Variation in germline mutagenesis in humans and other great apes

Michael Goldberg

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Kelley Harris, Chair

Richard McLaughlin

Philip Green

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Variation in germline mutagenesis in humans and other great apes

Michael Goldberg

Chair of the Supervisory Committee:
Kelley Harris, Assistant Professor of Genome Sciences
Department of Genome Sciences

All heritable variation begins with damage or copying mistakes affecting the DNA of germline, eggs, or embryos. Different DNA motifs and loci can have different mutation rates, and these rates can vary or change over time and genomic space. Here, I present work documenting how the spectrum of mutability of three-base-pair motifs has evolved rapidly during great ape diversification. I show that even as ape mutation spectra diverged from each other, there exists spatial mutation spectrum variation between different genomic regions that is largely conserved across all species. The mutational process can be deconvoluted into a mixture of fast-evolving signatures with uniform spatial distributions and conserved signatures that target specific regions. I also present work showing variation in *de novo* mutagenesis at short tandem repeats in different human families. Short tandem repeats are genomic elements with an elevated mutation

rate thought to largely result from polymerase slippage during replication. However, I find evidence that short tandem repeat mutation rates are associated with both paternal and maternal age at the birth of a proband. Because the maternal germline does not replicate after birth, the latter association supports a new possible damage-associated mutational pathway for these loci. Furthermore, I find that African genetic ancestry corresponds to a significantly higher short tandem repeat mutation rate, particularly at loci with high diversity. My findings may ultimately help determine the factors, either genetic or environmental, that contribute to temporal and spatial variation in germline mutagenesis.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Graduate school is a singular experience, particularly when completed during a historic pandemic. Obviously the journey requires and results in significant personal change – academically, socially, emotionally (and physically). Remarkably, these past five years have been some of the happiest and most fulfilling of my life, and I am so proud of the scientist I have become. It is thus bittersweet to close this chapter, reflect on this process, and acknowledge the people who have been instrumental in supporting me.

To Kelley – to an incredible mentor. It has been not just a joy, but an honor to work in your lab and learn science from you. You've taught me that science can be fun and playful, juggling concepts and tools and combining them in unexpected yet simple and powerful ways. Kelley strikes a great balance between classic and modern, with a mastery of a century of foundational math and theory of population genetics and an excitement for new technology to help solve old problems. I'm so sad to be leaving the lab, but I'm proud to have spent formative years becoming a scientist as a part of it. Thanks as well to all members of the Harris lab, past and present, for fascinating discussions and support.

To my committee – Bill Noble, Evan Eichler, Rick McLaughlin, Brian Browning, Sharon Browning, and Phil Green. Thank you for your support and guidance over the past 4 years. Your critical feedback has been instrumental in shaping my training as a scientist and the projects I've worked on.

To my friends – I love you so much. My graduate school experience has been a particularly solitary one, and you have supported me and reminded me of who I am and who I could be through this process. Amy, Elise, Deanna, Jackson, Fran, Isaac, Liz, Cailyn, Michelle, Brian, and so many more, both in and out of Seattle. I have friends with whom I host dinner parties, with whom I climb mountains, with whom I read poetry, with whom I share music, with whom I swim in lakes, with whom I rave. I'm lucky to have you all.

To my grandparents – Bubbe, Pa, Barbie, Alan. I love you all so much and wish you could have seen me graduate. Bubbe and Pa were both human geneticists, and it has been profoundly meaningful and bittersweet to follow in their footsteps.

To Lois – I can't say how much your presence defined my life and these past five years. Lois taught me high and low culture, to bid on contemporary art at auctions, and to shoot craps. Lois was a force of a woman, and I miss her daily.

To my mom and dad – thank you for everything. I am very close with my parents, and our relationship has matured as I have become an adult. That relationship is typified by the harmony we achieve while cooking oxtail stew at the cabin in the Methow, talking about trail conditions, and debating teaching strategies. I am so lucky to share so many values with you, and am grateful for shaping me into who I have become today.

And to Ellen – my sister. I cannot overstate how important your presence has been on my time here in Seattle. I am so lucky to be your brother and cannot imagine going through graduate school without your support. Spending the last five years with you here in Seattle has been one of the greatest joys of my life.

DEDICATION

This dissertation is dedicated to my parents, Lisa and Bennett, and my sister, Ellen.

Chapter 1. INTRODUCTION

Germline mutations are the ultimate source of genetic variation. Errors in faithful DNA replication in germline cells can cause deleterious pathologies but can also create traits beneficial to the carrier under certain selection regimes. The germline mutation rate, typically the number of expected mutations per year or per generation per base pair, is thus a parameter critical to understanding how species evolve and genetic disorders arise. However, germline mutation rate itself has rapidly evolved and differs even between humans and chimpanzees, our closest living relative [1,2]. Changes to mutation rate therefore represent a case of rapid evolution of a phenotype that affects evolutionary processes and health. However, we have limited knowledge about how mutation rate variation affects genomic function and conversely how changes to biological pathways affect mutation rate.

1.1 EARLY FINDINGS IN MUTATION RATE EVOLUTION

Observations of variation in mutation rate and theories on the evolutionary constraints of their genetic control date back nearly a century to when Sturtevant first found significant differences in the mutation rate between strains of *Drosophila* [3]. Controlling for variation in environment, observed differences in mutation rate between strains could only result from differences in genetic background. Mutation rate appeared to be under at least some genetic control, and could therefore evolve. Because most germline mutations are deleterious, Sturtevant theorized that genetic elements that increase mutation rate must be selected against [3]. Kimura formalized the selective dynamics of a mutation rate modifier in 1967, and critically showed with simulations that the selection against a mutator depends on its increase in mutation rate, linkage to its deleterious mutations, dominance, and distribution of fitness effects of its mutations [4]. Furthermore, high

mutation rates were more quickly selected against, but once mutation rate became sufficiently low, variance appeared less driven by selection. Selection against a very low mutation rate could either result from a population requiring a minimum amount of genetic diversity to evolve or an increasing physical cost to cellular machinery to correct all mutations with diminishing returns [4].

Evidence of mutation rate evolution also comes from phylogenetic studies where substitution rates vary between lineages. In the 1960s, when comparing the cytochrome C sequence of multiple species Margoliash found that the number of amino acid substitutions correlated with genetic distance [5]. Given the divergence time between lineages, one could calibrate the rate of substitution per unit time: a *molecular clock*. This concept was further supported by work on other protein alignments and proponents of the neutral theory of evolution [6,7]. However, by the 1980s, evidence supporting different evolutionary rates along different branches demonstrated that the molecular clock could be erratic. In particular, relatively few substitutions have accumulated on lineages leading to humans and other great apes; this phenomenon is known as the *hominoid slowdown* [1,8,9]. This observed variance in substitution rates was hypothesized to arise from genetic or environmental differences, or a combination thereof.

1.2 MUTATION SPECTRA AND SIGNATURES

Detecting changes to mutation rates and identifying the mechanisms underlying these changes is a complex problem. One successful approach is to use the fact that many mutational processes increase the rate of mutation only at specific patterns of nucleotides [10,11]. For example, error-prone DNA replication due to mutations in DNA polymerase ε lead to mutations in (C>A, A>C), specifically when the cytosine or the adenine is surrounded by thymines or other adenines (TCT>TAT, AAA>ACA), respectively [10,12]. These types of mutations that are most associated with a mutational process are known as a "*mutation signature*". Broadly, the local sequence of

DNA is highly predictive of its three-dimensional shape, which covaries with the set of molecular interactions at that sequence and their likelihoods [13–15]. Thus, classifying mutations by the site of mutation in addition to neighboring bases can help identify specific mutational processes. The relative rates of different mutations can be expressed as a single abundance histogram known as a "*mutation spectrum*", which represents a summation of the different mutational processes active in the studied genomic region. A common spectrum of single nucleotide substitutions will include the 96 unique types of mutations in a triplet context when accounting for reverse complementation (i.e., ACA -> ATA and TGT -> TAT are the same mutation) [10]. However, expanded sequence context can provide additional fine-scale information about highly specific mutation types when considering abundant mutational data [16]. Comparing the mutation spectra of different species, different individuals, or different regions of the genome can uncover even small differences in the activity of mutagens between samples. This type of analysis was popularized to analyze cancer genomes; the signatures of some mutagenic processes were identified by deconvoluting the mutation spectra from genomic sequencing data of many different cancer types [10]. When applied to germline mutation spectra, these analyses have revealed significant diversity in the spectra of mutations accumulating in different populations, families, or species [17–20].

Directly calculating the germline mutation rate to detect its evolution is difficult in multicellular organisms: variability in the quality of genomic sequencing often leads to variability in the number of mutations found in a given sample that could be used to calculate mutation rate [9,17,21,22]. Using mutation spectra to study the evolution of germline mutagenesis variation can partially solve problems inherent to calculating mutation rate because relative rates that comprise mutation spectra are more robust to sequencing heterogeneity [10]. Small shifts in the mutation spectrum, particularly in germline data, are therefore easier to detect than the corresponding

miniscule changes to the overall mutation rate. Furthermore, mutation rate varies spatially across the genome [23]. This variation has largely been classified as “cryptic," meaning unexplained by any known genomic feature or function [24] . In primates, this spatial variation appears to be conserved: the mutation rate of orthologous regions in different primate genomes correlate highly [23]. However, mutation spectrum analysis has recently helped uncover that regions that replicate later than others in S phase, known previously to accumulate more mutations, are specifically enriched for C>A and A>T mutations [24–26].

1.3 PARENTAL AGE EFFECTS AND OTHER MUTATORS

Broadly speaking, mutations have two primary sources: (A) errors in DNA replication that accumulate with each cell division and (B) damage to DNA that accumulates over a cell’s lifespan [27]. A classical view is that the majority of germline mutations in mammals arise from source (A): errors in DNA replication [28]. This assumption is based on the fact that the majority of germline mutations occur on the paternal lineage and their abundance correlates with paternal age [28]. While egg cell divisions cease early in female embryonic development, spermatagonial stem cells divide continuously to produce sperm after a male reaches puberty, thus increasing the likelihood of paternal replication-associated mutations linearly with paternal age. However, emerging evidence now indicates that damage may play a bigger role in germline mutagenesis than previously thought, as the mutation rate in young fathers is already high in humans, and the male bias is remarkably stable across life forms with different generation times [29–32]. Certain types of maternally derived mutations specifically correlate with maternal age (C>G, CpG transitions, e.g.); these mutation types have been linked to damage-associated pathways such as double-stranded DNA breaks and the spontaneous deamination of methylated CpGs [29,33,34].

One way to distinguish replication-dependent signatures from damage-dependent signatures is to examine *de novo* mutations (DNMs) in families where the age of each parent and each mutation's parent-of-origin are known. Examining fixed differences or polymorphisms between lineages can help detect variation in mutagenesis over long evolutionary timescales. However, these analyses lack the fine-scale resolution to differentiate effects of parental environment, life history traits, or specific genetic background, parameters possible to ascertain with DNMs. Although they capture mutations from only two meioses per child and thus have low statistical power, we can directly test the effects of environment and genetic background on mutagenesis. Furthermore, the power to detect segregating mutations correlates with allele frequency, which is affected by selection and drift. Additional forces such as GC-biased gene conversion (gBGC) and background selection result in variance in the likelihood of fixation between different loci and mutation types, thus complicating inference about mutagenesis [19,35–38].

To date, mutation spectrum analyses have uncovered variation in germline mutagenesis across diverse forms of life. Specific genetic or environmental mutators responsible for heritable variation are more challenging to identify with confidence; it appears as though mutagenesis is a highly polygenic trait with the possibility for many small-effect loci [39]. Abstractly, this contrasts to recombination rate, another genomic trait that demonstrates rapid evolution and high variance between lineages but appears largely controlled by a single *trans*-acting locus in most mammals [40–43]. Generation time differences certainly explain a large fraction of variance in mutation rate; in humans, the fraction was most recently estimated at 70%, although other studies have argued for higher fractions [44–46]. However, evidence exists that the strength of parental age effects on mutation rate may vary between families [46–48]. Recently, however, evidence for specific genetic mutators have been uncovered in humans and other mammals [49–52]. In mice, recent studies

have found quantitative trait loci at *Mutyh* and *Msh3* to be associated with C>A mutation and STR expansion rates, respectively [50,52]. In humans, a set of Amish families were recently discovered to harbor a lower mutation rate than other human pedigrees, indicating possible population-specific founder effects [53]. Another recent study found that some families harbor mutations in *TOP1* leading to a higher rate of germline TNT deletions [49].

1.4 MUTATION SPECTRUM VARIATION ACROSS THE GREAT APE PHYLOGENY AND GENOMIC SPACE

In chapter two, I detail work exploring variation in mutagenesis along evolutionary and functional axes in humans and other great apes. Great ape clades exhibit variation in the relative mutation rates of different three-base-pair genomic motifs, with closely related species having more similar mutation spectra than distantly related species. This pattern cannot be explained by classical demographic or selective forces, but implies that DNA replication fidelity has been perturbed in different ways along different branches of the great ape phylogeny. Here, we use whole-genome variation from 88 great apes to investigate whether these species' mutation spectra are broadly differentiated across the entire genome, or whether mutation spectrum differences are driven by DNA compartments defined by features including replication timing and ancient repeat content, that have particular functional features or chromatin states. We perform principal component analysis and mutational signature deconvolution on mutation spectra ascertained from compartments to identify multiple distinct mutational signatures. First, we find evidence for consistent species-specific mutational signatures that do not depend on which functional compartments the spectra are ascertained from. At the same time, we find that many compartments have their own characteristic mutational signatures that appear stable across the great ape

phylogeny. For example, in a mutation spectrum PCA compartmentalized by replication timing, the second PC (explaining 21.2% of variation) separates all species' late-replicating regions from their early-replicating regions. Our results suggest that great ape mutation spectrum evolution is not driven by epigenetic changes that modify mutation rates in specific genomic regions, but instead by trans-acting mutational modifiers that fairly uniformly affect mutagenesis across the whole genome.

1.5 VARIATION IN *DE NOVO* SHORT TANDEM REPEAT MUTAGENESIS IDENTIFIED IN PEDIGREES

In chapter three, I describe recent and ongoing work uncovering variation in germline mutagenesis of short tandem repeats (STRs). STRs are hotspots of genomic variability because of their high mutation rates, which have long been attributed to polymerase slippage during DNA replication. This model suggests that STR mutation rates should scale linearly with the number of cell divisions in male and female germlines. In particular, if STRs only mutate during cell division, their mutation rates should not scale with the age of the mother at her child's conception, since oocytes spend a mother's reproductive years arrested in meiosis II and undergo a fixed number of cell divisions prior that are independent of the age at ovulation. We tested this prediction using *de novo* mutation calls from the Simons Simplex Collection, a cohort of nearly 1600 human quad families, each consisting of two children plus parents whose ages at the birth of the children are known. Contrary to expectation, STR mutation rates covary with maternal age as well as paternal age, implying that some STR mutations are caused by DNA damage in quiescent cells rather than the classical mechanism of polymerase slippage. Our results echo the recent finding that DNA damage in quiescent oocytes is a significant source of *de novo* SNVs. However, we find that homopolymer STRs have a smaller maternal age effect than STRs of longer repeat unit lengths, and that the

maternal age effect is not confined to previously discovered hotspots of oocyte mutagenesis – an especially surprising observation in light of the prior belief in replication slippage as the dominant mechanism of STR mutagenesis.

Although we find no evidence that the proportion of maternal STRs varies among populations, we do find that STR mutation rates are overall significantly higher in families with African ancestry. However, further analysis suggests that some of this signal may be the result of allelic dropout, a bioinformatic artifact that often complicates STR calling. False positive *de novo* mutation calls can result from parental dropout of alleles that are called correctly in a child, and this is disproportionately likely to affect African families because of their higher heterozygosity. Nevertheless, stringent filtering to account for allelic dropout does not fully explain the observed difference in mutation rates.

Overall, our results suggest that STR mutagenesis cannot be fully explained by replication slippage but is influenced by the DNA damage affecting quiescent cells that has recently been shown to generate a significant fraction of point mutations. Better allelic dropout correction methods may be needed to mitigate the influence of heterozygosity on inferences of STR mutation rates. Yet, our observation of population differences in STR mutation rates suggests these rates may be influenced by genetic background or by environmental differences that correlate with ethnicity.

Chapter 2. MUTATIONAL SIGNATURES OF REPLICATION TIMING AND EPIGENETIC MODIFICATION PERSIST THROUGH THE GLOBAL DIVERGENCE OF MUTATION SPECTRA ACROSS THE GREAT APE PHYLOGENY

This chapter comprises work currently published in Genome Biology and Evolution [24].

2.1 INTRODUCTION

The pace of evolution and the healthspan of somatic tissue are both ultimately limited by the genomic mutation rate, which is a complex function of DNA damage susceptibility, polymerase fidelity, proofreading efficacy, and other factors [10,54]. Some regions of the genome accumulate mutations faster than others, such as DNA that replicates late in the cell cycle and motifs with certain epigenetic modifications [25,54–58]. Such mutation rate differences have the potential to confound efforts to infer patterns of purifying selection and background selection [59,60]. As a result, understanding how mutation rate varies across the genome is an important prerequisite for identifying modes and targets of natural selection [61–63]. Understanding the mutational landscape is similarly essential for predicting rates of deleterious *de novo* mutations in clinically relevant disease genes [64,65].

Some variation of mutation rate across the genome appears to be driven by features such as chromatin state, transcription, or, more broadly, genomic function [66,67]. For the purpose of this manuscript, we broadly consider a genomic compartment's "function" to be a set of shared molecular interactions that might or might not be critical to organismal fitness. However, even with this broad definition, a large component of the variance of the mutational landscape still

escapes association with any known motifs or functions and has thus been classified as “cryptic” [23,68,69]. This cryptic variation is conserved over relatively long timescales, being highly similar between human and macaque, which diverged ~25mya [70]. This conservation suggests that many regions of the genome have intrinsic or epigenetic features that affect their mutability and are functionally important enough to be maintained over time despite creating excess deleterious mutation load.

One powerful tool for disentangling the effects of selection and mutation rate variation is mutation spectrum analysis: a comparison of the relative abundance of specific types of mutations, often defined by their triplet context (i.e. AAA>ACA or ACG>ATG) [19]. When background selection removes genetic variation from a genomic region, it removes variants that are essentially sampled at random from the spectrum of variation that is present. Biased gene conversion can affect the mutation spectrum, but only in a specific way, favoring the retention of mutations from A/T to G/C over mutations from G/C to A/T [35]. A much broader variety of mutation spectrum changes can occur when the mutation rate increases as a result of alteration to the mechanisms of DNA damage or repair, most famously in cancer where cells’ housekeeping processes often break down [10]. For example, tumors that replicate their DNA with a defective polymerase ε accumulate high rates of TCT>TAT and TCG>TTG mutations [10,12]. Similar “mutational signatures” also occur in the normal human germline, where late-replicating DNA consistently accumulates proportionally more C>A and A>T mutations compared to DNA that replicates earlier during the cell cycle [26].

In addition to varying between regions of the genome, mutation rates and spectra also vary between different evolutionary lineages. Patterns of diversity point to a global mutation rate slowdown during hominoid evolution that has caused humans and closely related apes to

accumulate mutations more slowly than distantly related monkeys do; this slowdown has been studied since the 1980s [1,8,71]. A closer examination of ape mutation spectra recently revealed that every ape lineage has experienced changes in the relative mutation rates of some characteristic triplet motifs [18]. Even more surprisingly, closely related human populations have distinctive mutation spectra that provide enough information to classify individuals into continental ancestry groups [18]. Sometime during the 100,000 years since their descendants migrated out of Africa, Europeans experienced a temporary pulse of mutagenic activity that more than doubled the rate of TCC>TTC mutations [17,72,73].

Mutation spectrum divergence has the potential to shed light on both the mechanisms and overall speed of mutation rate evolution. In theory, many different biological mechanisms can cause mutation rates to evolve; these may be changes to cellular machinery (e.g., changes to a DNA repair protein) or to environmental/life history traits (e.g., longer generation time) [12,29]. In the event that all mutagenic processes generate diversity in a conserved and clocklike manner, mutation spectra are expected to stay constant over time. Conversely, if exposure to a particular mutagen goes up or a DNA repair mechanism becomes less efficient, this is expected to elevate the relative dosage of mutations in genomic motifs that are most vulnerable to damage by that mutagen or preferentially targeted by that DNA repair mechanism. Different genetic and environmental perturbations might cause similar changes in the overall mutation rate, but are less likely to elevate mutation rates in the same genomic sequence contexts.

In this study, we examine how mutation spectra covary across the genome and the ape phylogeny and find that great ape mutation spectra exhibit similar patterns of mutation spectrum divergence across both slowly and quickly mutating compartments of the genome. We find no evidence that species-specific mutator activity is correlated with chromatin state, ancient repeat

content, or replication timing during S phase, despite the fact that all of these variables correlate with stable mutational signatures that are consistently detectable across the great ape phylogeny. This implies that the rapid evolution of great ape mutation spectra has likely been driven by *trans*-acting mutators that do not preferentially target any specific genomic compartments, at least not compartments that are correlated with the variables examined in this study. Although there exist many differences between functionally divergent compartments in the spectra of mutations that accumulate, these differences appear to exhibit considerable stability between great ape species and are not likely responsible for the rapid mutation spectrum divergence. We find that such stability extends to species as divergent as humans and mice, where we find consistent mutational signatures present in genomic regions that are not homologous but are annotated with the same functional states as promoters, enhancers, or repressed regions.

2.2 RESULTS

2.2.1 *Quantifying the mutation spectrum differences between great ape species and subspecies*

Previous research utilizing the Great Ape Genome Project (GAGP) data showed that the germline mutation spectrum has evolved rapidly in great apes, leading to distinct species-specific spectra [18,74]. We first sought to recapitulate these results and measure for the first time how the differences between species compare to differences within species (Table S1).

