812-1819.
160. Han, M.V., Thomas, G.W., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* *30*, 1987-1997.

161. Marques-Bonet, T., Girirajan, S., and Eichler, E.E. (2009). The origins and impact of primate segmental duplications. *Trends Genet* 25, 443-454.
162. Carbone, L., Harris, R.A., Gnerre, S., Veeramah, K.R., Lorente-Galdos, B., Huddleston, J., Meyer, T.J., Herrero, J., Roos, C., Aken, B., et al. (2014). Gibbon genome and the fast karyotype evolution of small apes. *Nature* 513, 195-201.

# Gregg Thomas

[greggwct@gmail.com](mailto:greggwct@gmail.com)  
<https://gwct.github.io>

## EDUCATION

Doctor of Philosophy, Dual Degree:  
Informatics with a Bioinformatics track and Ecology, Evolution, and Behavior  
Indiana University  
Bloomington, IN  
July, 2019

Master of Science in Bioinformatics  
Indiana University  
Bloomington, IN  
May, 2013

Bachelor of Science in Biology  
Purdue University  
West Lafayette, IN  
May, 2010

## PUBLICATIONS

1. **Thomas GWC**, Dohmen E, Hughes ST, Murali SC, Poelechau M, Glastad K, ..., Chipman AD, Waterhouse RM, Bornberg-Bauer E, Hahn MW, Richards S. (in review). The genomic basis of Arthropod diversity.
2. Bentz A, **Thomas GWC**, Rusch DB, Rosvall KA. (in review). *De novo* transcriptome assembly of the tree swallow (*Tachycineta bicolor*): tissue-specific expression profiles and positive selection analysis.
3. **Thomas GWC** and Hahn MW. 2019. Referee: reference genome quality scores. *Genome Biology and Evolution*. 11(5):1483-1486.
4. Rogers J, ..., **Thomas GWC**, ..., Jolly CJ, Gibbs RA, Worley KC. 2019. The comparative genomics and complex population history of *Papio* baboons. *Science Advances*. 5(1).
5. Da Lage J-L, **Thomas GWC**, Bonneau M, Courtier-Orgogozo V. 2019. Evolution of salivary glue genes in *Drosophila* species. *BMC Evolutionary Biology*. 19:36.
6. Prost S, Armstrong EE, Nylander J, **Thomas GWC**, Suh A, Petersen B, Dalen L, Benz BW, Blom MPK, Palkopoulou E, Ericson PGP, Irestedt M. 2019. Comparative analyses identify genomic features potentially involved in the evolution of birds-of-paradise. *GigaScience*.

7. **Thomas GWC**, Wang RJ, Puri A, Harris RA, Raveendran, Hughes DST, Murali SC, Williams LE, Doddapaneni, Muzny DM, Gibbs RA, Abee CR, Galinski MR, Worley KC, Rogers J, Radivojac P, Hahn MW. 2018. Reproductive longevity predicts mutation rates in primates. *Current Biology*. 28(19):3193-3197.
8. Warren WC, García-Pérez R, ..., **Thomas GWC**, ..., Scharl M. 2018. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. *Nature Ecology and Evolution*. 2:669-679.
9. Schoville SD, Chen YH, ..., **Thomas GWC**, ..., Richards S. 2018. A model species for agricultural pest genomics: the genome of the Colorado potato beetle, *Leptinotarsa decemlineata* (Coleoptera: Chrysomelidae). *Scientific Reports*. 8(1931).
10. Palesch D, Bosinger SE, ..., **Thomas GWC**, ..., Silvestri G. 2018. Sooty mangabey genome sequence provides insight into AIDS resistance in a natural SIV host. *Nature*. 553:77-81.
11. **Thomas GWC**, Ather SA, and Hahn MW. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Systematic Biology*. 66(6):1007-1018.
12. **Thomas GWC**, Hahn MW, and Hahn Y. 2017. The effects of increasing the number of taxa on inferences of molecular convergence. *Genome Biology and Evolution*. 9(1):213-221.
13. Warren WC, ..., **Thomas GWC**, ..., Freimer NB. 2015. The genome of the vervet (*Chlorocebus aethiops sabaues*). *Genome Research*. 25(12):1921-1933.
14. **Thomas GWC** and Hahn MW. 2015. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Molecular Biology and Evolution*. 32(5):1232-1236.
15. Foote AD, Liu Y, **Thomas GWC**, Vinař T, ..., Gibbs RA. 2015. Convergent evolution of the genomes of marine mammals. *Nature Genetics*. 47(3):272-275.
16. Neafsey DE, Waterhouse RM, ..., **Thomas GWC**, ..., Besansky NJ. 2014. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science*. 347.
17. Montague MJ, ..., **Thomas GWC**, ... Warren WC. 2014. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc Natl Acad Sci USA*. 111(48):17230-17235.
18. Carbone L, ... **Thomas GWC**, ..., Gibbs RA. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature*. 513:195-201.

19. **Thomas GWC** and Hahn MW. 2014. The human mutation rate is increasing, even as it slows. *Molecular Biology and Evolution*. 31(2):253-257.
20. Han MV, **Thomas GWC**, Lugo-Martinez J, and Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution*. 30(8):1987-1997.