To minimize the effects of natural selection and read mapping errors on our mutation spectrum ascertainment, we defined a set of genomic regions, collectively called a "compartment", characterized as non-conserved and non-repetitive (NCNR). This NCNR compartment consists of 1.28Gb of the non-repetitive (annotated by RepeatMasker), non-coding human genome excluding both significantly conserved regions (p<0.05 in the PhastCons 44-way

primate alignment) and CpG islands (Table S2). In these compartments, we computed the relative abundances of each of the 96 triplet mutation types for each individual and each species (Figure 1A) using SNVs from the GAGP, following a number of filters. To ensure that shared genetic drift cannot inflate the appearance of mutation spectrum similarity among individuals that share many derived alleles, we computed mutation spectra using a sampling procedure that randomly counts each SNV toward the spectrum of only one of the individuals that carry it, rather than all such individuals (see Methods). This sampling method reduces the impact of GC-biased gene conversion, which primarily affects higher frequency alleles in regions specific to lineages (see Methods, Figure S1).

A principal component analysis (PCA) on mutation spectra shows clustering of individuals by species in a manner that recapitulates phylogeny. This pattern reveals the existence of distinct species-specific mutation spectra (Figure 1B). Mutation spectra of humans, chimpanzees (*Pan troglodytes*), and bonobos (*Pan paniscus*), which form a phylogenetic clade, separate from more distantly related gorillas (*Gorilla gorilla*) and Sumatran and Bornean orangutans (*Pongo abelii* and *Pongo pygmaeus*, respectively) along principal component 1 (PC1). PC2 separates humans from chimpanzees and bonobos in addition to separating gorillas from orangutans. The two orangutan species cluster closely together, which is unsurprising given that their divergence time (~483kya) is an order of magnitude more recent than that of humans and chimpanzees [74]. PC1 appears dominated by the proportion of A>C, and A>T and components of C>T mutations, while PC2 is dominated by the proportion of C>A and C>G mutations (Figure S2). The major A>C component of PC1, for example, corresponds to a 10-15% decrease in the fraction of A>C SNVs in gorillas and orangutans relative to humans, chimpanzees, and bonobos. PC2's C>A component corresponds to a 4% and 7% increase in the C>A SNV fraction in humans and

gorillas relative to the *Pan* and *Pongo* clades, respectively. These results are robust to subsampling of equal numbers of individuals across species (Figure S3). Although prior papers have reported homogeneity of de novo mutation spectra between non-human great apes, we find that these studies are underpowered to detect species differences of the magnitude we observe here [2,75] (See Methods, Table S3).

We even observed mutation spectrum differences among chimpanzee subspecies, as visible in the PCA: Western chimpanzees (*P. troglodytes verus*) overlap slightly less with bonobos along PC1 than other chimpanzee subspecies, which begin to separate from bonobos along PC2. The mutation spectrum differences between Western and non-Western chimpanzee subspecies is more clearly visualized in a PCA run on those individuals alone (Figure S4). The first PC demonstrates that the variance in mutation spectra between Western and non-Western chimpanzees significantly exceeds the variance among non-Western chimpanzees ($p \leq 2.2 * 10^{-16}$, two-sided t-test). Western chimpanzees experienced a population bottleneck when they diverged from the lineage ancestral to other chimpanzees, which might have accelerated their mutation spectrum divergence by allowing mutator alleles to drift to higher frequency [4,74,76–78] (Figure S4). Gorilla subspecies, orangutan species, and human populations exhibit more subtle mutation spectrum differences that are not visible when spectra are projected onto the principal axes of ape variation [18] (Figure S4). To quantify these mutation spectrum differences further, we embedded mutation spectrum histograms as points in a 96-dimensional space and computed distances between the using a standard Euclidean metric. As seen qualitatively in the PCA, we found that interspecific differences exceeded conspecific differences (Figure 1C). Furthermore, interspecific differences scaled with divergence time.

We undertook additional analyses to verify that the observed mutation spectrum differences are not likely caused by the tendency of natural selection to retain variation in certain sequence contexts. Although GC-biased gene conversion does favor the fixation of G/C alleles and the elimination of A/T alleles, none of the mutational signatures that vary in dosage between great ape species are consistent with the spectra of mutations that for which biased gene conversion selects (Figure S2). Moreover, when we stratify allele frequency within each species, we see the expected directional effect of biased gene conversion (Figure S5). However, rare alleles have spectra that are no more similar across species than more common alleles that are expected to be more profoundly affected by biased gene conversion. Furthermore, any mutation spectrum difference caused by the action of natural selection should affect high frequency alleles more than low frequency alleles, so we repeated several key analyses using only low frequency variants. We performed these replicate analyses using doubletons rather than singletons, since singletons are more vulnerable to confounding by sequencing error, and found no qualitative differences from the results obtained using the full frequency spectrum of genetic variants (Figure S6A-B).

We used non-negative matrix factorization (NMF) to explicitly infer which mutational signatures have changed in dosage along different branches of the ape phylogeny (Figure 1D). Similar to PCA, NMF is a model-free method used to determine major components of variance that underlie a matrix of data; the components NMF extracts from mutation spectrum matrices can be interpreted as mutation signatures, following the work of Alexandrov et al., 2013. We ran NMF on the matrix of individual NCNR mutation spectra using Helmsman, specifying the model to infer $K = 6$ signatures [79]. Figure 1D shows the dosage of each signature for every individual in the GAGP, grouped by species. Although each signature is present in every individual, the

signatures clearly demonstrate lineage-specific dosage. For example, S3 is present at moderate dosage in all non-human species but appears to have increased in relative rate in humans following the divergence of humans from the human-chimpanzee-bonobo common ancestor (Figure S7). S4, conversely, has decreased 2-fold on the branch leading to human-chimpanzee-bonobo; S2 has decreased nearly 3-fold on the branch leading to orangutans. These results support a prior hypothesis that the dosage of one or more mutational signature has changed along each branch in the great ape phylogeny [18].

2.2.2 *A mutational signature associated with DNA replication timing is conserved among great apes*

Differences in replication timing explain a substantial portion of the variation in somatic and germline mutation rate across the genome [25,55]. Compared to regions that replicate early in S phase, late replicating regions tend to have a higher overall mutation rate, and in humans they particularly harbor a higher rate of A>T and C>A mutations [26]. The established correlation between late replication timing and elevated mutation rate implies that replication timing QTLs (rtQTLs) may be examples of *cis*-acting mutation spectrum modifiers, with late-replication alleles acting to increase the load of a late-replicating mutational signature in the surrounding DNA. We analyzed late- and early-replicating compartments of the genome to determine whether replication timing had a similar effect on the mutation spectrum across great apes.

In an attempt to replicate the late-replication signature reported by Agarwal and Przeworski, we defined early and late replication timing compartments to be the earliest- and latest-replicating quartiles of the genome identified by RepliSeq in human lymphoblastoid cell lines [55] (Figure 2A, Table S2). We then generated separate mutation spectra from the early-replicating and the late-replicating compartments of each individual genome, normalizing the

spectra by the triplet composition of each compartment. We ran a PCA on a matrix containing three mutation spectra derived from each human genome in the GAGP: two derived from early- and late-replicating compartments and the third derived from the NCNR. compartment (Figure 2B, Figure S8) and observed that PC1 separated out these spectra as a function of replication timing. To determine the principal triplet mutation types driving the differences in mutation spectra between compartments, we generated heatmaps of the log odds ratio enrichments of each mutation type occurring in the late versus the early replication timing compartments (Figure 2C, Figure S9). For this analysis, we counted the number of segregating sites within a species to generate a single 96-dimensional vector for each species and compartment (rather than a vector for each individual and compartment, see Methods). As expected, late-replicating regions were enriched for C>A and A>T mutations.

After replicating the reported effect of replication timing on the human mutation spectrum, we set out to determine if the action of this signature appeared conserved among species. To this end, we identified the ape genomic compartments that aligned to the human early-replicating and late-replicating compartments and ran a PCA on a matrix of the early replicating, late replicating, and NCNR compartment mutation spectra of all individuals (Figure 2D). PC1 separates the spectra by phylogeny and species identity, while PC2 separates them along an axis that aligns with replication timing. PC3 separates human from chimpanzee and bonobo spectra (not shown). The direction of separation of early and late replication timing compartments is similar across all species, implying the action of a conserved mutational process to a consistent compartment of the genome. The most parsimonious explanation for this pattern is that the replication timing landscape is largely conserved across great apes and that late replication exerts a similar mutagenic effect in all great ape species. As above, we obtained consistent results using only

doubletons (Figure S6C,D). A PCA of individual mutation spectra, using only doubletons, recreates similar clustering patterns to those presented in Figure 2. Furthermore, the second PC again captures the mutational signature of late replication timing (Fig S6D). The main effect of restricting to rare variants is that it makes humans appear less displaced from chimps along the axis of differentiation by replication timing. These results suggest that differences between species in biased gene conversion or demographic history have a minimal effect on the observed trends (Figure S6).

To further quantify the conservation of the replication timing mutational signature, we tested the correlation of the log odds of late vs. early replication compartments between each pair of species, thereby quantifying the similarity between each species' late vs. early replication timing mutational heatmaps. The correlations between every pair of species' replication timing heatmaps were highly significant (Figure S9, Table S4). Our results show that late replication timing is associated with a conserved mutational signature across great apes. Moreover, these mutational patterns suggest that the genomic landscape of replication timing is broadly conserved across species, biasing all genomes toward C>A and A>T mutations in compartments that are directly orthologous to the human late-replicating compartment (Figure S10). The conservation of this mutation signature contrasts with the rapid evolution of the species signatures. We see a few hints of interactions between replication timing and species identity – for example, CG>GG mutations are depleted in late replication timing regions, and the depletion appears slightly stronger in non-human apes than in humans. However, these nonlinear effects are small and higher resolution sequencing data will be needed to verify whether they are true biological signals.

The layering of rapidly evolving species signatures over conserved replication timing signatures was further supported by NMF decomposition. We ran NMF independently for each species on matrices of individual mutation spectra from the early replicating and late replicating compartments, after normalizing for compartment-specific nucleotide content. Each NMF was run to extract $K = 7$ signatures. Following the methodology of Alexandrov et al. 2013, we then grouped similar signatures by their loadings using hierarchical agglomerative clustering (Figure S11). The clustering of these signatures, even following independently-run NMFs, supports the deep conservation of a mutation signature associated with replication timing that spans the great ape phylogeny. Furthermore, the loadings of these signatures are enriched for C>A and A>T mutation types, once again supporting our presented results.

2.2.3 *Great ape genetic variation appears to be shaped by a conserved landscape of cis-acting mutational modifiers*

We found that all ape DNA, regardless of replication timing, has a consistent mutation spectrum bias that we call a species-specific signature. Late-replicating regions of chimpanzee genomes contain a mixture of both a chimpanzee-specific signature and the same late-replication signature that is found in human genomes (Figure S9). We see little evidence of any mutational signature unique to late-replicating chimpanzee DNA that is not also found in early-replicating chimpanzee DNA or in late-replicating regions of other ape genomes. Furthermore, we see no evidence that species-specific signatures have a rate or dosage that depends on replication timing.

Using published annotations of the human genome, we defined several more overlapping functional compartments to characterize the phylogenetic distribution of spatially localized mutational signatures. We used RepeatMasker to delineate repetitive vs. non-repetitive DNA and

used ENCODE chromHMM output (the intersection of heterochromatic regions in nine cell types) to annotate several types of heterochromatin [80]. Another compartment we annotated consists of ancient repeats, which have the potential to mutate differently from higher complexity DNA via several mechanisms, including the formation of non-B-DNA secondary structures and the editing activity of antiviral enzymes such as APOBECs [81–83] (Table S2).

We ran a PCA for each species comparing the mutation spectra of all eight compartments and observed similar topologies of compartment separation in each species. In all cases, the vector separating late-replicating from early-replicating compartments is nearly aligned with the first principal component, which explains 23.3% to 34.1% of the total variance. PC2, which explains 8.2%-13.5% of the total variance, is similarly aligned with the separation between repetitive and nonrepetitive compartments. Finally, PC3 (4.0%-8.7% variance explained) separates the ERVs from the other compartments to a much greater extent than either of the first PCs (Figure 3A-F, Figure S12). The similarities of the independent PCAs across all species of great apes imply conservation of the *cis*-regulated mutational signatures associated with repetitive content, methylation, and replication timing. Each compartment shows a similar degree of separation between species, with high degrees of correlation between the positioning of compartments in different species' PCAs (Table S5).

We identified only one genomic compartment whose localized mutational signature appears to be distributed non-uniformly across species: maternal mutation hotspots first identified using human trio data [33] (Table S2). Maternal mutation hotspots are genomic regions that are enriched for *de novo* C>G mutations that arise on the maternal lineage and whose rate has an unusually strong correlation with maternal age. These hotspots exist in chimpanzees and, to a lesser extent, gorillas, but their signal is nearly absent from orangutans. Our mutation PCAs

similarly show that human genomes have higher levels of a compartment-specific mutational signature in these regions (Figure S13). The separation of NCNR to maternal mutation hotspot mutation spectra compartments decays with phylogenetic distance from humans, recapitulating findings from Jonsson et al. [33]. Although a *trans*-acting protein might be involved in the creation of C>G mutations at maternal hotspots, perhaps due to error-prone repair of double strand breaks, the targeting of damage toward specific regions fits the profile of a *cis*-acting targeting factor. Either this *cis* targeting factor or a *trans* interacting partner has been intensifying its mutagenic effects along the evolutionary lineage leading to humans, causing extra mutations to accumulate in a localized pattern. Although differences in maternal age at conception may be a partial explanation for these observations, generation time differences cannot fully explain the patterns across all great apes. The strength of this signature certainly decreases with generation time among humans, *Pan* clade, and gorillas (29, 24, and 19 years respectively). However, orangutans have a higher average maternal age (25 years) at conception than that of gorillas and the *Pan* clade but exhibit the weakest dosage of the C>G signature [2].

The compartment that harbored a signature whose dosage was the strongest among the presently investigated compartments are CpG islands, genomic regions ranging from 200-2000bp long that are enriched for CpG dinucleotides [84,85] (Table S2). CpG islands are the only compartment we identified whose differentiation from the NCNR compartment explains a greater proportion of mutation spectrum variance than differentiation between species across these regions. Outside of these CpG islands, CpGs are often methylated to 5-methylcytosine, which mutate to TpG at a rate ten times higher than unmethylated CpGs. In contrast, CpG islands are hypomethylated and are often situated in conserved 5' promoters and genic regions. We hypothesized that, due to their lack of CpG>TpG mutations and overall conservation, a

compartment containing CpG islands would demonstrate a contrasting mutation spectrum relative to that of the NCNR compartment. A PCA of individual mutation spectra from the NCNR and a CpG compartment from all GAGP individuals demonstrated that differentiation of the CpG island compartment exceeds the magnitude of spectrum differentiation between species (Figure S14). This is unsurprising given the unique mutational properties of CpG islands compared to the rest of the genome.

2.2.4 *Endogenous retroviruses carry a distinct mutational signature conserved across all great ape species*

ERVs are a class of repetitive, transposable DNA elements that duplicate themselves in a copy and paste manner. The act of duplication into new regions of the genome can disrupt function; for example, integration into the coding region of a gene could result in a complete knock-out of gene function. Therefore, ERV activity is restricted by a number of known mechanisms, including hypermethylation, inhibition of integration, and hypermutation.

In light of these mechanisms that target ERVs, we were intrigued by the fact that ERVs separated from other genomic compartments along the third principal component of our mutation spectrum analysis (4-8% variance explained). ERVs bear an excess load of a unique mutational signature that appears to be largely conserved among great apes (Figure 2B,D,F) and is previously undescribed, to our knowledge.

To determine whether any component of the ERV signature could be caused by high rates of methylation and heterochromatization, we directly compared the mutation spectrum of the ERV compartment to that of nonrepetitive heterochromatin. We calculated the log ratio enrichments and depletions of the 96 mutation types in ERVs relative to nonrepetitive heterochromatin for each species and found an enrichment for CpG C>G mutations and a depletion of TAA>TTA

mutations in ERVs that appears conserved in all species other than *Pongo pygmaeus* (Figure 4A). This comparison shows that ERVs' high rate of CpG>CpT transitions is likely caused by their heterochromatic status, but that heterochromatinization cannot explain the other components of the ERV signature. Furthermore, we determined that differences in 7-mer nucleotide content between the two compartments explained some, but not all, of the ERV-specific enrichment for CG>GG mutation types (Figure S15).

We hypothesized that the CpG C>G mutational signature could result from the high and variable rates of CpG hydroxymethylation (hmC) of ERVs, which has been recently shown to increase rates of C>G mutations [86]. To test this hypothesis, we compared the mutation spectra of ERVs with versus without evidence of hmC CpG, based on hmC-specific sequencing of human embryonic stem cells [87] (Table S2). ERVs with hmC showed a significant enrichment for CpG C>G mutations compared to ERVs without hmC in all six species, supporting the hypothesis of hmC-related mutagenesis in ERVs (Figure 4B, Table S6). We assessed the robustness of the CpG C>G mutational signature to differences in mapping quality, compartment size, and species-specific nucleotide content, finding that the CpG C>G mutational signature was robust to all quality control tests (Figure 4C, Figures S16-17).

2.2.5 *Conserved mutational signatures are associated with functional genomic elements in distantly related mammalian species*

After observing that compartment-associated mutational signatures appear to be conserved among great ape species, we hypothesized that such conservation might extend to even more distantly related species and tested this hypothesis using annotations of epigenetic function that were generated by the ENCODE project for humans and mice. Specifically, we split the human and mouse genomes into six different functional compartments based on chromHMM

annotations generated from epigenetic assays run in ESCs and mESCs, respectively [80,88] (Table S7). These chromHMM-defined compartments do not necessarily correlate with the various compartments we analyzed across great apes. We posited more generally that mutational signatures associated with specific compartments that experience specific molecular interactions (DNA binding proteins, histone modifications, chromatin state, e.g.) are relatively stable among similar genomic regions in different species. Across these compartments, we calculated normalized mutation spectra using publicly available whole-genome polymorphism data from 2504 diverse humans and 67 wild-caught *Mus musculus* and *Mus spretus* individuals [89,90]. Ancestral states for human polymorphisms were determined with regards to an inferred reconstructed genome representing the most recent common ancestor of humans; ancestral states for mouse were polarized based on an aligned rat reference genome (rn6) [89]. Polymorphisms were subject to filters similar to those applied to the great ape data: sites that either failed quality filters, were missing from >20% of haplotypes, or had a minor allele frequency of 1/2N (e.g. singletons) were excluded from analysis. Sites fixed in the sampled haplotypes were also excluded. We employed a similar randomization strategy to avoid structure due to shared variation, but summed together the mutation spectra of humans from the same subpopulation and mice from the same sampling site and subspecies to avoid sparseness (n=26, 9 respectively; see Methods).

Two separate matrices containing mutation spectrum data from all compartments in each respective species were decomposed using PCA, and we observed several commonalities between human and mouse in the spatial arrangement of the chromHMM compartments within the span of the first two principal components (Figure 5). Unsurprisingly, mouse species separate to a greater degree than do human populations, but within a subspecies the relative positioning of

spectra from different functional compartments mirrors that observed in humans, especially after a roughly 30-degree rotation of the PC axes. In humans, PC1 largely separates promoters from other compartments, and PC2 separates transcribed regions. Enhancers and insulators cluster together, indicating mutation spectra more similar relative to other compartments. These same features are true of the mouse PCA, except that the vector separating promoters from other compartments is intermediate between PC1 and PC2. Mutation spectra from the heterochromatin or repressed compartments cluster more closely to insulators and enhancers respectively in mouse and human respectively; this phenomenon could be due to annotation differences. Furthermore, the loadings of PC1 and PC2 are significantly correlated between the two separate PCAs (Pearson's $\rho = 0.486, 0.555$, $p = 5.14*10^{-7}, 4.44*10^{-9}$, respectively) (Figure S18). As expected, CpG transition rates are weighted heavily along PC1, likely reflecting the unique methylation patterns that regulate promoter function, but other types of transitions and transversions also appear to have different rates among these functional compartments. Both C>A and A>T, the mutation types associated with late replication timing in great apes, appear negatively associated with PC2, a pattern that is consistent with the tendency of transcribed regions, which are positively displaced along PC2, to occur in early replicating regions. Although ChromHMM-defined compartments are not necessarily orthologous between human and mouse, their shared epigenetic markings and functions appear to be enough to cause a common set of sequence motifs to be particularly vulnerable to mutation.

2.3 Discussion

Despite considerable documented evidence of mutation spectrum variation across genomic space and phylogenetic time, little was previously known about the covariation of mutation spectra along spatial and temporal axes. Our results show that such covariation is negligible, at least in

great apes: spatial and temporal mutation spectrum variation are largely orthogonal to one another. Replication timing, repeat content, and other functional categories have consistent mutational biases across all great ape species. At the same time, each species has a distinctive mutation spectrum bias that affects all functional compartments we have analyzed. There exist some exceptions to this general rule, most notably in the compartments of the genome that accumulate a maternal-age-related signature in humans that is attenuated in chimpanzees and gorillas and nearly nonexistent in orangutans. Nevertheless, our results show that mutation spectrum divergence between ape species is mostly driven by processes that act promiscuously across the genome. Determining how broadly this conclusion applies to species beyond great apes will be an exciting avenue for future work.

Our results provide evidence that mutation rates in late replicating regions are elevated due to a mechanism whose activity pattern over great ape evolution has been largely conserved. The distinct signatures associated with different great ape species demonstrate that the simplest possible “hominoid slowdown” model is likely not sufficient to explain all differences between great ape mutation rates. On one hand, some papers have suggested that increasing reproductive age is enough to explain differences in mutation rates that have been inferred by analyses of the branch lengths of great ape phylogenies. However, we see no evidence that differences between great ape species’ mutation spectra can be explained by varying the dosage of a single mutational signature associated with increased parental age, or even separate signatures associated with paternal and maternal ages of the kinds that have been inferred from human de novo mutation data.

To understand why species-specific mutational signatures are not obviously biased toward particular genomic regions, it is helpful to consider previous theoretical work on the dynamics of

alleles that drive mutation rate evolution [3,4,77,78]. Although a genetic modifier of the mutation spectrum will not necessarily change the mutation rate, the most parsimonious scenario is for new mutations that alter the mutation spectrum to slightly increase the overall mutation rate by impairing the faithful replication of particular motifs. Kimura, Lynch and others have noted that selection against alleles that increase the mutation rate is driven by selection against the excess deleterious variation that such alleles beget. However, most new variants created by a *trans*-acting mutator will be on different chromosomes or distant parts of the same chromosome that will immediately recombine with other genetic backgrounds. The only deleterious mutations that are likely to cause selection against the mutator are mutations that happen to occur in a small window that maintains high linkage disequilibrium with the mutator locus. In contrast, a mutator that affects mostly neighboring DNA will tend to stay linked to more of the deleterious mutations it creates, making it more susceptible to purifying selection and loss from the population. A *trans* acting rate modifier that targets specific genomic regions might not experience strong linked selection itself, but the associated genomic targets might experience selective pressure to stop attracting targeted mutagenic activity.

Some species-specific signatures might be the footprints of environmental mutagens, but *trans*-acting genetic modifiers are more parsimonious explanations for signatures that affect larger clades of multiple species. The mutations we analyzed here are all segregating variants that originated long after modern ape species had become reproductively isolated, and environmental exposures are not likely to have respected phylogenetic boundaries for millions of years after the completion of ape speciation. Fixed differences between polymerases, DNA repair factors, and/or their regulatory elements are more likely to be responsible for differences in mutation spectra that respect phylogenetic structure and act consistently across the genome.

We have noted that all ape species exhibit some internal mutation spectrum substructure, with Western chimpanzees being the most distinctive subspecies. Western chimpanzees are different from other chimpanzee subspecies in several ways, including lower levels of bonobo gene flow, a higher load of transposable elements, and a stronger population bottleneck in their recent history. Both transposable elements and accelerated genetic drift may have hastened this lineage's rate of mutation spectrum drift.

Although selection, drift, and demographic history certainly affect genome-wide genetic diversity and variation in diversity across the genome, these forces are not reasonable explanations for the patterns of mutation spectrum divergence we present in this manuscript. Biased gene conversion selects for mutations from A/T to G/C, but none of the mutational signatures that vary between compartments or species fit this simple profile. In coding regions, selection generally allows synonymous mutations to reach higher frequencies than nonsynonymous mutations, but coding regions comprise a nonexistent to negligible fraction of the various compartments analyzed here, and the universality of the genetic code prevents selection against nonsynonymous substitutions from generating any distinctive species-specific differences. It is generally not plausible to suppose that the same mutational signature would be selected for in a single lineage across all the noncoding genomic compartments we analyze here, as this would require mutations in many triplet contexts to have fitness effects that were somehow consistent across the whole genome.

Although identifying the causal fixed differences still represents a challenging unsolved problem, the insights from this paper will allow us to narrow the field of possible mutation spectrum modifiers to exclude ones that target only subsets of the genome. Focusing on ERVs in detail, we were able to use functional genomic annotations to link this compartment's CG>GG

mutational signature to hydroxymethylation of CpG sites. Examination of a broader set of functional genomic data may facilitate the interpretation of other localized signatures and bring us closer to understanding their causality.

Previous work estimated that 80% of spatial mutation rate variation could be explained by letting mutation rates depend on an extended 7-mer sequence context [16,91]. Since 7-mer composition differs between genomic compartments, these extended sequence context models likely derive some of their predictive power from the effects of *cis*-acting mutational modifiers. However, we have shown that compartment annotations provide extra information about mutability above and beyond what we can tell from extended sequence context alone, at least in the case of the ERV hydroxymethylation signature. An important avenue for future work will be to examine the converse possibility and determine how much of the dependence of mutability on extended sequence context can be explained by genomic compartmentalization.

A small proportion of the genome is expected to vary in mutation rate between individuals due to the presence of rtQTLs, but our results suggest that the coarse shape of the replication timing landscape is stable across the great ape clade [55]. The genomic distribution of the replication timing signature could even be leveraged to estimate the extent of TAD variation within and between species. More generally, if mutational signatures of chromatin states and various epigenetic modifications prove stable and interpretable over large phylogenetic clades, they represent a valuable source of information about the evolution of chromatin structure and function in non-model organisms where only genome sequences are available.

2.4 METHODS

2.4.1 *SNV filtering*

We ascertained mutation spectra from a set of high-coverage great ape SNVs that were previously called and filtered by Prado-Martinez et al. [74]. For each species and compartment, we collated the set of biallelic SNVs falling within the genomic segments that comprise that compartment. Ancestral states were assigned using a parsimony approach. Briefly, a biallelic site segregating within a genus (*Homo*, *Pan*, *Gorilla*, or *Pongo*) was polarized to the allele fixed in all other genera. Sites segregating in multiple genera, sites with multiple fixed alleles, and sites with more than two alleles in a single genus were excluded due to their inconsistency with the assumptions of no balancing selection and only one mutation event per site. Singletons were also excluded due to their higher likelihood of sequencing error. We used the inferred ancestral base to classify 3' and 5' neighboring nucleotides. For 3-mer mutational analyses, we excluded SNVs whose 3' and 5' neighboring nucleotides were an 'N' in the hg18 reference and SNVs demonstrating evidence of recurrent mutation in the great ape lineage; we expanded this filter to include the three 3' and 5' neighboring nucleotides for 7-mer mutational analyses. Finally, we removed SNVs out of Hardy-Weinberg Equilibrium with excess heterozygosity (using an exact test, $p<0.05$). Excess heterozygosity at a locus could indicate a cryptic segmental duplication with a single, fixed mutation in a copy. We also excluded SNVs with ≥ 0.5 derived allele frequency to avoid mutational classes with an elevated risk of ancestral state misidentification.

2.4.2 *Computing the mutation spectra of individuals and species*

The PCA analyses in this paper require the computation of mutation spectra from individual genomes, whereas the complementary heat map analyses involve calculating aggregate mutation

spectra from larger samples. Each analysis employs the filtering system described above and ultimately involves counting the number of filtered derived alleles, classified by 3-mer context. However, slightly different calculation details are involved in the two cases.

The aggregate mutation spectrum of a species *S* is obtained from a set of counts $C(m_1,S),\ldots,C(m_{96},S)$ where $m_1,\ldots,m_{96}$ are the 96 3-mer mutation type categories AAA>C, …, TCT>T. The count $C(m_1,S)$ is the total number of SNVs segregating in species *S* that fall into the mutational equivalence class m_1. To compare spectra across samples with different amounts of variation, these mutation type counts are normalized to obtain a 96-dimensional histogram with frequency categories summing to 1.

A mutation spectrum can be similarly calculated from a particular individual *I* as the distribution of 3-mer mutation types across the derived alleles present in *I*'s genome. Homozygous derived alleles are given twice the weight of heterozygous alleles such that the spectrum is the average of the spectra one would compute from the two phased haplotypes making up *I*'s diploid genome.

When individual mutation spectra are computed in this way, two types of derived alleles can contribute to spectrum covariance between individuals *I* and *J*. The first type are pairs of derived alleles that occur at separate loci in *I* and *J* but belong to the same mutation equivalence class. The second type are derived alleles inherited by both *I* and *J* from a common ancestor. To maximize our power to detect mutation spectrum evolution and distinguish it from shared genetic drift, we devised a randomization strategy to eliminate the second source of signal while preserving the first.

This randomization strategy involves computing the mutation spectrum of individual *I* from only a subset of the derived alleles present in *I*'s genome. If one copy of a particular derived

allele is present in *I*'s genome and has frequency *k/2N* in the GAGP panel, the allele will be counted toward *I*'s mutation spectrum with probability 1/*k*. Conversely, this derived allele will be counted toward the mutation spectrum of exactly one ape haplotype that carries it, with the identity of that haplotype chosen uniformly at random.

2.4.3 *Comparing mutation spectra across genomic compartments*

Comparing mutation spectra between regions of the genome required accounting for differences in compartment size and nucleotide content. Larger compartments naturally had more mutations than smaller ones; it was therefore necessary to compare mutation fractions rather than raw counts. Furthermore, differences in nucleotide content between compartments could bias our comparison and calculation of local mutation rates. For example, a particular compartment could have a relatively high count of AAA>ACA SNVs, but this high count might be caused by the compartment having many occurrences of the triplet AAA, and therefore more opportunities for an AAA>ACA to occur. Thus, we rescaled the number of mutations for each compartment by the nucleotide content of the NRNC compartment before calculating fractions. To calculate the rescaled rate $r(m)$ of mutation m_i: $\{m_1, \ldots, m_{96}\}$ corresponding to triplet t_t: $\{t_1, \ldots, t_{32}\}$ and compartment C:

$$r(m_i) = \frac{R_{i,C}}{\sum_j R_{j,C}}$$

$$R_{i,C} = \#m_{i,C} * \frac{\#t_{t,\mathrm{NCNR}}}{\#t_{t,C}}$$

We calculated triplet content of each compartment by sliding a 3bp window, 1bp at a time, across each compartment. Triplets whose central mutation was contained within the boundaries

of a compartment segment but whose 3' or 5' flanking mutations fell beyond and triplets with N's were excluded.

Several statistical analyses comparing two different mutation spectra required count data rather than frequency data (e.g., Chi-square tests). We devised a slightly different rescaling strategy for these counts to avoid artificially inflating the mutation counts. To calculate the rescaled count of mutation m_i: $\{m_1, \ldots, m_{96}\}$ corresponding to triplet t_t: $\{t_1, \ldots, t_{32}\}$ and compartment C_1 in preparation for comparison to compartment C_2, we scaled down the raw count of mutation m in the compartment where m is more abundant, rather than scaling up the count of m in the compartment where it is more abundant:

$$R_{i,C_2} = \begin{cases} \#m_{i,C_1} * \frac{\#t_{t,C_2}}{\#t_{t,C_1}}, \#t_{t,C_1} \geq \#t_{t,C_2} \\ \#m_{i,C_1}, \#t_{t,C_1} < \#t_{t,C_2} \end{cases} \quad (2.1)$$

The same rescaling is used for compartment C_2, switching the subscripts accordingly.

2.4.4 *Statistical analyses*

We generated plots and performed statistical analyses in R (version 3.1.0) using scripts available at https://github.com/harrispopgen/gagp_mut_evol.

We ran **PCAs** on matrices ($k \times C$ rows by 96 columns, for k individuals and C compartments) of rescaled 3-mer mutation rates calculated for each individual and each compartment using the *prcomp* method. Some PCAs were run on individuals from all species; others were only run on individuals from a single species. The matrices were centered and scaled, as is standard for *prcomp*. The **PCA loading heatmaps** display the weights associated with each of the 96 3-mer mutation types for a given PC. The **Euclidean distance ridge plots** were similarly generated with individual mutation spectrum data that required no rescaling since

the analysis considered only a single compartment. We plotted the distribution of Euclidean distances based on a 2×96 matrix comparing the mutation counts between two different individuals of either a single or two different species. For comparisons within species, we ran *k* choose 2 tests (*k* being the number of individuals for a given species); for comparisons between two species, we ran $k \times l$ tests (*k* and *l* being the numbers of individuals in both species respectively).

The **log-odds heatmaps** were generated to display the relative enrichment or depletion of specific mutation types when comparing two compartments directly to each other. For a given species, we plotted the log transform of the ratio between the rescaled mutation rates of two compartments.

The **7-mer content-corrected heatmap** required 7-mer mutation and nucleotide content from various compartments, which were generated using the 3-mer mutation and nucleotide methods and equally expanding context on the 5' and 3' side of the central base. Each original 3-mer mutation $m_{3,k}$: {AAA>ACA, AAA>AGA … TCT>TTT} is a collapsed equivalence class of 256 unique 7-mer mutations $m_{7,i,x}$: {AAAAAAA>AAACAAA, AAAAAAC>AAACAAC, … TTAAATT>TTACATT}. To explicitly re-weight the counts of each 3-mer mutation $m_{3,i,C}$ in compartment *C* using the ratio of 7-mer content in *C* $s_{s,C}$: {AAAAAAA, AAAAAAC, … TTTCTTT} to that of compartment *C'*:

$$R_{7,i,C} = \sum_l \#m_{7,l,C} * \frac{\#s_{s,C\prime}}{\#s_{s,C}} \tag{2.2}$$

$$r\left(m_{3,i}\right) = \frac{R_{7,i,C}}{\sum_j R_{7,j,C}} \tag{2.3}$$

We used this method to rescale 3-mer mutations from the ERV compartment to match the 7-mer content from the nonrepetitive heterochromatin compartment; the heatmap in Figure S17 presents the log ratio of the rescaled ERV mutation and the non-rescaled nonrepetitive heterochromatin mutation spectra.

We ran **correlation analyses** to quantify the similarities between the mutation spectrum heatmaps between species. Each mutation spectrum heatmap comprised the log ratio between each of the 96 mutation types between two compartments for a single species. To calculate the similarity between heatmaps for two different species, we ran a Pearson correlation test on the paired vectors of log odds.

2.4.5 *Compartments*

We defined compartments based on published annotations of genomic features. Each compartment was a list of genomic segments in a bed file format. The following is a list of the compartments used in our analyses and a short description of how we generated them.

NCNR: (non-conserved, non-repetitive) the entire hg18 genome, excluding repetitive elements defined by repeatMasker, conserved regions in primates based on the phastCons 44-way multi-species alignment, CpG islands from the UCSC genome browser, and coding exons from refGene.

ERVs: all ERVs in repeatMasker run on hg18, excluding those classified specifically as mammalian long-terminal repeats (MaLRs). MaLRs are believed to be largely inactive in great apes, unlike ERVs which are still active in several species.

LINEs: all LINEs in repeatMasker run on hg18

Heterochromatin: the intersection of the heterochromatin domain called in the hg18 chromHMM run on 9 different cell types from ENCODE (Gm12878, H1 HESCs, HepG, Hmec, Hsmm, Huvec, K562, Nhek, and Nhlf), minus repetitive elements defined in repeatMasker.

Early/late replicating regions: the genomic quartiles that replicate earliest and latest during S phase were ascertained using replication timing data from Koren et al., 2012 [55]. In that manuscript, the fine-scale replication timing of regions in the genome was determined by sequencing human lymphoblastoid cell lines at S1/G phase; read depth over a region in the genome corresponded to its average relative replication timing. The read depths were measured at specific genomic positions. We calculated average replication timing for non-overlapping 20kb window that included at least one measurement of replication timing and calculated replication timing quartiles. The earliest and latest quartiles were used as compartments.

Early/late replicating, repetitive/non-repetitive regions were the subsets of the replication timing compartments that overlapped with or excluded repeatMasker-annotated repeats, respectively.

Human maternal mutation hotspots: defined in Jonsson et al., 2017 as regions whose *de novo* mutation rate strongly associates with maternal age, lifted over to hg18 [33].

ERVs ± 5hmC: ERV compartment, split into segments that had or lacked evidence of ≥ 1 5hmC site, based on Tet-assisted bisulfite sequencing of human ESCs [87].

2.4.6 *Quality control analyses*

We ran a number of analyses to test the robustness of our methods and findings.

Determining how GC-biased gene conversion contributes to the separation of species and compartments in PCA plots

GC-biased gene conversion (gBGC) is the nonreciprocal copying of short DNA tracts between homologous chromosomes during meiosis, which at heterozygous sites tends to retain G/C (strong, S) alleles more often than A/T (weak, W) alleles [92–94]. This process causes G/C derived alleles to have higher substitution rates and frequencies than A/T alleles. If the strength of gBGC were not conserved across the great ape phylogeny, we might expect species' and compartments' mutation spectra to separate along an axis defined by the ratio of W-to-S and S-to-W mutation types, with species and compartments that experience the most gene conversion having the highest proportions of S-to-W variants. Such a gradient would be expected to dissipate, however, if we constructed mutation spectra using a higher proportion of rare variants, which are younger than common variants and have had less time to be influenced by the effects of gBGC. To this end, we tabulate mutation spectra in the following way: each SNV present in k haplotypes is counted toward the mutation spectrum for only one randomly selected haplotype; this process is described above ("computing the mutation spectra of individuals and species") and hereafter referred to as "randomization." We observe that individuals cluster by species in a PCA run on NCNR mutation spectra whether we employ randomization or count each SNV toward every haplotype on which the derived allele appears. Furthermore, the principle component loadings are not consistent with the expected signature of gBGC. Many of the mutation types that have different mutation fractions in different ape species are W-to-W or S-to-S mutation types that are expected to be unaffected by gBGC (Figure S2). Even if we compute mutation spectra entirely from rare doubleton variants, we observe similar species separation to what we observe using more common variants, despite these rare variants being more weakly affected by gBGC.

The rates of gBGC covary across the genome with recombination rate. The locations of recombination and gBGC hotspots change rapidly and are often species-specific due to evolution of the gene *PRDM9*. Thus, comparing mutation spectra between species in locations where gBGC rates are most divergent should (A) indicate an upper bound on the effects of gBGC rate evolution on our mutation spectrum analyses and (B) test our capacity to minimize its effects through the randomization method. We therefore examined the mutation spectra of species-specific recombination hotspots with and without our random sampling method (Figure S1). These analyses showed that, without random sampling, differences in mutation spectra within recombination hotspots are in fact dominated by the *cis*-acting effects of GC-biased gene conversion (gBGC). Figure S1A shows a PCA of individual mutation spectra from the NCNR compartment and two compartments containing the genomic regions whose recombination rates fall within the upper and lower genome-wide deciles in humans (Kong et al., 2010); this spectrum is not thinned by randomization. We can clearly see the mutation spectra cluster by species, but we can also see high-recombination-rate and low-recombination rate compartments separate along PC2, especially in humans where the ascertained recombination hotspots are active. PC2's loadings are dominated by A>G and A>C mutations, which are both W>S mutations and likely indicate a human-specific enrichment for gBGC. In contrast, the separation between high recombination and NCNR compartments is mostly attenuated in Figure S1B with the use of the randomization method.

Comparing SNV to de novo mutation spectra

Several papers have recently reported homogeneity among the de novo mutation (DNM) spectra of great apes. Our analyses, however, demonstrate that mutation spectra generated from

SNVs vary significantly between species. We therefore compare our SNV to DNM spectra from Besenbacher et al. (2019) to assess the difference in results and find that DNM spectra are underpowered to detect the species signatures we find in SNV spectra.

We test the likelihood that the distribution of the observed DNM counts in an individual is pulled from the SNV spectrum of a given species *s*. We calculate the expected probabilities *p* in [$p_{A>C,s}$, $p_{A>G,s}$, … $p_{C>T,s}$], given the number of SNVs segregating in species *s* in the GAGP of mutation type $m_{i,s}$ in [$m_{A>C}$, $m_{A>G}$, … $m_{C>T}$] by the following equation:

$$p_{i,s} = \frac{\#m_{i,s}}{\sum_j \#m_{j,s}} \tag{2.4}$$

We then calculate the log likelihood that the spectrum of DNMs for individual *x* ($m_{i,x}$) assuming a multinomial distribution. This likelihood represents how well the DNM spectrum for an individual fits the SNV spectrum of a given species.

$$P_{s,x} = \sum_i \#m_{i,x} * \log(p_{i,s}) \tag{2.5}$$

To determine whether a SNV spectrum fit a DNM spectrum significantly better than others, we calculated the significance of the differences in likelihoods for a given individual when compared to chimpanzees, orangutans, and gorillas (x : [*C, O, G*]). We simply calculated the fold range in fit as:

$$\exp\left(\text{range}(P_{C,x}, P_{O,x}, P_{G,x})\right) \tag{2.6}$$

A "significance" threshold was set at a fold range of 20x, to replicate a standard α value of 0.05. The table below shows that none of these values approach 20 (Table S1). We conclude that the DNM spectra are too sparse to demonstrate clear species differences.

Testing the robustness of PCA clustering to species representation

We tested the robustness of the clustering of individuals by species in the NRNC PCA to differences in number of individuals sequenced per species. We down-sampled the number of individuals per species to match those of humans ($n = 9$), excluding the two orangutan species who were grouped as a single super-species group for this analysis. The clustering patterns in the PCA remained.

We recreated the several plots from Figs. 1 and 2 using a rarer subset of variants and found no qualitative differences when compared to the original figures (Figure S6A). To avoid the potential quality issues related to relying on singletons, we used only doubletons (DAF = 2/2*N*, for *N* individuals sequenced from each species) in this analysis. The PCA of the individual mutation spectra across all species in the NCNR compartment alone still demonstrates the clustering of individuals by species (Figure S6B). The distribution of Euclidean distances between the NCNR mutation spectra of individuals within and between species still demonstrates the correlation of mutation spectrum distance with divergence time (Figure S6B). The PCA of individual mutation spectra from the NCNR, early replication timing, and late replication timing compartments still demonstrates similar trends as observed in figure 2D (Figure S6C). The first PC, representing the greatest axis of variance among all spectra, separates spectra according to

phylogeny and represents species signatures; the second PC separates compartments by their replication timing. The loadings for PC2 are enriched for C>A and A>T mutation types, corresponding with the known late replication timing signature (Figure S6D). We attribute the spread observed in chimpanzees in the replication timing doubleton PCA below to be largely an effect of noise imposed by down-sampling.

Testing the CG>GG signature in ERVs

We determined the enrichment for CG>GG mutations in ERVs compared to nonrepetitive heterochromatin was unaffected by the differences in mapping quality between the two compartments, noting that mutations in repetitive regions are more difficult to call confidently. To determine the potential confounding effect of mapping quality on the CG>GG signature, we compared the distribution of the mapping quality value of the variants in the ERV and nonrepetitive heterochromatin compartments (MQ field in the GAGP vcf). Density distributions of the mapping qualities for the two compartments were highly overlapping within each species (Figure S20).

We also determined that the CG>GG signature was unaffected by the different 'shapes' of the ERV and nonrepetitive heterochromatin compartments, i.e. the distribution of segment lengths and overall size of compartment. For each ERV compartment segment for a given chromosome, we reassigned the coordinates randomly within the nonrepetitive heterochromatin compartment, preserving segment length. In the event a randomized compartment was chosen to overlap one or more "N" bases, a new compartment was resampled. We did not filter for overlapping randomized segments, but assumed that, given that the ERV compartment was a fraction of the size of the heterochromatin compartment (127 Mb and 430 Mb respectively),

collisions would be unlikely enough to that their sparsity would not bias our findings. This randomization process to create 'ERV-like' compartments was bootstrapped 100 times. We then calculated nucleotide content and mutation spectra for each of the 100 bootstrapped compartments. We generated the log-odds of the CG>GG mutation type between the bootstrapped compartments and nonrepetitive heterochromatin compartment (each normalized to each other, using the Chi-square normalization method as described above), then compared those values to the same statistic of the original two compartments (Figure 4C) for all species. The observed CG>GG enrichment in the ERV compartment lies outside of the null distribution in all species except Bornean orangutans, for which the confidence interval of the observed log-odds overlaps the null distribution. Given that the observed CG>GG fraction still falls outside of the null distribution, we conclude that Bornean orangutans show the same trend as seen in other species, but that trend is not significant on its own. This lower enrichment in Bornean orangutans may be a result of their low sample size (2N = 10) and the paucity of observed segregating variation (65% that of Sumatran orangutans, 30% that of gorillas, e.g.). The enrichment for CG>GG mutations comparing ERVs to nonrepetitive heterochromatin is significantly stronger than the enrichment in any of the bootstrapped, 'ERV-like' compartments (Table S6).