## PRESENTATIONS

1. **Reproductive longevity predicts mutation rates in primates**  
Population, Evolutionary, and Quantitative Genetics Conference,  
Madison, WI  
Platform talk  
May 19, 2018
2. **The evolution of the genes and genomes of 76 arthropod species**  
Evolution Meeting, Portland, OR  
Regular talk  
June 26, 2017
3. **The evolution of the genes and genomes of 76 arthropod species**  
Arthropod Genomics Symposium, Notre Dame University, South  
Bend, IN  
**Invited talk**  
June 9, 2017
4. **Gene-tree reconciliation with MUL-trees for polyploidy analysis**  
Evolution Meeting, Austin, TX  
Regular talk  
June 19, 2016
5. **Accounting for sequencing error in phylogenetics**  
Society of Systematic Biologists, University of Michigan, Ann Arbor,  
MI  
Lightning Talk  
May 21, 2015
6. **Inferring molecular convergence from genomic data**  
Midwest Ecology and Evolution Conference, Indiana University,  
Bloomington, IN  
Contributed talk  
March 28, 2015
7. **Convergent evolution of the genomes of marine mammals**  
Society for Molecular Biology and Evolution, San Juan, Puerto Rico  
Contributed talk  
June 12, 2014

## RESEARCH EXPERIENCE

### *Research Assistant*

Laboratory of Matthew Hahn  
School of Informatics, Computing, and Engineering  
Department of Biology  
Indiana University, Bloomington, IN

2012 – present

- Developed a method to estimate genome assembly and annotation error from gene count data using CAFE's error model function (caferror).
- Studied patterns of convergent evolution in marine mammals and echolocating mammals.
- Devised a method to infer the presence and mode of polyploidy from gene tree topologies (GRAMPA).
- Modeled and observed mutation rate patterns in primates, including single nucleotide mutations and structural variants, by sequencing families of owl monkeys and macaques.
- Led the comparative phylogenetic portion of the i5K pilot project which involved analyzing the genomes of 76 arthropods.
- Wrote software to annotate genomes with quality scores (Referee).
- Participated in several collaborations by performing comparative analyses, such as phylogeny reconstruction and assessment, gene family analysis, and positive selection scans.

## TEACHING EXPERIENCE

### *Student Mentor*

School of Informatics, Computing, and Engineering  
Department of Biology  
Indiana University, Bloomington, IN

2014 –2019

Provided guidance to high school and undergraduate students in conceptualizing evolution by involving them in various computational projects, providing a basis in programming, data analysis, and scholarship.

- Jelena Nguyen, Indiana University: CEWiT Research Experience for Undergraduate Women (Fall 2018 to Spring 2019).
- Arthi Puri, Indiana University: Computer Science Independent Study (Fall 2017 to Spring 2018).
- S. Hussain Ather, Indiana University: Computer Science Independent Study (Spring 2016 to Spring 2017).
- Nana Addo, Indiana University: Jim Holland Summer Science Research Program (Summer 2014).

### *Teaching Assistant*

2011 – 2016

School of Informatics, Computing, and Engineering  
Department of Biology  
Indiana University, Bloomington, IN

Taught lab sessions, led class discussions, graded assignments, and met with students individually to assist them.

- INFO-I211: Information Infrastructure (Fall 2014, Spring 2016).
- BIOL-Z620/INFO-I590: SNP Discovery and Population Genetics (Fall 2014).
- INFO-I308: Information Representation (Fall 2011, Spring 2012).

## PROFESSIONAL SERVICE

### *Graduate Student Advisor*

Indiana University Bioinformatics Club  
Indiana University, Bloomington, IN

2012 –2014

Served as a co-founding member and treasurer (2012 only) to raise awareness of bioinformatics and associated opportunities for undergraduate and graduate students by facilitating group projects and discussions, tours, and social events.

### *Reviewer*

- Molecular Biology and Evolution
- New Phytologist
- Pacific Symposium on Biocomputing, 2019
- PLoS One

## AWARDS

### *Genetics, Cellular, and Molecular Sciences Training Grant*

Department of Biology  
Indiana University, Bloomington, IN

2014 –2015

## SOFTWARE

### **Referee: Reference genome quality scores**

<https://gwct.github.io/referee>

- This software uses genotype likelihoods from reads mapped back to their assembly to calculate a quality score for every position in the assembled genome.

### ***Drosophila* 25 species phylogeny**

<http://dx.doi.org/10.6084/m9.figshare.5450602>

- As part of a larger project, I inferred the phylogeny of 25 *Drosophila* species and

published it standalone on FigShare as a resource for others to use.

### **GRAMPA: Gene-tree Reconciliation Algorithm with MUL-trees for Polyploid Analysis**

<https://gwct.github.io/grampa.html>

- Given a singly-labeled species topology and a set of corresponding gene-trees, this software can infer if any whole genome duplications have occurred and, if so, infer the mode of polyploidization and the placement on the phylogeny.

### **I5K Phylogenomics Website**

<https://i5k.gitlab.io/ArthroFam>

- With the vast amount of data involved in the i5K pilot project, I developed this website to organize and share the phylogenetic and comparative results with colleagues.

### **GWCT: Genome-Wide Convergence Tester**

<https://github.com/gwct/gwct>

- Software written to count convergent, divergent, and unique substitutions in sequence data.

### **caferror**

<https://hahnlab.github.io/CAFE/>

- Part of CAFE version 3, I wrote this program to use CAFE's error modeling function to estimate genome assembly and annotation error.