We tested the effect of species-specific nucleotide content on our rescaling method. The GAGP data are aligned to hg18; therefore, mutations in genomic regions in a non-human species that do not exist or do not map well in humans (e.g., new repetitive elements) were absent from our analyses. Our mutation rescaling method, however, relies on compartment nucleotide content determined from the hg18 reference, therefore including regions specific to humans and absent in other species. We compared the log-odds of each mutation type between the ERV and nonrepetitive heterochromatin compartment in all six species with those calculated by rescaling

mutation counts using the nucleotide content of the segments of a given compartment that successfully lifted over to each respective species (lifted from hg18 to gorGor4, panPan1, panTro2, and ponAbe2 for gorilla, bonobo, chimpanzee, and both orangutans, respectively, using default liftOver settings). The log-odds values within species are highly significantly correlated (all $\rho \geq 0.95$, Figure S19).

2.5 Data availability

No new data were generated or analyzed in support of this research. The processed mutation spectra and analysis pipelines are available in a public GitHub repository at https://github.com/harrispopgen/gagp_mut_evol.

2.6 Acknowledgements

This work was supported by the National Institute of General Medical Science at the National Institutes of Health (T-32 GM081062 to M.E.G., 1R35GM133428-01 to K.H.); the Burroughs Wellcome Fund (a Career Award at the Scientific Interface, to K.H.); the Pew Charitable Trusts (Biomedical Scholarship, to K.H.), the Searle Scholars Program (Career Award, to K.H), and the Alfred P. Sloan Foundation (Research Fellowship, to K.H). We thank Evan Eichler, Phil Green, Sharon Browning, and members of the Harris lab for helpful discussions. We also thank Noah Snyder-Mackler and Aylwyn Scally for manuscript comments and Shwetha Murali for technical assistance.

2.7 FIGURES

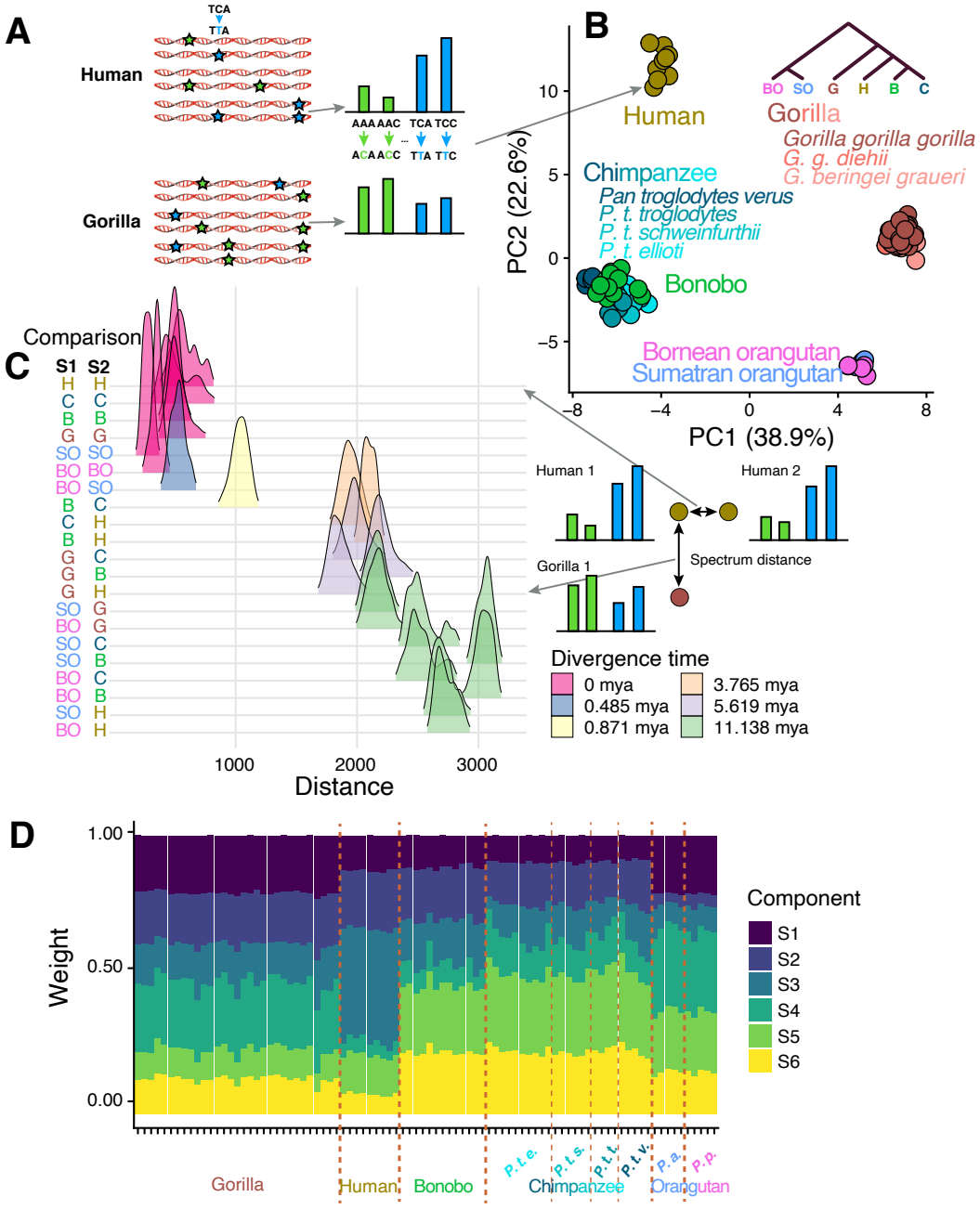

A
Human
Gorilla
TCA
TTA
AAA AAC TCA TCC
ACA ACC TTA TTC
B
Human
Gorilla
Chimpanzee
Bonobo
Bornean orangutan
Sumatran orangutan
Gorilla gorilla gorilla
G. g. diehii
G. beringei graueri
Pan troglodytes verus
P. t. troglodytes
P. t. schweinfurthii
P. t. ellioti
BO SO G H B C
PC2 (22.6%)
PC1 (38.9%)
C
Comparison
S1 S2
Distance
Human 1
Human 2
Gorilla 1
Spectrum distance
Divergence time
0 mya
0.485 mya
0.871 mya
3.765 mya
5.619 mya
11.138 mya
D
Weight
Component
S1
S2
S3
S4
S5
S6
Gorilla
Human
Bonobo
Chimpanzee
Orangutan

Figure 2.1. Covariance of species-specific and replication timing mutation spectra in great apes

A. SNVs segregating within a species were counted to generate a triplet mutation spectrum for each individual in the GAGP. We include SNVs found in non-conserved, non-repetitive (NCNR) regions of the genome, which we collectively call the NCNR compartment.

B. PCA of NCNR compartment mutation spectra reveals clustering of individuals by species. Each point represents the NCNR compartment mutation spectrum from a single individual in the GAGP; colors represent species, while shades of a color represent subspecies (applies only to gorillas and chimpanzees).

C. Mutation spectra are more similar between individuals of the same species than between individuals from different species. We plotted the Euclidean distances between triplet mutation spectra of all possible pairs of individuals in the GAGP within and between species (see Methods).

D. When nonnegative matrix factorization (NMF) is used to infer mutational signatures that best explain variation among ape mutation spectra, we see more variation of signature composition between species than within species and identify signatures that change in dosage on specific phylogenetic tree branches. For example, signature 1 appears to have decreased in dosage on the branch ancestral to humans, chimps, and bonobos.

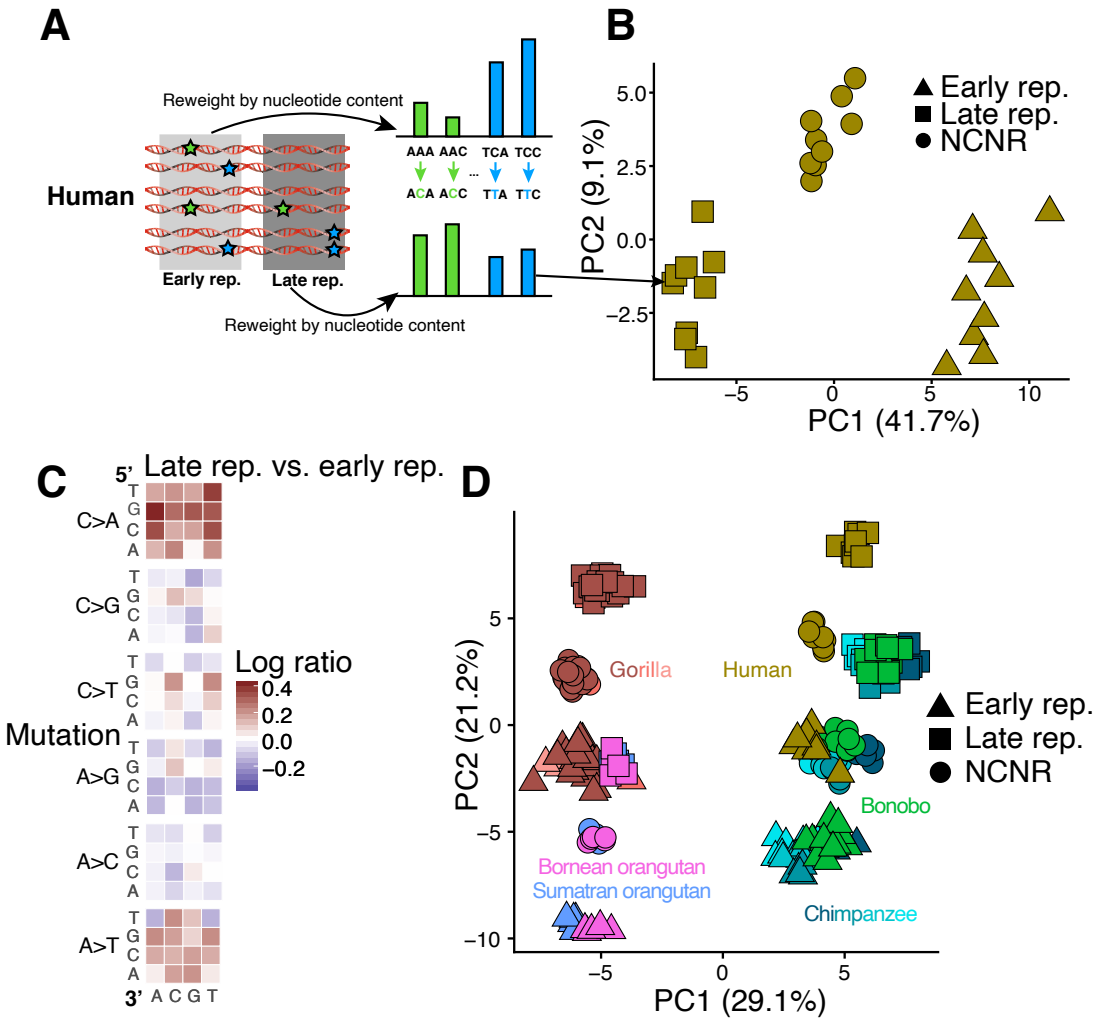


Figure 2.2. Signatures associated with replication timing appear conserved among great apes

A. We calculated separate mutation spectra for each individual in compartments that replicate early versus late in S phase and re-weighted each spectrum by the trinucleotide content of the associated compartment.

B. Each point in this PCA represents the mutation spectrum from a single individual's NCNR, early replicating, or late replicating compartment. The observed gradient results from differences in mutation spectra between early-replicating and late-replicating compartments.

C. A heatmap of the log ratios of triplet mutation fractions in humans shows an enrichment for C>A and A>T mutations in the late replicating compartment compared to the early replicating compartment. This mutation signature recapitulates recently described late replication timing signature in humans. To generate the species mutation spectra, we counted the number of SNVs with triplet context segregating within a species that occurred in each compartment. The triplet mutation fractions were normalized by compartment nucleotide content.

D. Mutation spectra from late and early replication timing and NCNR compartments from each individual in the GAGP separate along a first PC associated with phylogeny and a second PC associated with replication timing. The mutation signature associated with late replication timing appears conserved among all great apes. Different shades of each species' color represent subspecies, as in Figure 1D.

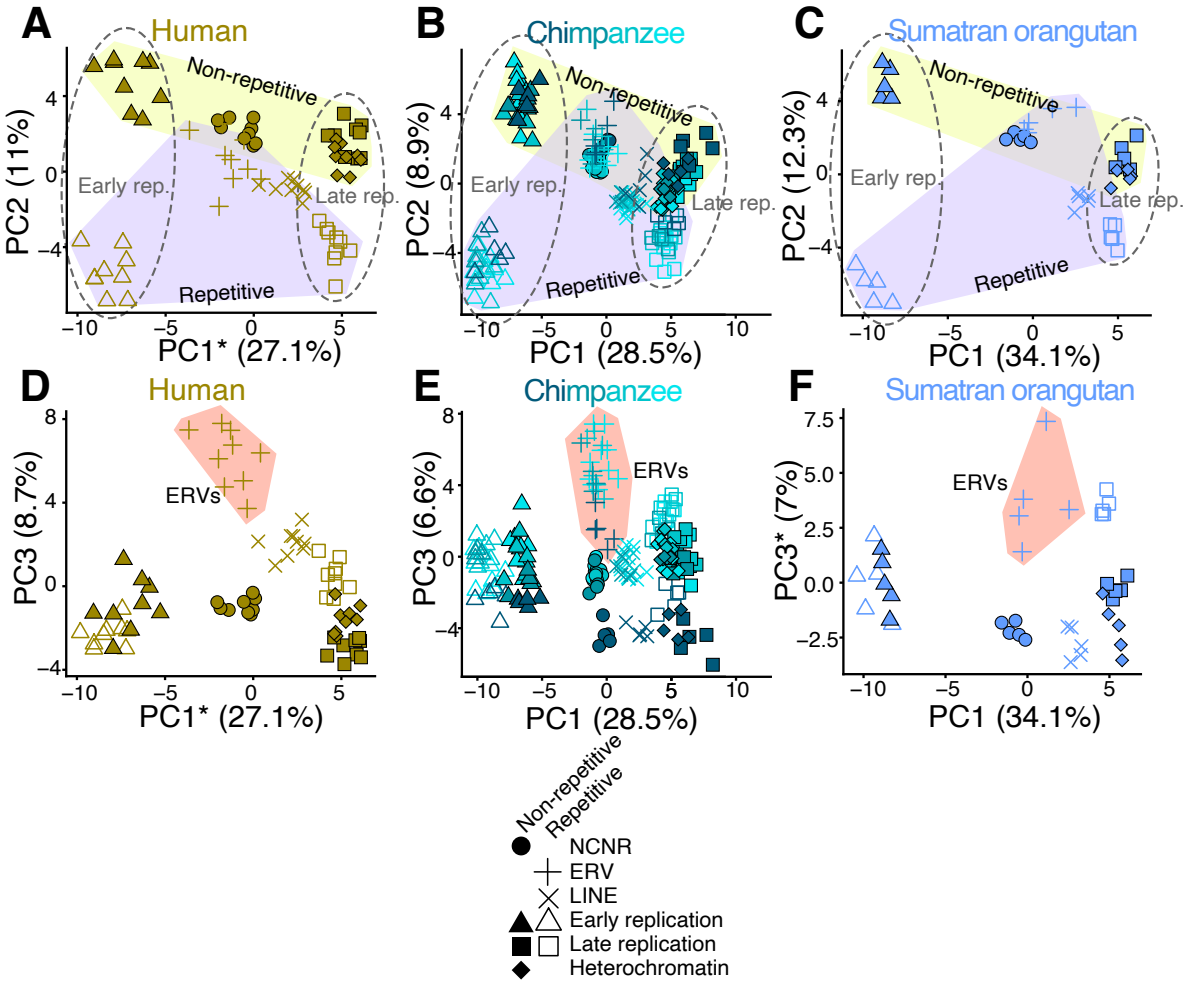


Figure 2.3. Conserved axes of mutation spectrum variance among great apes

A-F. We defined eight overlapping functional compartments to test for evolution of mutation spectrum modifiers along axes of chromatin accessibility, replication timing, and repetitive content. We then ran a PCA on the individual mutation spectra for all eight compartments for each species separately (only human, chimpanzee, and Sumatran orangutan shown). For all species, PC1 and PC2 separate compartments along gradients that correspond to replication timing and repetitive content, respectively (dotted lines vs. shaded polygons, A-C). PC3 separates ERVs from other compartments (shaded polygons,

D-F). The similarities of these independent PCAs across all species implies conservation of *cis*-acting mutational signatures. *Axis inverted for readability.

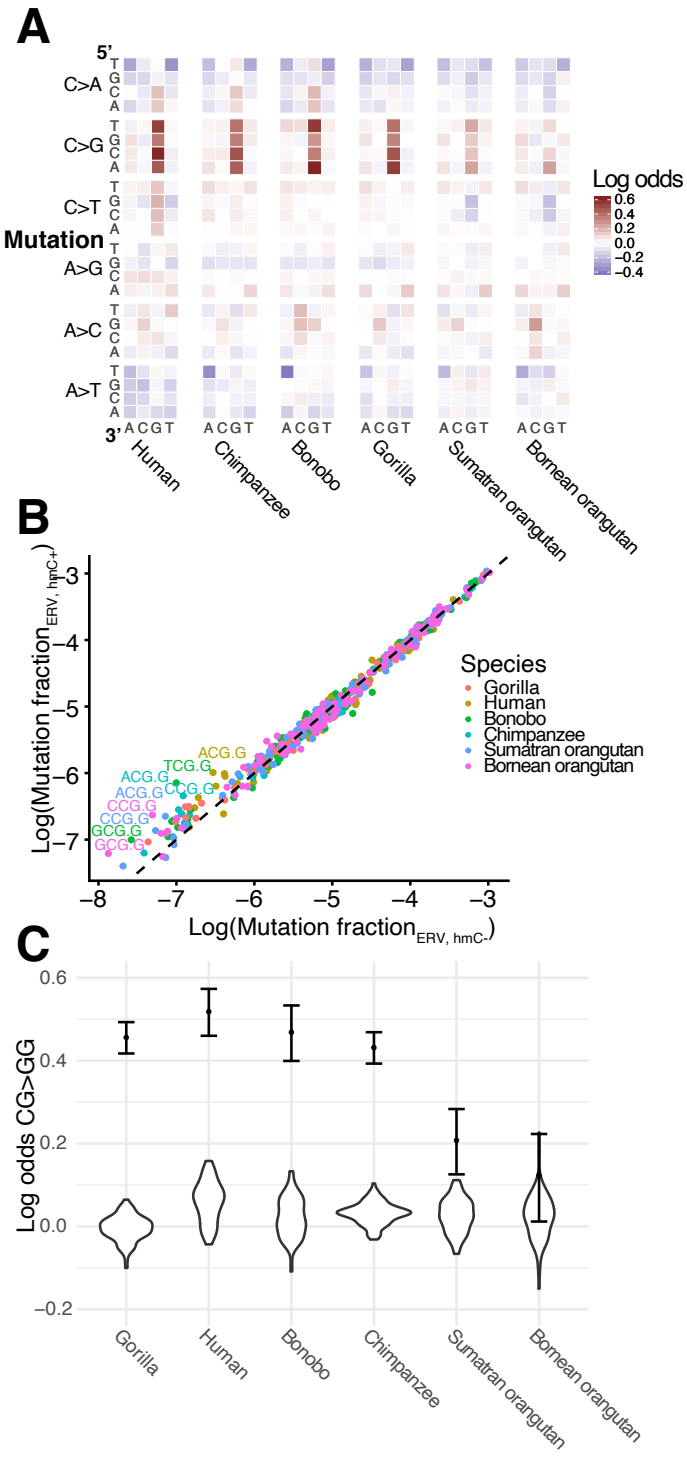

A
5'
3'
C>A
C>G
C>T
A>G
A>C
A>T
Mutation
T G C A
A C G T
Log odds
0.6
0.4
0.2
0.0
−0.2
−0.4
Human
Chimpanzee
Bonobo
Gorilla
Sumatran orangutan
Bornean orangutan
B
Log(Mutation fraction ERV, hmC+)
Log(Mutation fraction ERV, hmC-)
−3
−4
−5
−6
−7
−8
ACG.G
TCG.G
ACG.G
ACG.G
CCG.G
CCG.G
CCG.G
GCG.G
GCG.G
Species
Gorilla
Human
Bonobo
Chimpanzee
Sumatran orangutan
Bornean orangutan
C
Log odds CG>GG
0.6
0.4
0.2
0.0
−0.2
Gorilla
Human
Bonobo
Chimpanzee
Sumatran orangutan
Bornean orangutan

Figure 2.4. A hydroxymethylation-related CG>GG mutation signature distinguishes ERVs from other compartments

A. Heatmaps of log odds of triplet mutation spectra comparing ERV to nonrepetitive heterochromatin compartments for each species show a significant ERV-specific CG>GG mutation signature.

B. Enrichment of CG>GG mutation types in ERVs with hydroxymethylation compared to ERVs without. Points represent the fraction of each triplet mutation in ERVs with and without hydroxymethylation calculated from SNVs segregating within each species (*y* and *x* axis, respectively). Mutation types that fall along the $y = x$ line occur equally frequently in both compartments. Mutation type labels are included only for mutation types whose log ratio of ERV hmC+:ERV hmC- exceeds 0.4. Points are colored by species.

C. The CG>GG mutation signature in ERVs is robust to the size and shape of the ERV compartment. We created 100 "ERV-like" compartments by sampling segments corresponding to the size of those in the ERV compartment from random locations within the nonrepetitive heterochromatin compartment. The distribution of the log odds of CG>GG mutations between these ERV-like to the original nonrepetitive heterochromatin compartment are violin plots. The log odds of CG>GG mutations between the original ERV and nonrepetitive compartment are plotted as dots for reference, with 95% confidence interval.

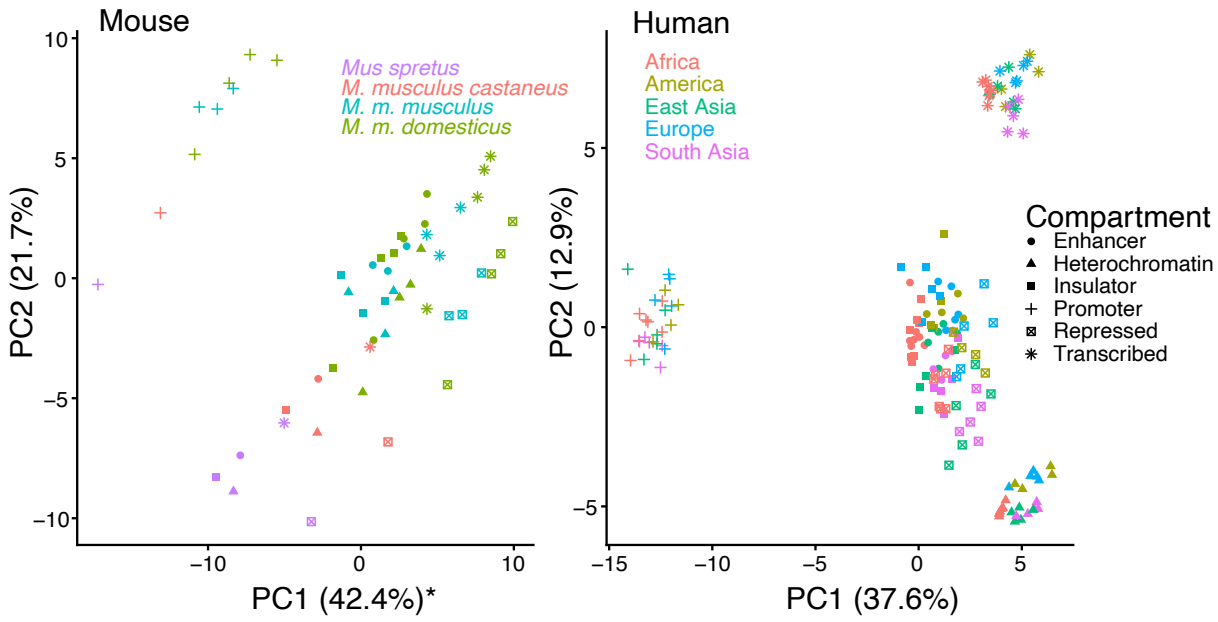


Figure 2.5. Structure of mutation spectrum variation is conserved between mouse and human

A-B. We ran PCA separately for mutation spectrum data from mouse (A) and human (b) individuals, whose genomes were split into compartments annotated by chromHMM (PC1 of mouse reversed for readability). The relative positioning of compartment mutation spectra within a species or subspecies are similar, particularly after a roughly 30 degree rotation for the mouse PC1 and PC2. For both species, PC1 separates promoters from other compartments, while PC2 distinguishes transcribed regions. The loadings of PC1 and PC2 are significantly correlated between the two separate PCAs (rho = .486, 0.555, $p = 5.14*10^{-7}$, $4.44*10^{-9}$, respectively) (Supplemental Figure S18).

Chapter 3. VARIATION IN *DE NOVO* SHORT TANDEM REPEAT MUTAGENESIS IDENTIFIED IN PEDIGREES

3.1 INTRODUCTION

Short tandem repeats (STRs) are genomic elements composed of repeating nucleotide motifs of 1-20 base pairs [95]. Also known as micro- and minisatellites, their unusually high mutation rates enabled studies of genetic diversity in the pre-genomic era [96,97]. Simple repeats compose nearly 9% of the human genome, and their mutagenesis is implicated in developmental disorders [95,98–100]. Despite this classical importance, STRs have fallen behind relative to single nucleotide polymorphisms (SNPs) in terms of our understanding of how mutation rates vary in humans as a function of life history and population of origin [17,18].

Both SNP and STR mutation rates depend strongly on paternal age [45,101]. This dependency was classically explained by the fact that male germ cells replicate throughout the reproductive lifespan, assuming that most mutations originate as errors in replication [1,28]. In the case of SNPs, recent work has cast doubt on this model and has begun to support DNA damage as a substantial source of SNPs [1,28–30]. The proportion of SNPs inherited from the paternal lineage, known as *a*, remains strikingly stable with parental age, which should not be the case if most mutations are replicative in origin and the ratio of paternal to maternal cell divisions increases with age [29]. Although the female germline does not continue to replicate after birth, the ratio of male to female germline mutations appear to remain largely constant regardless of developmental stage.

Although we have less information on how life history impacts STR mutagenesis, the high mutation rate at these loci has long been attributed primarily to polymerase slippage during replication [101,102]. Replication slippage typically causes STR expansions or deletions by one

or more repeat units [95,103–105]. Past studies have found no association of STR mutation rates with maternal age but have found a strong association with paternal age (these associations are called maternal and paternal age effects) [101,104]. At face value, these age associations seem to bolster the theory that STRs mutate only during S phase DNA replication. Given the lack of maternal germline replication after the future mother's birth, we would expect replication-driven mutations to exhibit a paternal age effect but no maternal age effect [28]. However, previous studies of STR parental age associations draw upon much less data than comparable studies of single nucleotide mutation rates.

Mutagenesis has been shown to vary with genetic ancestry as well as parental age in humans, other great apes, and many other species [17,18,24,51,106]. This variation might result from genetic modifiers of mutagenesis, and might also be impacted by environmental factors that correlate with genetic background. Teasing apart these contributing factors will be a critical part of elucidating whether different populations and species vary in their susceptibility to genetic disease. Although continental ancestry differences have not been shown to covary with modern day germline mutagenesis in humans, two recent studies have found evidence for genetic mutators alleles that affect DNMS in specific individuals: a lower mutation rate in an Amish population, and a mutator allele that affects indels at TNT repeats in some families [49,53]. Furthermore, the strength of parental age contribution to SNP mutation rate can vary between families, possibly attributed to genetic background [47,48]. It is not currently known whether STR mutation rates vary among populations or families in any way that is not explainable by reproductive age.

One reason why so little is known about human STR mutation rate variation is the complicating effect of bioinformatic errors, which affect STR detection to a greater extent than

SNP detection. Different STR alleles can have different probabilities of detection depending on repeat length, flanking sequence composition, and other factors [95,104]. These detection asymmetries can lead to allelic dropout, a phenomenon in which one allele at a heterozygous locus is not successfully sequenced and called, resulting in a false homozygous genotype call. In older studies that genotyped STRs using PCR technology, allelic dropout was most commonly associated with a polymorphism in the binding site of a primer used in PCR amplification, resulting in selective amplification failure of the linked haplotype [107–109]. However, modern next generation short read sequencing (NGS) is amplification-free but suffers from other technological biases that interfere with the genotyping of certain other STR alleles [110,111]. In PCR-free NGS data, allelic dropout disproportionately affects long alleles that short reads have difficulty spanning with sufficient coverage [95].

In this study, we sought to leverage modern NGS data to measure variation of STR mutation rates as a function of parental age and population of origin, as well as to estimate the impact of allelic dropout on STRs as a function of characteristics such as allele length and repeat unit size. To do so, we analyzed a dataset of *de novo* mutations (DNMs) at STR loci found in 1593 families comprising two parents and two children (quads), one of whom has been diagnosed with autism spectrum disorders (ASD) without prior family history (simplex case) [112]. These families were recruited, in part, to study how rare and *de novo* mutations contribute to ASD as a subset of the Simons Simplex Collection (SSC) [112]. Recent findings show that STR DNMs are more numerous and are more selectively deleterious than STR DNMs in unaffected siblings, and thus may contribute significantly to ASD [95]. This work is expected to fill in a blind spot about how age and genetic variation impact autism risk.

3.2 RESULTS

3.2.1 *Short tandem repeat mutation rates do not perfectly track cell divisions but covary with maternal age*

We looked at STR *de novo* mutations in a cohort of 1593 quad families recruited as a part of the Simons Simplex collection based on the diagnosis of a simplex case of ASD in a single child [112]. These probands have been shown to harbor an enrichment of *de novo* mutations that may affect neurodevelopment and lead to ASD [110]. We used mutation calls from Mitra et al., 2021. As described in Mitra et al.'s methods, the families were sequenced to 35x coverage, and diploid genotypes at STR loci were inferred using GangSTR [95,112]. This database of STR DNMs is the largest compiled to date. DNMs were called using MonSTR, a likelihood-based caller with an accuracy of 90%, calculated based on capillary electrophoresis [95]. Stringent filters were placed on the genotype and DNM calls as described in Mitra et al., 2021. Mutations were phased to their parental lineages based solely on the observed alleles in parents and child without explicitly recreating the surrounding haplotype. If one allele in a child could be assigned unambiguously to one parent's lineage, the mutation was phased to the other parent's lineage.

We separately regressed the number of maternally and paternally phased mutations per child against their mother's and father's ages at their birth, respectively, using Poisson regressions with identity link functions. Unless otherwise specified, all further Poisson regressions in this study used an identity link function. Parental age at the time of a child's birth served as a proxy for parental age at the time of conception. Our analysis confirms the strong paternal age effect on the rate of paternally phased STR mutations that has been reported in this and other STR DNM datasets (Fig 3.1). More surprisingly, we also observe a weak but significant effect of maternal age on maternally-inherited STR DNMs (Fig 3.1).

Prior studies have found no significant maternal age effect on the STR DNM rate, but most have used several orders of magnitude fewer loci and DNMs [101,104]. To determine if these prior studies were simply underpowered to discover the maternal age effect, we repeated our regression-based inference of the DNM rate after restricting to the subset of loci analyzed in a prior study by Sun, et al. that found no maternal age effect [104]. Using this subset, we were unable to find a significant maternal age effect, indicating that previous studies may indeed have been under-powered (Supplementary figure 1) ($P = 5.16*10^{-7}$, 0.91 for paternal, maternal age effects; linear regression). Maternal age remains a significant predictor of the maternal STR mutation rate even when we include paternal age as a covariate, indicating that phasing errors are not likely to explain the signal (P = 2.93e-05).

Recent studies of de novo SNVs have found that the fraction of phased mutations arising from the paternal lineage, known as *a*, does not depend on parental age; instead, the father contributes about 3/4 of mutations regardless of the total mutation load or the ages of the parents [29]. In contrast to this, we find a significant positive correlation between α and paternal age, indicating that the paternal STR mutation rate appears to accelerate over time relative to the maternal mutation rate (Supplementary Figure 2).

Part of the maternal age effect on single nucleotide substitutions has been attributed to a higher rate of post-zygotic mutagenesis in older mothers [29]. To examine whether the same might be true for STRs, we tested for covariance between maternal age and the mutation rate of STRs occurring on paternally inherited chromosomes, controlling for variance in paternal age [29]. After controlling for paternal age, any correlation of mutation rate on paternally inherited chromosomes with maternal age could indicate that older mothers harbor a higher postzygotic mutation rate. To do this, we sampled pairs of children born to fathers of the same age (+/- 6

months) and calculated the differences in maternal ages as well as the additional number of STR DNMs phased to the father identified in the child born to the older mother. We detected no significant positive correlation between maternal age and paternally phased STR DNMs, conditional on equal paternal age. This analysis provided no evidence that advanced maternal age causes additional post-zygotic mutations to accumulate on paternally inherited chromosomes (one-sided Spearman's correlation test, $P = 1$, $\rho = -1.15*10^{-2}$).

Certain mutational pathways display biases towards expansions or deletions [11]. Although deletion mutations (loss of repeat unit(s)) are less common than expansion mutations (gain of repeat unit(s)) in the maternal lineage, we observed that their mutation rate is more strongly associated with maternal age than the rate of expansion mutations. Poisson regressions found maternal age at birth to be significantly associated with the maternal deletion rate but not the maternal expansion rate ($P = 3.02*10^{-5}$, 0.0629; slopes of 0.022, 0.015 DNM/year, respectively) (Fig 3.2). In contrast, we found paternal age to be significantly associated with both paternal expansions and deletions ($P = 4.75*10^{-34}$, $1.57*10^{-24}$; slopes of 0.09, 0.094 DNM/year, respectively). Although the paternal age effect is not significantly different between deletions and expansions, a higher proportion of expansion DNMs accumulate before birth than deletions (P = 1.26e-10, 0.753; effects = 2.50, 0.0037 for intercepts, slopes, respectively; ANCOVA assuming Poisson distribution with identity link). The fraction of expansions and deletions that phased to the paternal lineage were both separately associated with paternal age ($P = 2.7*10^{-3}$, $1.42*10^{-3}$; slopes of $3.01*10^{-3}$, $2.47*10^{-3}$ per year) (Supplementary figure 2). In addition, we see evidence that maternal age seems to have a nonlinear effect on the deletion rate that accelerates over time. A Poisson model with a log link function, explicitly modeling an exponential effect of maternal age, fit the data better than the model with an identity link function ($\Delta AIC = -1$). This

stronger association with older maternal age mirrors maternal age effects on damage-associated C>G single nucleotide substitution DNMs [29].

The above regression analyses all suggest that STR insertions and deletions are differently impacted by the mutagenic effects of increasing parental age. However, because the phasing method did not explicitly infer the haplotype surrounding the STR, some ambiguity exists in the directionality (i.e. expansion vs. deletion) and size of STR DNMs. Each *de novo* STR mutation was phased to its parent of origin using a parsimony approach that assumes the most probable mutation is the one that changes a parental allele by the smallest possible number of repeat lengths. For example, in a trio with maternal, paternal, and proband genotypes of [12, 12], [10, 13], and [12, 11], respectively, the "12" allele is most likely inherited from the mother whereas the "11" allele is most likely a mutant of the paternal "10" allele that was changed by an insertion of one repeat unit. While this assumption is parsimonious, the "11" allele might also be a deletion of two repeat units from the parental "13" allele. However, if the child's genotype were [12, 9], the mutation would be unambiguously a deletion (although of ambiguous size). In cases where multiple parental alleles would result in the minimum change in allele length but in opposite directions, maternal inheritance is algorithmically favored over paternal and mutations arising from shorter parental alleles over longer ones. For example, if the trio's genotypes were instead [12,12], [20, 22], and [12, 21], the mutation would be called as an expansion from the paternal "20" allele to the proband's "21" allele. On the other hand, given genotypes [9,12], [9,10], [9,11], the proband's "11" allele would be called as a deletion mutation from the maternal "12" allele. Although this parsimony strategy no doubt resulted in some ancestral state misidentification, we note that the observed association of maternal age with the maternal deletion rate and paternal age association with the deletion and expansion rates were both robust

to filtering away DNMs that could have arisen through either an expansion or deletion from a parental allele.

The maternal age effect appears to be driven primarily by deletion DNMs affecting non-homopolymer STR loci. Maternal age dependence also appears weaker when we restrict to DNM categories that are likely enriched for false positives. One such category is homopolymer STRs, which are more prone to genotyping errors than other STR types though make up the majority of mutations in our dataset [95]. The homopolymer deletion mutation rate has no detectable maternal age effect in this dataset (P = 0.412, Poisson regression), and the maternal age effect is significantly stronger for non-homopolymers ($P = 9.453*10^{-3}$; effect size of 0.0161; ANCOVA). The latter is also true for the paternal age effect on deletions in non-homopolymers ($P < 2*10^{-16}$, effect size of 0.08, ANCOVA). Another STR category that is likely enriched for false positives are mutations that change the allele by more than one repeat unit, and indeed we observe that the maternal age effect was strongest on mutations that changed an STR length by a single repeat unit. Although mutations that result in a larger difference in allele length exist, many are likely to be false positives caused by allelic dropout in a heterozygous parent given the low prior likelihood of such mutations. We found that the maternal age effect remained significant when running the same Poisson linear regression, including only mutations of +/- 1 repeat unit in size, then further filtering for only those mutations of unambiguous directionality and size (P = 1.13*10-7, 0.00591, respectively).

To further characterize the mutational signature associated with maternal age at STR loci, we examined how maternal age affected the distribution of replication timing at mutant STRs. Although mutations that accumulate with maternal age are not likely to originate as replication errors, their profile might still appear to depend on replication timing given that replication

timing covaries with chromatin state and the types and frequencies of certain molecular interactions. Using replication timing data from embryonic stem cells, we calculated the replication timing for each STR DNM [113]. Maternal age was significantly associated with a later mean replication timing of deletions but not of expansions ($P = 7.34*10^{-4}$, 0.153, respectively; coefficients of a generalized linear model) (Supplementary figure 3). This observation contrasts with an earlier report that the replication timing of single nucleotide DNMs in older fathers skews towards earlier replicating regions than DNMs from younger fathers [114].

Certain genomic regions in humans have been classified as maternal age hotspots because they have exceptionally elevated mutation rates in the children of older mothers with a particular enrichment of C>G mutations [33]. These regions appear to be hotspots for double-stranded breaks in older oocytes, leading to damage-associated mutagenesis. We ran a Poisson regression with an identity link function to measure the maternal age effect within these hotspot regions but found no significant maternal age effect ($P = 0.5$), likely due to the relatively small number of STRs occurring in these regions [33].

3.2.2 *Higher STR mutation rate observed in children of African ancestry after controlling for high allelic dropout in parents*

Motivated by previous studies that observed variation of SNP mutation spectra among human populations, we tested for population variation in the STR mutation rate [16–18]. Although the individuals in this cohort self-reported race, we based our analyses upon maximum likelihood ancestries inferred using a reference panel of diverse genomes [115]. For each individual in the SSC, we projected SNPs onto $k = 8$ components generated by running ADMIXTURE on a combined reference panel of 1000 Genomes Project and Simons Genome Diversity Project individuals [115–117]. We annotated the components based on the majority

continental ancestry of reference individuals. Individuals in the SSC were assigned ancestry based on their highest ancestry component. We found that the SSC comprises a majority of individuals with predominantly European ancestry, but also includes a number of individuals from other continental ancestry groups (Supplementary Figure 4). Inferred ancestries showed high concordance with self-reported race (Supplementary Table 1).

We observe nearly 20 more STR DNMs per child in children of predominantly African ancestry compared to children of other ancestry groups (Fig. 3.3). Although African ancestry is associated with slightly lower STR DNM call quality, this effect is not strong enough to explain the observed difference in mutation rates. A Poisson generalized linear model regressing the number of STR DNMs per child against inferred ancestry and parental age at birth while accounting for covariance in mean DNM call quality finds that ancestry is still significantly predictive (coefficients are significant to $P < 2 * 10^{-16}$ for all ancestries). We observe that African ancestry is still significantly associated with the number of mutations after we restrict the analyses to higher quality mutations (Supplementary figure 5).

Environmental factors such as diet and pollution have been hypothesized to affect the germline mutation rate [53,118]. We cannot rule out the possibility that a higher mutation rate in families with African ancestry is driven by environmental differences from other families that affect mutagenesis independently of genetic background. We hypothesized that such an environmental divide, if it exists, might correlate with socioeconomic status, which is summarized as a total household income in the SSC metadata. However, we found no association of SES with mutation rate, either independently or when also accounting for parental age and genetic ancestry (Supplementary figure 6) ($P = 0.1414$, Pearson's correlation test; $P > 0.804$, Poisson regression).

The excess DNMs observed in children of African ancestry are detectable across STR loci and across many different mutational categories. Although the ratio of the African to the non-African mutation rate is highest for STR expansions, the rate of deletions also appears to be significantly higher in children of African ancestry (Supplementary figure 7) ($P = 1.04*10^{-36}$, $1.33*10^{-14}$, respectively; Wilcoxon rank sum test). An elevated African mutation rate is also observed across different STR repeat unit lengths, but the effect is strongest for homopolymers (Supplementary figure 8) ($P = 3.87 * 10^{-38}$, $1.88*10^{-7}$, $1.35*10^{-2}$, 0.051, and 1 for repeat units of 1, 2, 3, 4, and ≥ 5, respectively; Wilcoxon Rank Sum test with Bonferroni correction).

Since mutations are the source of genetic diversity, we expect that loci of higher diversity may on average have higher locus-specific mutation rates than lower diversity loci. Indeed, we observe a higher mutation rate per locus where both parents are heterozygous compared to sites where both parents are homozygous (Fig 3.3B). The mutation rate at sites with two heterozygous parents is not significantly different among ancestries ($P = 0.23$, Wilcoxon rank sum test). However, we observe that loci where one parent is homozygous and the other is heterozygous have a higher mutation rate than loci where both or neither parent are heterozygous (Fig 3B). At these loci, we find that the DNM is more likely to originate on the chromosome inherited from the homozygous parent rather than the heterozygous parent; this imbalance is greatest in children of African ancestry (Fig 3C). In order to test whether this observation could be a result of bioinformatic error, we stratified mutations into two quality bins: those whose quality scores fell below or above the median quality score. Mutations of higher quality display a higher likelihood of arising in the heterozygous parent's chromosome than mutations of lower quality, but a slight imbalance remains for children of African ancestry (Fig 3.3D). Interestingly, mutations

exclusively with a posterior probability greater than 0.9999 – ostensibly the highest quality set of mutations – still phase more frequently to the heterozygous parent (Supplementary figure 9).

We hypothesized that some or all of the excess *de novo* mutations in Africans might be the result of allelic dropout events, which can lead to false positive DNMs if an allele inherited by a child from a parent heterozygous at that locus is not successfully genotyped in the parent. This causes the parent to be falsely genotyped as homozygous such that the allele appears to have arisen *de novo* in the child. This could explain our observation of an elevated rate of mutations transmitted from the homozygous parent at loci where the other parent is heterozygous, as we observe in Fig 3B.

We considered trying to minimize allelic dropout by excluding DNMs that violate the infinite sites model by generating an allele already segregating in another SSC family. However, in practice only about 1% of the 175K STR DNMs obey the infinite sites model, making such a filter impractical. However, filtering out mutations that appear to change an allele's length by more than one repeat unit are enriched for allelic dropout artifacts because such mutations are less likely *a priori*. We measured this by looking at DNMs in children with one heterozygous parent and one homozygous parent and measuring their deviation from the expectation that 50% on average should phase to the homozygous parent. In practice, we found that filtering out DNMs that changed the ancestral allele by more than one repeat unit reduced the proportion of mutations mapping to the homozygous parent from an odds ratio of 0.76 to an odds ratio of 0.67 (Supplementary figure 10). Applying this filter in addition to requiring a posterior greater than the 50^{th} percentile (0.91) still resulted in significantly more DNMs associated with African ancestry, though it also decreased the mean number of DNMs by nearly an order of magnitude (mean of 55 to 10 DNMs per child) ($P = 8.00 * 10^{-5}$) (Supplementary figure 11).

These observations indicate that parental allelic dropout may be at least in part responsible for the seemingly higher rate of DNMs we observe in children with African ancestry. However, the maternal age effect is unlikely to be affected by allelic dropout, as we neither expect nor observe (P=0.74, 0.58 for mothers and fathers, respectively) any significant correlation between heterozygosity and parental age at conception. We still observe a significant maternal age effect on the number of DNMs after applying the aforementioned filters on mutation quality and size (P = 0.00862, Poisson regression).

Children diagnosed with ASD were found to have a higher STR mutation rate than sibling controls in the original analyses of the SSC STRs, and we found the mutation rate difference between autism-affected and unaffected siblings to be similar across ancestries [95] (Supplementary figure 12). We tested for ancestry-specific effects in children with or without an ASD diagnosis by regressing the number of DNMs against the child's ancestry, ASD diagnosis, and an interaction variable between ASD diagnosis and ancestry (Poisson regression with identity link function). While ancestry and ASD diagnosis were significantly independently predictive of DNM rate, no interactions were significant (P > 0.28 for all ancestries).

3.3 Discussion

Our analyses of the parental age effects on STR mutagenesis support the hypothesis that STR mutagenesis is more replication dependent than SNP mutagenesis, but do not support the classical view that replication slippage is the sole cause of STR mutations. The fraction of phased mutations deriving from the paternal lineage increases with paternal age, supporting a hypothesis that the ratio of paternal to maternal mutation rates is lower pre-puberty than post-puberty. This is ostensibly a result of the continued replication of the paternal germline after puberty. Nevertheless, the significant maternal age effect cannot be a result of polymerase

slippage during S-phase replication, as the maternal germline does not replicate after birth [28]. The stronger maternal age association with deletion DNMs at non-homopolymer repeats may provide a clue as to the source of these mutations, though our current understanding of STR mutational signatures is not advanced enough to decode it. Certain mutation signatures deconvoluted from somatic mutation spectra in tumor cells, COSMIC ID4 and ID12, both primarily comprise deletions at short repeats other than homopolymers. Although ID12 is of unknown etiology, TOP1 transcription-associated mutagenesis was recently implicated as the cause of ID4, particularly at TNT motifs [11,49]. Future work may examine how maternal age affects rates at certain nucleotide motifs.

We found African ancestry to be associated with a significantly elevated mutation rate amongst SSC families, although this signal may be at least partially an artifact of allelic dropout. Classically, allelic dropout was thought to most strongly affect long STR alleles and homopolymers; we observed, however, that the strength of the effect of ancestry on mutation rate did not covary with allele length and was not limited to homopolymers [95]. This implies that if the ancestry effect on mutation rate is entirely an artifact of allelic dropout, this dropout is more pervasive than the pattern described based on older PCR studies and is not predicted by the same locus attributes. Notably, African ancestry is not associated with an elevated single nucleotide *de novo* mutation rate in the SSC (Supplementary figure 13) ($P = 0.6319$, Wilcoxon rank sum test), though this does not rule out an effect of ancestry on the mutation rate at STRs. Validations with additional sequencing technologies and possibly additional cohorts will likely be needed to determine how many of the mutation calls in this STR dataset are false positives and whether the observed effect of African ancestry is biological or artifactual.

Methods that use models of STR mutagenesis can help estimate the burden of this mutation type on ASD and other developmental disorders [95]. Our results indicate that preliminary studies of STR mutations in the SSC have likely underestimated the effects of allelic dropout, and they also imply that ancestry may affect the DNM burden at STR loci. Higher-fidelity technologies to detect STR DNMs will be needed to better infer the deleterious contribution of these mutation types.

3.4 FIGURES

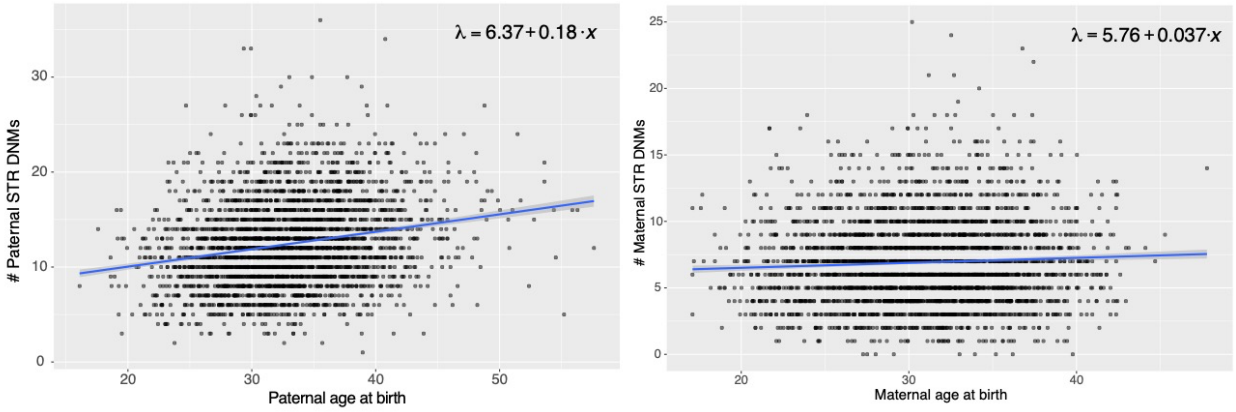


Figure 3.1. Paternal and maternal age are both associated with higher STR *de novo* mutation rates

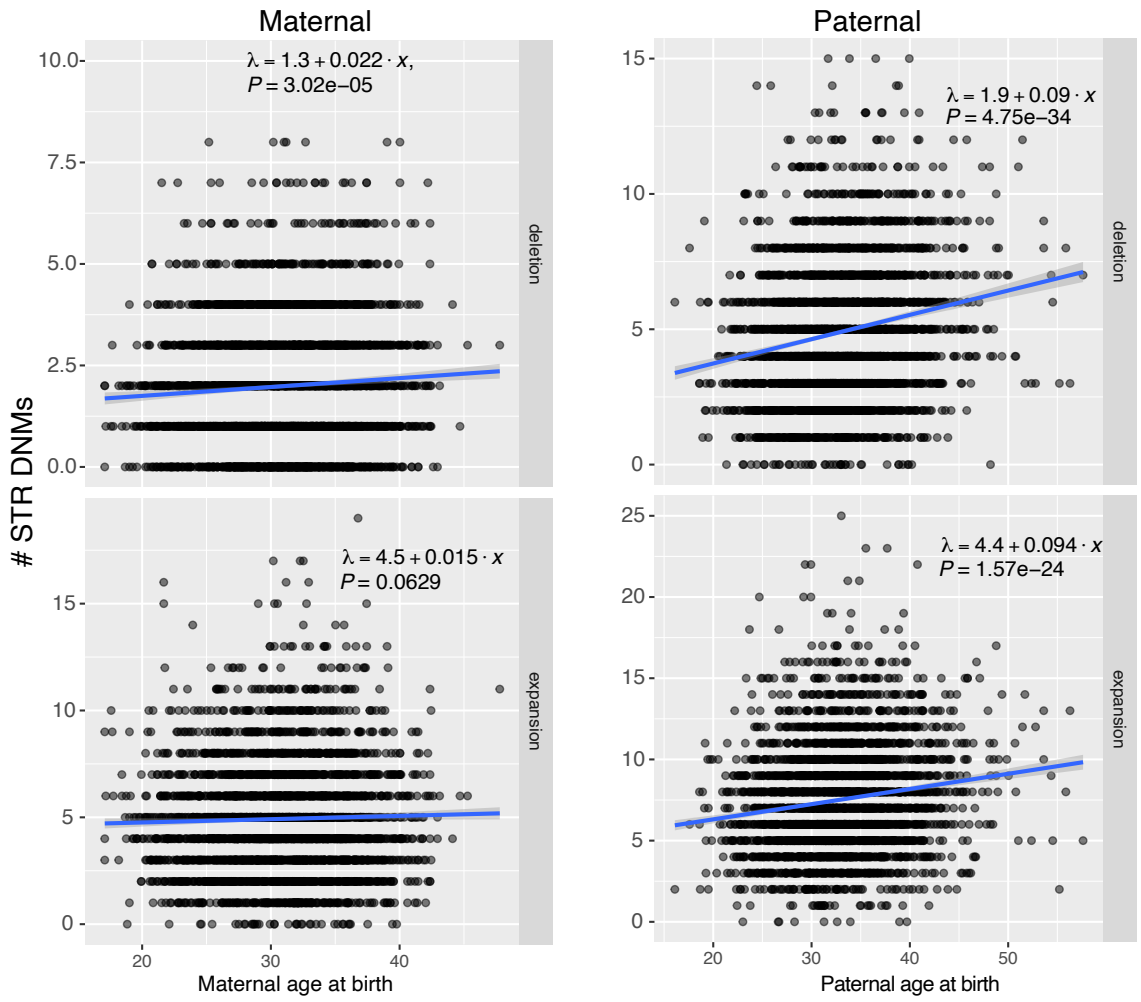


Figure 3.2. Maternal and paternal age are significantly associated with STR mutation rates, particularly the STR deletion rate

Poisson regressions with identity link functions modeled the number of maternally and paternally derived deletion and expansion DNMs in a child as a function of the maternal and paternal age at birth, respectively. P-values are shown for the estimated slope coefficients.

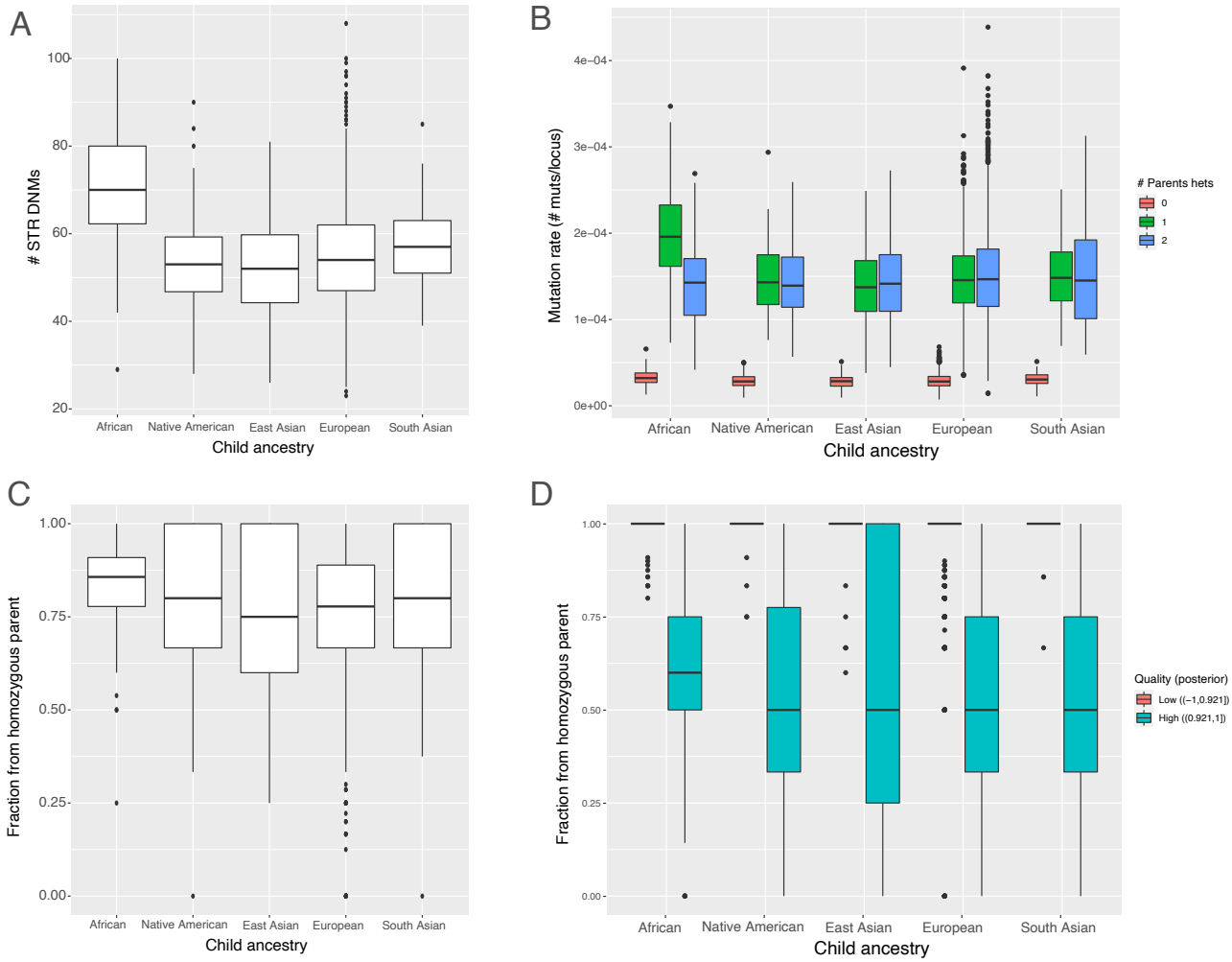


Figure 3.3. Variation of STR mutation rates with genetic ancestry

A: The number of STR DNMs in children in the SSC, grouped by inferred genetic ancestry. Children of predominantly African genetic ancestry harbor significantly more mutations than children of other ancestries (Wilcoxon rank-sum test, $P = 1.04 * 10^{-36}$). B: Mutation rate significantly differs as a function of ancestry only at the set of sites where at least one parent is homozygous (Wilcoxon rank-sum tests, $P = 0.23$, $1.13 * 10^{-26}$, $2.87 * 10^{-7}$ for 0, 1, 2 parents homozygous, respectively). C: At sites where one parent is homozygous and the other is heterozygous, mutations more frequently phase to chromosome inherited from the homozygous parent. We filtered for mutations where parents shared no alleles, thus giving a hypothetically

equal probability of phasing to either lineage. Each observation represents the fraction of these STR DNMs in one SSC child phased to the homozygous parent. D: Posterior probability, as a proxy for call confidence and quality, covaries with the balance in inheritance. Of the mutations examined in panel C, higher quality mutations phase more evenly to either parent.

Chapter 4. CONCLUSIONS

The research I have presented in chapters 2 and 3 contribute to a larger field that endeavors to describe the cause and consequence of evolution of germline mutagenesis. Results from my work help quantify the extent of variation in germline mutagenesis and point to possible mutators and their mechanisms. I hope that these findings may help inform future work to further study the questions I have asked over the past four years.

4.1 GENETIC CAUSES AND CONSEQUENCES OF THE RAPID EVOLUTION OF GERMLINE MUTAGENESIS

In some ways, I view the trajectory of my subfield of mutation rate evolution as parallel to research on evolution of recombination rate, particularly in mammals and likely throughout most vertebrates. Much of meiotic recombination occurs at recombination hotspots, but the location and usage of hotspots differ drastically between species; variation is observed even between strains of mice or human populations [41–43,119]. Like mutation rate, recombination rate is a rapidly evolving genomic trait that both influences and is influenced by genetic variation. There exist experimental systems and statistical tools to test hypotheses in both fields [42,43,120]. However, unlike mutation rate evolution, much of the variation in recombination hotspot usage is linked to a single *trans*-acting locus: evolution the gene *PRDM9* in humans and its ortholog in other lineages [40–43]. The protein encoded by *PRDM9* harbors a rapidly evolving zinc finger domain, whose structure determines DNA binding motif preferences; different alleles of the gene confer different hotspot usage. Binding by *PRDM9* rapidly erodes hotspots, which in turn alters the distribution of fitness effects for different *PRDM9* alleles and encourages selection for a different motif.

On the other hand, much of mutation rate variation between species and within the genome remains cryptic, and few efforts have identified single genetic loci responsible for this variation [70,121]. The research documented in chapter 2 aimed originally to identify possible large-effect *cis*- or *trans*-acting mutation spectrum modifiers segregating along different great ape lineages [24]. Genome-wide spectrum divergence in polymorphisms had already been established since the most recent common ancestor of great apes, resulting in mutation signatures unique to lineages. Compartments, genomic regions syntenic across lineages that share a unique feature such as chromatin state or function, helped us inquire whether mutation signatures linked to specific features had diverged between lineages and thus contributed to the rapid evolution of these species signatures. However, we found that compartment signatures and their contributions to mutation spectra were largely constant across all lineages and genomic regions, supporting a hypothesis that these signatures evolved before the most recent common ancestor of great apes. Our results showed that chromatin state and genomic function are certainly mutation spectrum modifiers, but we found no evidence that they vary in lineage-specific manners among great apes. This research continues to push the field of mutation spectrum evolution away from the simpler story of recombination rate evolution; though the trait evolves rapidly, its genetic contribution is more likely due to myriad small-effect size mutators and cannot be easily pinned down to a small set of easily identifiable loci. Furthermore, we observed a tight correlation between genetic distance and mutation spectrum distance in great ape individuals. This phenomenon seems most likely to result from the accumulation and fixation of many small mutator alleles along each lineage. Life history traits that covary with genetics distance, such as generation time, likely contribute to mutation spectrum variation, but our methods were not powered to disentangle these from genetic causes [2].

A surprising benefit of our results could be derived from the striking conservation of these compartment-specific mutation signatures. Variation in mutation rate between lineages confounds evolutionary models that require mutation rate as a parameter and frequently treat it as a constant [1,19]. This discrepancy can lead to uncertainty in molecular clock-based estimates of evolutionary history such as divergence times between lineages [1,106]. However, more reliable molecular clocks can be parametrized on mutation types whose rates are more stable over time [9,19,21]. Further extending these models to include mutations from conserved mutation signatures could be an extension of our results.

4.2 NEXT GENERATION SEQUENCING AND BEYOND

Much of the work presented in chapters 2 and 3 leverages short read next generation sequencing (NGS), a relatively cheap and highly scalable technology. However, inference from short read sequencing has limitations as evidenced by constraints to the analyses and findings in chapters 2 and 3. In chapter 2, our work leveraged short read whole-genome NGS from great ape individuals mapped to the human reference genome [74]. The single reference genome facilitated easily executing analyses on syntenic regions, but barred us from analyzing newer non-syntenic regions in non-human lineages. Furthermore, mutations in low complexity regions, such as recent segmental duplications, are challenging to identify from short read sequencing due to an inability to uniquely map reads. Mutagenesis of these regions has been shown in humans to be markedly different than mutagenesis of higher complexity regions [110,122]. Future work could leverage new high quality long read species-specific reference genomes and annotations by either simply remapping the original short reads or generating new long read genomes of individuals.

Short read sequencing poses clear challenges to calling genotypes and DNMs at STRs, data on which analyses in chapter 3 are based. As mentioned above, repeats are difficult to sequence with short read sequencing due to homology with other genomic regions. Furthermore, long STR alleles may be more likely to experience allelic dropout due to lower coverage [95]. New long read sequencing technologies such as PacBio continuous long read and high-fidelity (HiFi) or Oxford Nanopore Technologies (ONT) can help resolve low complexity regions of the genome, and have recently led to a marked increase in the number of DNMs found in a human pedigree [122]. Furthermore, applying multiple sequencing technologies to the same family better distinguished true DNMs from inherited variants that dropped out while genotyping the parent of origin. At a small scale, analyzing differences in STR DNM calling between different technologies in the same family may help better parametrize callers that use short read sequencing and design more precise filters to account for allelic dropout. More broadly, though costly, large numbers of pedigrees could be sequenced with long read technologies to better estimate STR DNM rates and their variation as a function of parental age and genetic background.

4.3 IDENTIFYING THE MECHANISM OF MUTATORS

One goal of research in germline mutagenesis is to ultimately identify the specific mechanism of mutators. Although many environmental and genetic variables have been found to correlate with particular mutation signatures, the exact molecular mechanisms largely remain unknown [11]. For example, the exact molecular mechanisms of the well-documented parental age or replication timing effects on mutation rate are still unclear [25,29,123]. Much of the work in chapters 2 and 3 identifies genomic or environmental correlates of mutation rate and spectrum, and these analyses certainly help narrow down possible mutational pathways responsible for this

variance. However, the exact mechanisms of these mutators still remains largely unclear. In chapter 2, we reported an enrichment for CpG>GpG mutations at hydroxymethylated sites in ERVs in humans and other great apes. Functional follow-up could include studying mutation spectra in syntenic regions in species without hydroxymethylation, or an experimental study using a model in which hydroxymethylation can be induced. Similarly, we found that both paternal and maternal age affected the mutation rate of non-homopolymer STRs significantly more than homopolymer STRs. Although homopolymers are notoriously challenging to sequence, and the signal of parental age effects could be purely obscured by the resulting noise, several possible signatures have been identified in both cancer and germline mutations that may help indicate molecular mechanism [11,49]. COSMIC ID signature 4 describes an enrichment for deletions at non-homopolymers and was recently associated with *TOP1* activity [49]. Measuring variance in *TOP1* activity in germline or somatic reproductive tissue (e.g. testis tissue) as a function of age may help further elucidate the exact mechanism of this parental age effect.

BIBLIOGRAPHY

1. Goodman M. Rates of molecular evolution: The hominoid slowdown. Bioessays. 1985;3:9–14.

2. Besenbacher S, Hvilsom C, Marques-Bonet T, Mailund T, Schierup MH. Direct estimation of mutations in great apes reconciles phylogenetic dating. Nat Ecol Evol. 2019;3:286–92.

3. Sturtevant AH. Essays on Evolution. I. On the Effects of Selection on Mutation Rate. The Quarterly Review of Biology. 1937;12:464–7.

4. Kimura M. On the evolutionary adjustment of spontaneous mutation rates. Genet Res. 1967;9:23–34.

5. Margoliash E. Primary structure and evolution of cytochrom c. Proc Natl Acad Sci USA. 1963;50:672–9.

6. Zuckerkandl E, Pauling LB. Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B, editors. Horizons in Biochemistry. New York: Academic Press; 1962. p. 189–225.

7. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge University Press; 1983.

8. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012;13:745–53.

9. Moorjani P, Amorim CEG, Arndt PF, Przeworski M. Variation in the molecular clock of primates. Proceedings of the National Academy of Sciences. 2016;113:10607–12.

10. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.

11. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.

12. Shinbrot E, Henninger EE, Weinhold N, Covington KR, Göksenin AY, Schultz N, et al. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. Genome Res. 2014;24:1740–50.

13. Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, et al. Deconvolving the Recognition of DNA Shape from Sequence. Cell. 2015;161:307–18.

14. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein–DNA recognition. Nature. 2009;461:1248–53.

15. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. Proceedings of the National Academy of Sciences. 2013;110:6376–81.

16. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nature Genetics. 2016;48:349–55.

17. Harris K. Evidence for recent, population-specific evolution of the human mutation rate. Proc Natl Acad Sci USA. 2015;112:3439–44.

18. Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. McVean G, editor. eLife. 2017;6:e24284.

19. Hwang DG, Green P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proceedings of the National Academy of Sciences. 2004;101:13994–4001.

20. Carlson J, DeWitt WS, Harris K. Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation. Current Opinion in Genetics & Development. 2020;62:50–7.

21. Moorjani P, Gao Z, Przeworski M. Human Germline Mutation and the Erratic Evolutionary Clock. PLOS Biology. 2016;14:e2000744.

22. Bergeron LA, Besenbacher S, Turner T, Versoza CJ, Wang RJ, Price AL, et al. The Mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. eLife. 2022;11:e73577.

23. Terekhanova NV, Seplyarskiy VB, Soldatov RA, Bazykin GA. Evolution of local mutation rate and its determinants. Mol Biol Evol. 2017;msx060.

24. Goldberg ME, Harris K. Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny. Corbett-Detig R, editor. Genome Biology and Evolution. 2022;14:evab104.

25. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. Human mutation rate associated with DNA replication timing. Nat Genet. 2009;41:393–5.

26. Agarwal I, Przeworski M. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. Proc Natl Acad Sci USA. 2019;116:17916–24.

27. Poulos RC, Olivier J, Wong JWH. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. Nucleic Acids Research. 2017;45:7786–95.

28. Drost JB, Lee WR. Biological basis of germline mutation: Comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. Environmental and Molecular Mutagenesis. 1995;25:48–64.

29. Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. Proc Natl Acad Sci USA. 2019;116:9491–500.

30. Wu FL, Strand AI, Cox LA, Ober C, Wall JD, Moorjani P, et al. A comparison of humans and baboons suggests germline mutation rates do not track cell divisions. Barton NH, editor. PLoS Biol. 2020;18:e3000838.

31. Lindsay SJ, Rahbari R, Kaplanis J, Keane T, Hurles ME. Similarities and differences in patterns of germline mutation between mice and humans. Nat Commun. 2019;10:4053.

32. de Manuel M, Wu FL, Przeworski M. A paternal bias in germline mutation is widespread across amniotes and can arise independently of cell divisions [Internet]. Evolutionary Biology; 2022 Feb. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.02.07.479417

33. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017;549:519–22.

34. Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, et al. New observations on maternal age effect on germline de novo mutations. Nat Commun. 2016;7:1–10.

35. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics. 2001;159:907–11.

36. Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. Quantification of GC-biased gene conversion in the human genome. Genome Res. 2015;25:1215–28.

37. Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. eLife. 2018;7:e36317.

38. McVicker G, Gordon D, Davis C, Green P. Widespread Genomic Signatures of Natural Selection in Hominid Evolution. Nachman MW, editor. PLoS Genet. 2009;5:e1000471.

39. Milligan WR, Amster G, Sella G. The impact of genetic modifiers on variation in germline mutation rates within and among human populations [Internet]. Evolutionary Biology; 2021 Aug. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.08.25.457718

40. Cavassim MIA, Baker Z, Hoge C, Schierup MH, Schumer M, Przeworski M. *PRDM9* losses in vertebrates are coupled to those of paralogs *ZCWPW1* and *ZCWPW2*. Proc Natl Acad Sci USA. 2022;119:e2114401119.

41. Parvanov ED, Petkov PM, Paigen K. *Prdm9* Controls Activation of Mammalian Recombination Hotspots. Science. 2010;327:835–835.

42. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science. 2010;327:836–40.

43. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. Science. 2010;327:876–9.

44. Kaplanis J, Ide B, Sanghvi R, Neville M, Danecek P, Coorens T, et al. Genetic and chemotherapeutic influences on germline hypermutation. Nature [Internet]. 2022 [cited 2022 May 15]; Available from: https://www.nature.com/articles/s41586-022-04712-2

45. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488:471–5.

46. Goldmann JM, Hampstead JE, Wong WSW, Wilfert AB, Turner TN, Jonker MA, et al. Differences in the number of de novo mutations between individuals are due to small family-specific effects and stochasticity. Genome Res. 2021;31:1513–8.

47. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, et al. Timing, rates and spectra of human germline mutation. Nat Genet. 2016;48:126–33.

48. Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, et al. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. eLife. 2019;8:e46922.

49. Reijns MAM, Parry DA, Williams TC, Nadeu F, Hindshaw RL, Rios Szwed DO, et al. Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. Nature. 2022;602:623–31.

50. Sasani TA, Ashbrook DG, Beichman AC, Lu L, Palmer AA, Williams RW, et al. A natural mutator allele shapes mutation spectrum variation in mice. Nature [Internet]. 2022 [cited 2022 May 13]; Available from: https://www.nature.com/articles/s41586-022-04701-5

51. Jiang P, Ollodart AR, Sudhesh V, Herr AJ, Dunham MJ, Harris K. A modified fluctuation assay reveals a natural mutator phenotype that drives mutation spectrum variation within Saccharomyces cerevisiae. eLife. 2021;10:e68285.

52. Maksimov M, Ashbrook DG, BXD Sequencing Consortium, Villani F, Colonna V, Mousavi N, et al. A novel quantitative trait locus implicates *Msh3* in the propensity for genome-wide short tandem repeat expansions in mice [Internet]. Genomics; 2022 Mar. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.03.02.482700

53. Kessler MD, Loesch DP, Perry JA, Heard-Costa NL, Taliun D, Cade BE, et al. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. Proc Natl Acad Sci USA. 2020;117:2560–9.

54. Sima J, Gilbert DM. Complex correlations: replication timing and mutational landscapes during cancer and genome evolution. Current Opinion in Genetics & Development. 2014;25:93–100.

55. Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, et al. Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. The American Journal of Human Genetics. 2012;91:1033–40.

56. Liu L, De S, Michor F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nature Communications [Internet]. 2013 [cited 2019 Sep 3];4. Available from: http://www.nature.com/articles/ncomms2502

57. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518:360–4.

58. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488:504–7.

59. Hellmann I, Prüfer K, Ji H, Zody MC, Pääbo S, Ptak SE. Why do human diversity levels vary at a megabase scale? Genome Research. 2005;15:1222–31.

60. Martincorena I, Seshasayee ASN, Luscombe NM. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature. 2012;485:95–8.

61. Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. A Neutral Explanation for the Correlation of Diversity with Recombination Rates in Humans. The American Journal of Human Genetics. 2003;72:1527–35.

62. Keightley PD, Eöry L, Halligan DL, Kirkpatrick M. Inference of Mutation Parameters and Selective Constraint in Mammalian Coding Sequences by Approximate Bayesian Computation. Genetics. 2011;187:1153–61.

63. Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MAF. Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. Proceedings of the National Academy of Sciences. 2008;105:10051–6.

64. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, et al. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. Cell. 2012;151:1431–42.

65. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet. 2012;13:565–75.

66. Ananda G, Chiaromonte F, Makova KD. A genome-wide view of mutation rate co-variation using multivariate analyses. Genome Biol. 2011;12:R27.

67. Li C, Luscombe NM. Nucleosome positioning stability is a significant modulator of germline mutation rate variation across the human genome [Internet]. Genomics; 2018 Dec. Available from: http://biorxiv.org/lookup/doi/10.1101/494914

68. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nature Reviews Genetics. 2011;12:756–66.

69. Johnson PLF, Hellmann I. Mutation Rate Distribution Inferred from Coincident SNPs and Coincident Substitutions. Genome Biology and Evolution. 2011;3:842–50.

70. Tyekucheva S, Makova KD, Karro JE, Hardison RC, Miller W, Chiaromonte F. Human-macaque comparisons illuminate variation in neutral substitution rates. Genome Biol. 2008;9:R76.

71. Li W-H, Tanimura M. The molecular clock runs more slowly in man than in apes and monkeys. Nature. 1987;326:93–6.

72. Speidel L, Forest M, Shi S, Myers S. A method for genome-wide genealogy estimation for thousands of samples [Internet]. Genetics; 2019 Feb. Available from: http://biorxiv.org/lookup/doi/10.1101/550558

73. DeWitt WS, Harris KD, Harris K. Joint nonparametric coalescent inference of mutation spectrum history and demography [Internet]. Evolutionary Biology; 2020 Jun. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.06.16.153452

74. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013;499:471–5.

75. Thomas GWC, Wang RJ, Puri A, Harris RA, Raveendran M, Hughes DST, et al. Reproductive Longevity Predicts Mutation Rates in Primates. Current Biology. 2018;28:3193-3197.e5.

76. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, et al. Evolution and diversity of copy number variation in the great ape lineage. Genome Research. 2013;23:1373–82.

77. Lynch M. The Cellular, Developmental and Population-Genetic Determinants of Mutation-Rate Evolution. Genetics. 2008;180:933–43.

78. Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 2016;17:704–14.

79. Carlson J, Li JZ, Zöllner S. Helmsman: fast and efficient mutation signature analysis for massive sequencing datasets. BMC Genomics. 2018;19:845.

80. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.

81. Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova N, Abeysinghe SS, et al. Breakpoints of gross deletions coincide with non-B DNA conformations. Proceedings of the National Academy of Sciences. 2004;101:14162–7.

82. Guiblet WM, Cremona MA, Cechova M, Harris RS, Kejnovská I, Kejnovsky E, et al. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res. 2018;28:1767–78.

83. Harris RS, Petersen-Mahrt SK, Neuberger MS. RNA Editing Enzyme APOBEC1 and Some of Its Homologs Can Act as DNA Mutators. Molecular Cell. 2002;10:1247–53.

84. Gardiner-Garden M, Frommer M. CpG Islands in vertebrate genomes. Journal of Molecular Biology. 1987;196:261–82.

85. Larsen F, Gundersen G, Lopez R, Prydz H. CpG islands as gene markers in the human genome. Genomics. 1992;13:1095–107.

86. Supek F, Lehner B, Hajkova P, Warnecke T. Hydroxymethylated Cytosines Are Associated with Elevated C to G Transversion Rates. Duret L, editor. PLoS Genetics. 2014;10:e1004585.

87. Yu M, Hon GC, Szulwach KE, Song C-X, Zhang L, Kim A, et al. Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. Cell. 2012;149:1368–80.

88. Pintacuda G, Wei G, Roustan C, Kirmizitas BA, Solcan N, Cerase A, et al. hnRNPK Recruits PCGF3/5-PRC1 to the Xist RNA B-Repeat to Establish Polycomb-Mediated Chromosomal Silencing. Molecular Cell. 2017;68:955-969.e10.

89. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.

90. Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Ž, et al. Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Sci Data. 2016;3:160075.

91. Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, The BRIDGES Consortium, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nature Communications [Internet]. 2018 [cited 2019 Sep 3];9. Available from: http://www.nature.com/articles/s41467-018-05936-5

92. Eyre-Walker A. Recombination and mammalian genome evolution. Proceedings of the. 1993;7.

93. Eyre-Walker A. Evidence of Selection on Silent Site Base Composition in Mammals: Potential Implications for the Evolution of Isochores and Junk DNA. :10.

94. Holmquist GP. Chromatin Flavors, and Their Functional Features. American Journal of Human Genetics. 1992;51:17–37.

95. Mitra I, Huang B, Mousavi N, Ma N, Lamkin M, Yanicky R, et al. Patterns of de novo tandem repeat mutations and their role in autism. Nature. 2021;589:246–50.

96. Spencer CC, Neigel JE, Leberg PL. Experimental evaluation of the usefulness of microsatellite DNA for detecting demographic bottlenecks. Molecular Ecology. 2000;9:1517–28.

97. Weber JL, Wong C. Mutation of human short tandem repeats. Hum Mol Genet. 1993;2:1123–8.

98. Sherman SL, Jacobs PA, Morton NE, Froster-Iskenius U, Howard-Peebles PN, Nielsen KB, et al. Further segregation analysis of the fragile X syndrome with special reference to transmitting males. Hum Genet. 1985;69:289–99.

99. Gatchel JR, Zoghbi HY. Diseases of Unstable Repeat Expansion: Mechanisms and Common Principles. Nat Rev Genet. 2005;6:743–55.

100. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53.

101. Forster P, Hohoff C, Dunkelmann B, Schürenkamp M, Pfeiffer H, Neuhuber F, et al. Elevated germline mutation rate in teenage fathers. Proc R Soc B. 2015;282:20142898.

102. Klintschar M, Dauber E-M, Ricci U, Cerri N, Immel U-D, Kleiber M, et al. Haplotype studies support slippage as the mechanism of germline mutations in short tandem repeats. Electrophoresis. 2004;25:3344–8.

103. Amos W, Kosanović D, Eriksson A. Inter-allelic interactions play a major role in microsatellite evolution. Proc R Soc B. 2015;282:20152125.

104. Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. Nat Genet. 2012;44:1161–5.

105. Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. The American Journal of Human Genetics. 1998;62:1408–15.

106. Chintalapati M, Moorjani P. Evolution of the mutation rate across primates. Current Opinion in Genetics & Development. 2020;62:58–64.

107. Navidi W, Arnheim N. Using PCR in preimplantation genetic disease diagnosis. Human Reproduction. 1991;6:836–49.

108. Walsh PS, Erlich HA, Higuchi R. Preferential PCR amplification of alleles: mechanisms and solutions. Genome Res. 1992;1:241–50.

109. Gagneux P, Boesch C, Woodruff DS. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Mol Ecol. 1997;6:861–8.

110. Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell. 2017;171:710-722.e12.

111. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol. 2018;19:121.

112. Fischbach GD, Lord C. The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron. 2010;68:192–5.

113. Ding Q, Edwards MM, Wang N, Zhu X, Bracci AN, Hulke ML, et al. The genetic architecture of DNA replication timing in human pluripotent stem cells. Nat Commun. 2021;12:6746.

114. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of de novo mutations in humans. Nat Genet. 2015;47:822–6.

115. Wilfert AB, Turner TN, Murali SC, Hsieh P, Sulovari A, Wang T, et al. Recent ultra-rare inherited variants implicate new autism candidate risk genes. Nat Genet. 2021;53:1125–34.

116. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538:201–6.

117. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.

118. Mathieson I, Reich D. Differences in the rare variant spectrum among human populations. Girirajan S, editor. PLoS Genet. 2017;13:e1006581.

119. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467:1099–103.

120. Ségurel L, Leffler EM, Przeworski M. The Case of the Fickle Fingers: How the PRDM9 Zinc Finger Protein Specifies Meiotic Recombination Hotspots in Humans. PLoS Biol. 2011;9:e1001211.

121. Hodgkinson A, Ladoukakis E, Eyre-Walker A. Cryptic Variation in the Human Mutation Rate. Barton NH, editor. PLoS Biol. 2009;7:e1000027.

122. Noyes MD, Harvey WT, Porubsky D, Sulovari A, Li R, Rose NR, et al. Familial long-read sequencing increases yield of de novo mutations. The American Journal of Human Genetics. 2022;109:631–46.

123. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015;521:81–4.

124. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. Nature Genetics. 2002;31:241–7.

APPENDIX A

SUPPLEMENT FOR CHAPTER 2

Genus	Species	Subspecies	Common name	Population	N
Homo	sapiens	-	Human	African	3
Homo	sapiens	-	Human	Non-African	6
Pan	troglodytes	ellioti	Nigeria-Cameroon chimpanzee	-	10
Pan	troglodytes	schweinfurthii	Eastern chimpanzee	-	6
Pan	troglodytes	troglodytes	Central chimpanzee	-	4
Pan	troglodytes	verus	Western chimpanzee	-	5
Pan	paniscus	-	Bonobo	-	13
Gorilla	beringei	graueri	Eastern lowland gorilla	-	3
Gorilla	gorilla	diehli	Cross river gorilla	-	1
Gorilla	gorilla	gorilla	Western lowland gorilla	-	27
Pongo	abelii	-	Sumatran orangutan	-	5
Pongo	pygmaeus	-	Bornean orangutan	-	5

Table S1: Distribution of analyzed individuals in the GAGP by genus, species, subspecies, and population. Individuals were sequenced to a mean 25-fold coverage [74].

Compartment	**bp (hg18)**
NCNR	1230819358
Early replication	683958817
Late replication	610749359
ERV	126631021
Heterochromatin	429839988
LINE	552512327
Early replication, non-repetitive	349059126
Early replication, repetitive	330655407
Late replication, non-repetitive	311710121
Late replication, repetitive	296248298
Human recombination coldspots	254546577
Human recombination hotspots	254370558
ERVs, hmC+	16318060
ERVs, hmC-	110312961
Human maternal hotspots	234167393
CpG islands	29082042

Table S2: Sizes of compartments examined in great apes.

Individual	Species	Max. diff. in likelihood
Aris	Orangutan	2.82
Carl	Chimpanzee	1.39
Dennis	Chimpanzee	2.61
Dylan	Chimpanzee	1.46
Efata	Gorilla	1.10
Marlies	Chimpanzee	1.48
Marlon	Chimpanzee	1.31
Mutasi	Gorilla	1.09
Pat	Chimpanzee	1.66
Ruud	Chimpanzee	3.09

Table S3: Fold-differences in goodness-of-fit for the de novo mutation (DNM) spectrum of each individual from Besenbacher et al. (2018) with single nucleotide variant (SNV) spectra from orangutans, gorillas, and chimpanzees. The mutations in DNM spectra are too sparse to confidently distinguish between the goodness-of-fit to any particular species' SNV spectrum.

Species 1	Species 2	ρ	P-value
Human	Chimpanzee	0.82	5.31E-25
Human	Bonobo	0.81	1.13E-23
Human	Gorilla	0.77	5.11E-20
Human	Sumatran orangutan	0.89	5.19E-33
Human	Bornean orangutan	0.84	1.59E-26
Chimpanzee	Bonobo	0.94	9.27E-45
Chimpanzee	Gorilla	0.94	1.80E-46
Chimpanzee	Sumatran orangutan	0.91	1.24E-37
Chimpanzee	Bornean orangutan	0.85	4.69E-28
Bonobo	Gorilla	0.93	5.97E-42
Bonobo	Sumatran orangutan	0.93	1.83E-41
Bonobo	Bornean orangutan	0.91	1.12E-37
Gorilla	Sumatran orangutan	0.87	5.12E-31
Gorilla	Bornean orangutan	0.86	9.76E-29
Sumatran orangutan	Bornean orangutan	0.95	1.54E-47

Table S4

A table of Pearson's correlation tests of the log ratios of mutation types comparting late to early replication timing compartments between each pair of species. P-values are uncorrected.

Species 1	Species 2	ρ	P-value
Human	Chimpanzee	0.96	1.49E-16
Human	Bonobo	0.97	5.98E-17
Human	Gorilla	0.94	5.93E-14
Human	Sumatran orangutan	0.91	1.09E-11
Human	Bornean orangutan	0.93	3.86E-13
Chimpanzee	Bonobo	0.99	2.14E-22
Chimpanzee	Gorilla	0.98	9.75E-21
Chimpanzee	Sumatran orangutan	0.96	8.02E-16
Chimpanzee	Bornean orangutan	0.97	1.10E-16
Bonobo	Gorilla	0.98	7.64E-20
Bonobo	Sumatran orangutan	0.96	5.41E-16
Bonobo	Bornean orangutan	0.98	7.89E-19
Gorilla	Sumatran orangutan	0.96	1.59E-15
Gorilla	Bornean orangutan	0.96	2.78E-16
Sumatran orangutan	Bornean orangutan	0.98	2.29E-21

Table S5

PCAs run on individual mutation spectra from 8 different compartments are highly correlated among species, demonstrating uniform dosages of species mutational signatures. We calculated the midpoint of all individuals' mutation spectra for each compartment in PC1-3 and for each species. We then calculated the distance between the midpoints for each pair of compartments, computing a total of $\binom{8}{2} = 24$ distances for each species. We ran Pearson correlation tests between the paired vectors of distances between each combination of species, performing $\binom{6}{2} = 15$ tests; uncorrected P-values and estimates of ρ are listed above.

Species	P-value
Human	1.90E-05
Chimpanzee	2.90E-12
Bonobo	2.03E-07
Gorilla	3.58E-09
Sumatran orangutan	0.00245204
Bornean orangutan	0.00014367

Table S6

Table of the P-values from a Chi-square test for different rates of CG>GG mutation types comparing the ERV hmC+ to hmC- compartments. P-values are uncorrected.

Compartment	**Size (human, hg19)**	**Size (mouse, mm10)**
Promoter	45669288	37074578
Enhancer	109737822	112146530
Insulator	21727890	18373998
Transcribed	106414354	148527127
Repressed	36676448	181894919
Heterochromatin	1845415220	146461481

Table S7: Sizes of chromHMM compartments in mouse and human in bp.

SUPPLEMENTARY FIGURES

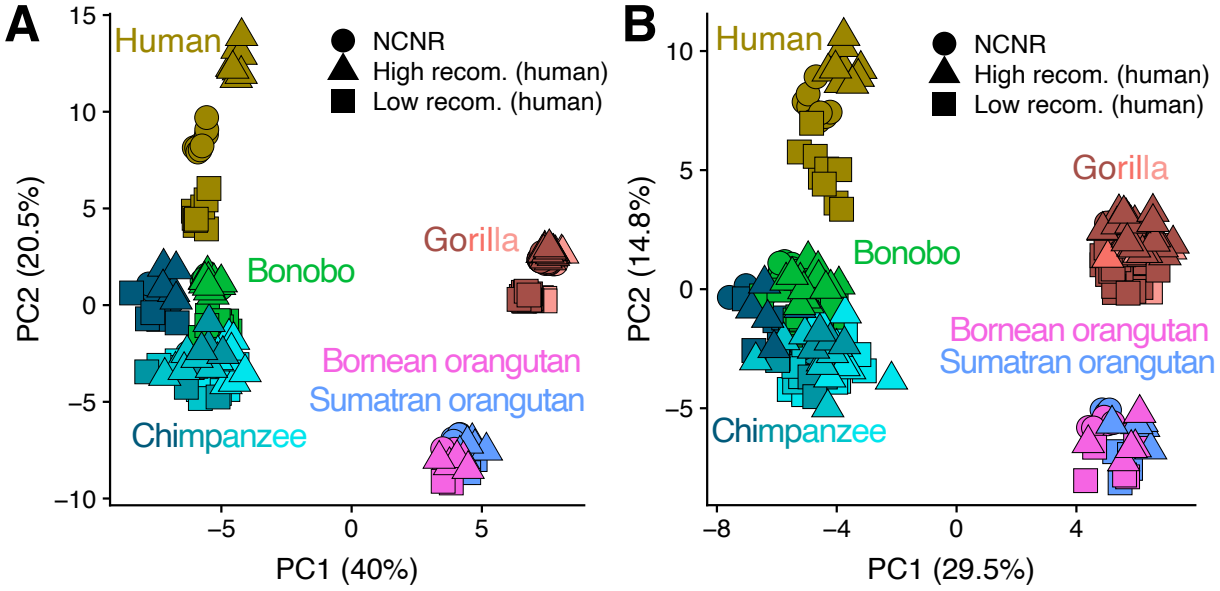


Figure S1: The effects of gBGC on mutation spectrum variation in recombination hotspots and coldspots is minimized by randomization method. We defined two compartments as those representing the top and bottom 10% of the genome ranked by local recombination rate in humans. The SNVs included in the spectra presented in (B) were randomly assigned to a single haplotype; this method was not applied in (A). The recombination hotspot mutation spectra in humans separate cleanly in (A); this separation decreases with divergence time from humans. These trends align with the rapid evolution of recombination hotspots, and therefore local rate of gBGC. The differentiation of recombination hotspot mutation spectra in (B) nearly disappears with the randomization method, thus demonstrating its minimizing effect on gBGC.

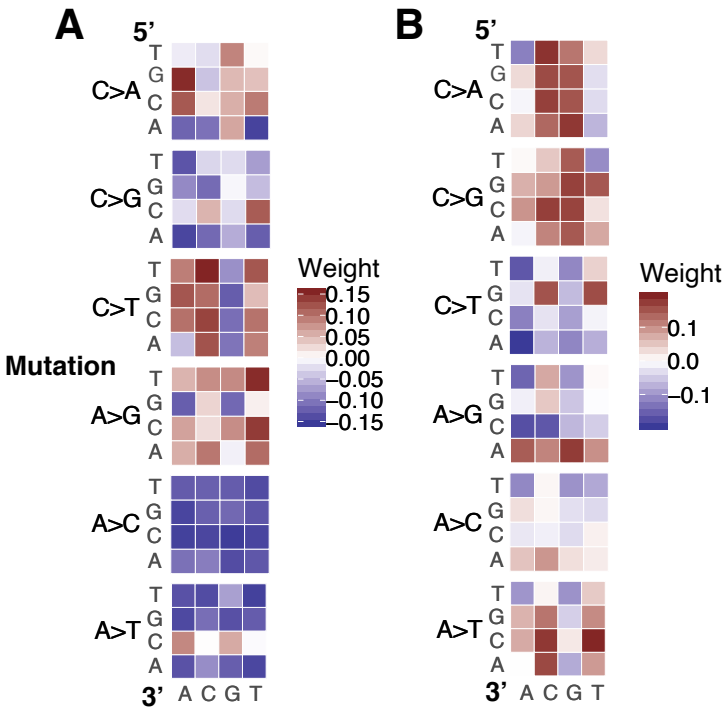


Figure S2: The loadings of PC1 (A) and PC2 (B) from PCA of NCNR spectra across all 88 individuals in the GAGP.

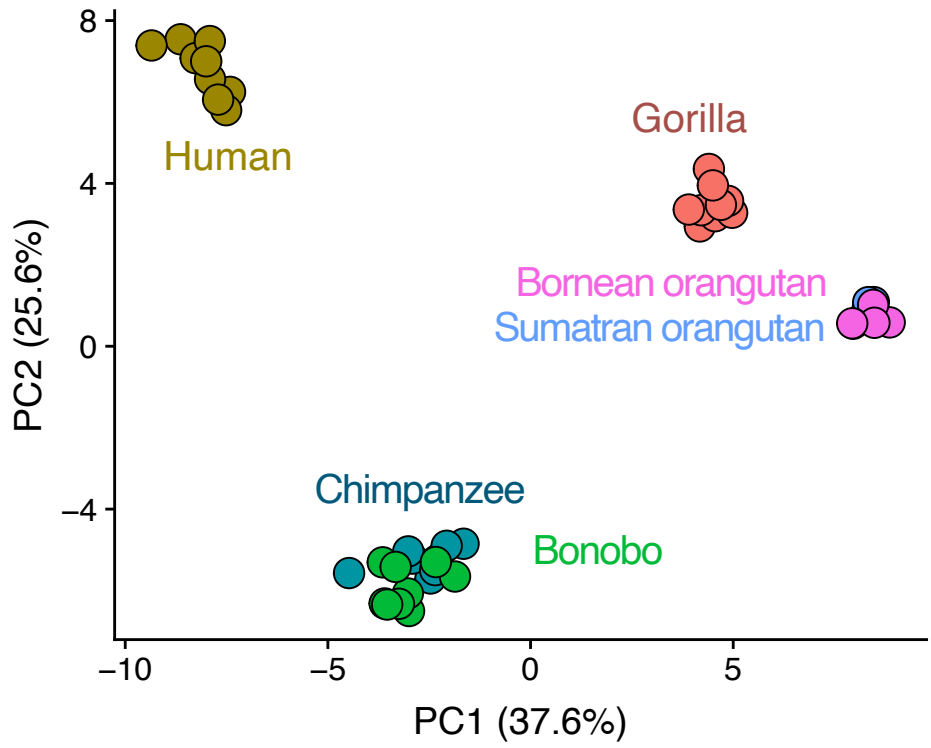


Figure S3

NCNR PCA clustering is robust to differences in species representation. We down-sampled each species (or both orangutan species, grouped together) to nine individuals and generated a PCA.

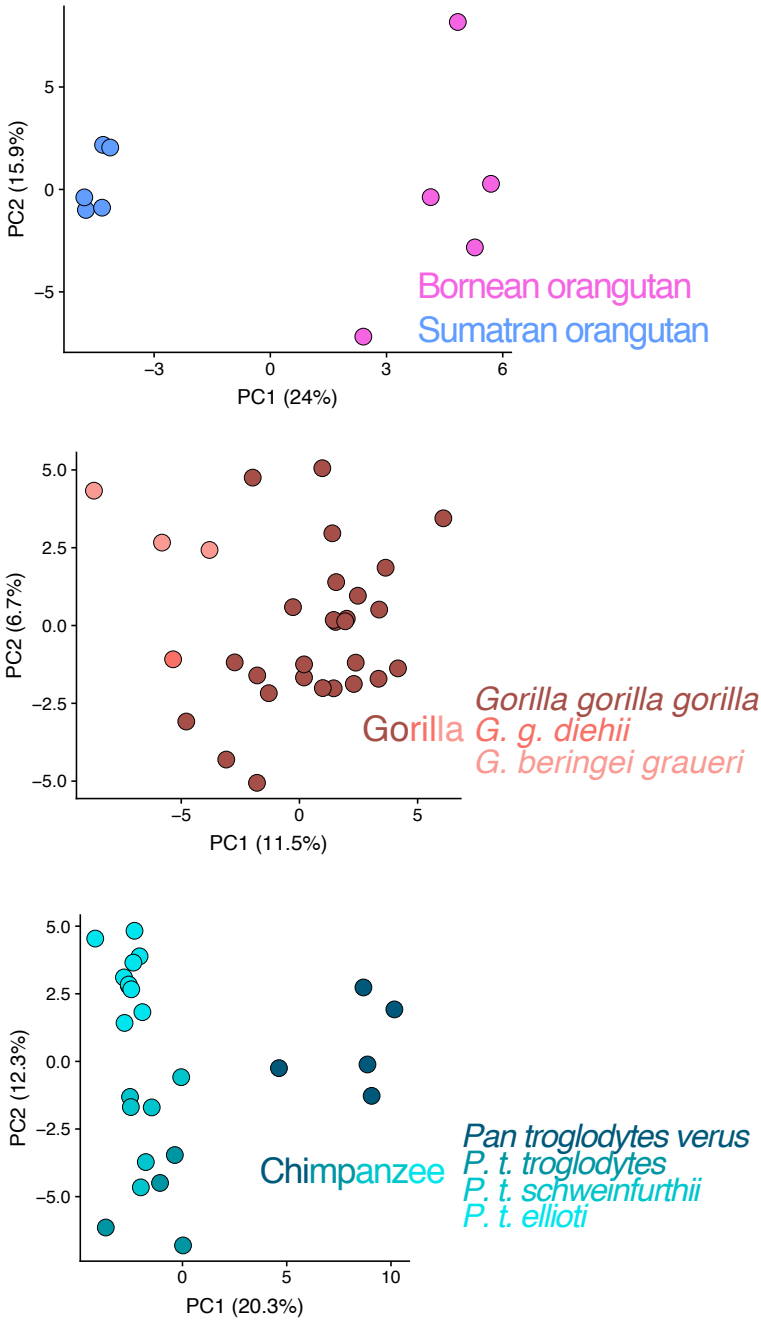


Figure S4

PCAs of the NCNR compartment for the orangutan clade, gorillas, and chimpanzees demonstrate finer-scale separation and clustering of individuals by subspecies.

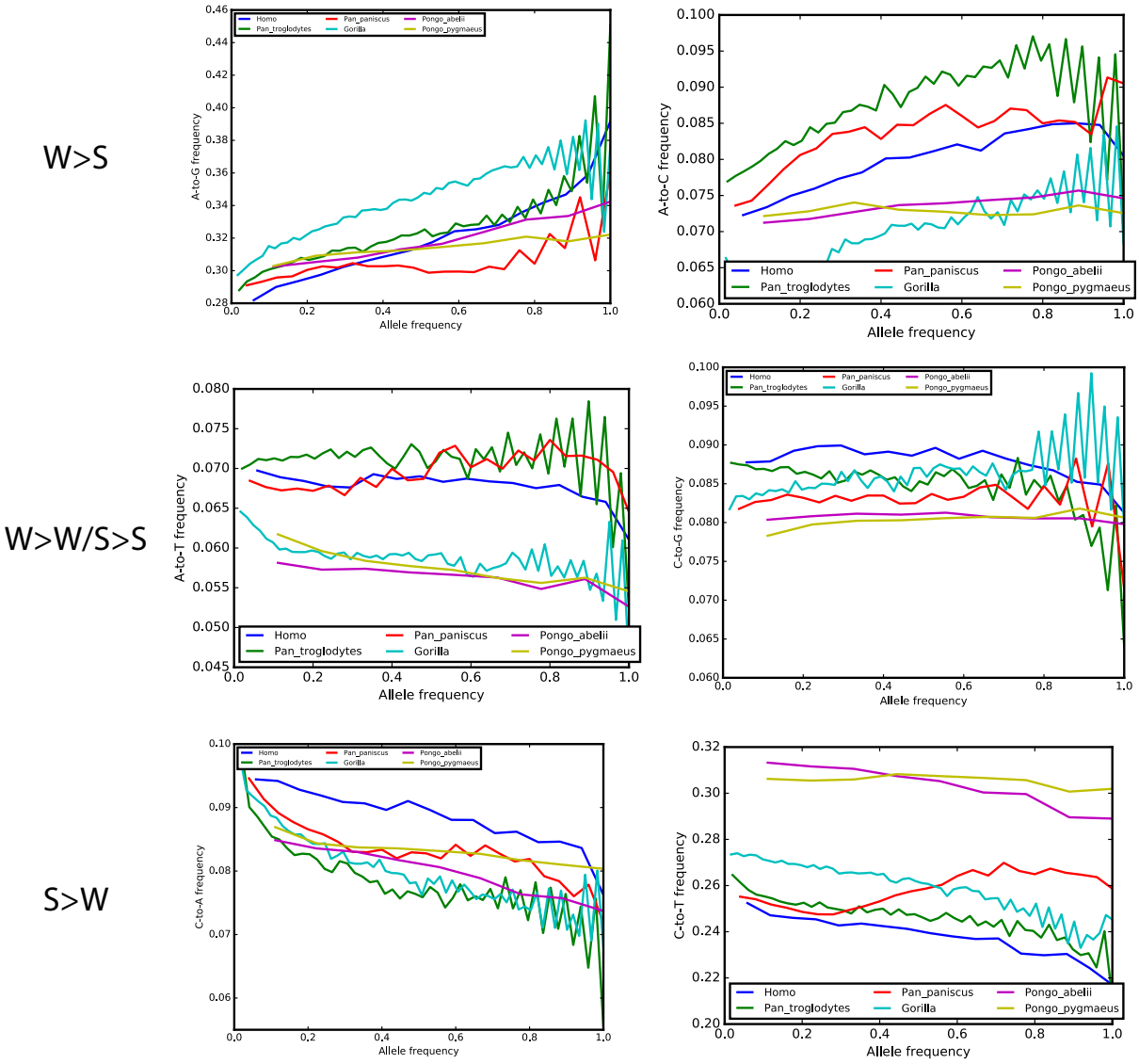


Figure S5

Effects of gBGC on mutation spectra as a function of allele frequency across great apes. These graphs show how each species' non-context-dependent mutation spectrum varies as a function of allele frequency. In general, high allele frequency SNVs show enrichment for W>S mutations and depletion for S>W mutations; this trend, in concordance with the effects of gBGC, is consistent across all species. These plots suggest that gBGC may have skewed the spectra of

high frequency alleles to a greater extent in some species, but the spectra of low frequency alleles (which are less affected by gBGC) have spectra that are no more similar across species than high frequency alleles.

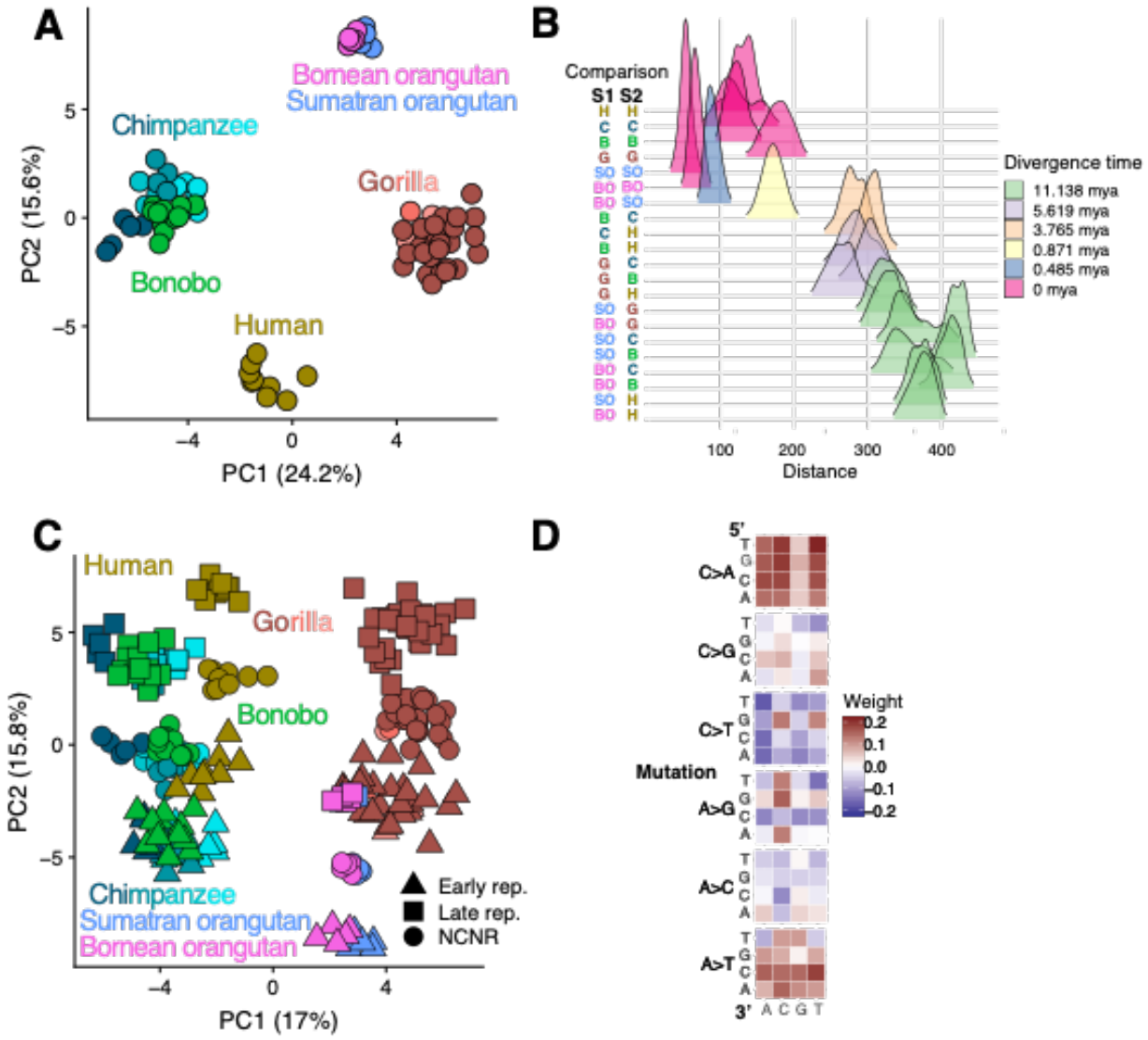


Figure S6: Rare variants demonstrate similar mutation spectrum trends to common variants

A. NCNR mutation spectra for 88 individuals in the GAGP cluster by species, demonstrating rapid, lineage specific evolution of mutation spectra. Different colors imply different species, and different shades of chimpanzees and gorillas are different subspecies.

B. Euclidean distances between mutation spectra of intra- and inter-specific pairs of individuals demonstrates an association of mutation spectrum distance with divergence time.

C. PCA of mutation spectra from NCNR, early replication, and late replication timing compartments for 88 individuals shows separation along orthogonal “phylogenetic” and “replication timing” axes.

D. The loadings of PC2, which roughly corresponds to the “replication timing” axis, shows an enrichment for C>A and A>T mutation types. Late replication timing regions are known to be enriched for this mutation signature; this recapitulates results from using rare and common variants together.

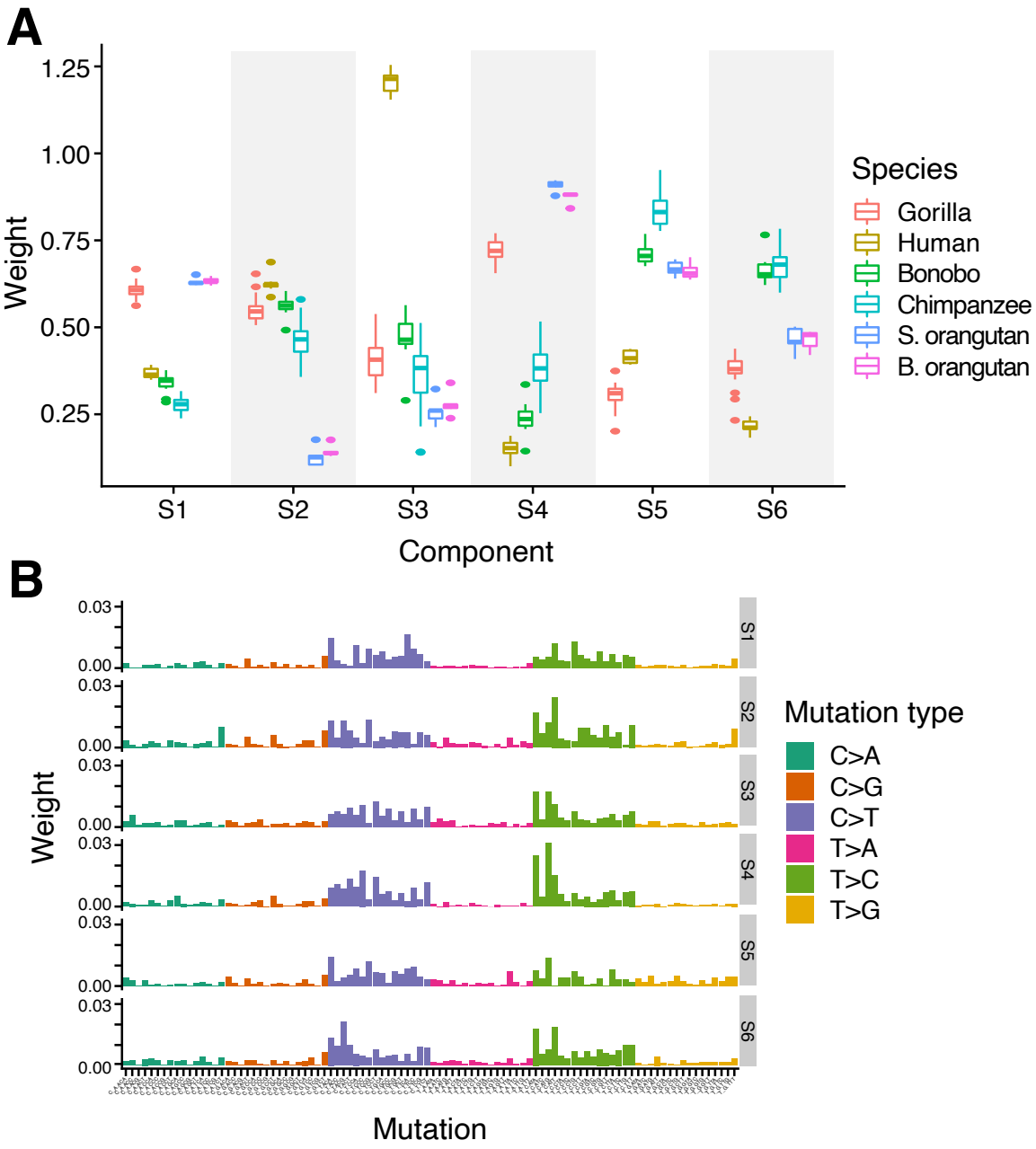


Figure S7: NMF identifies rapidly-evolving germline mutation signatures that are specific to great ape lineages.

A. The dosages of mutation signatures appear lineage-specific. The boxplots shows the distribution of each signature's weight for each species.

B. Different signatures are enriched for different mutation types.

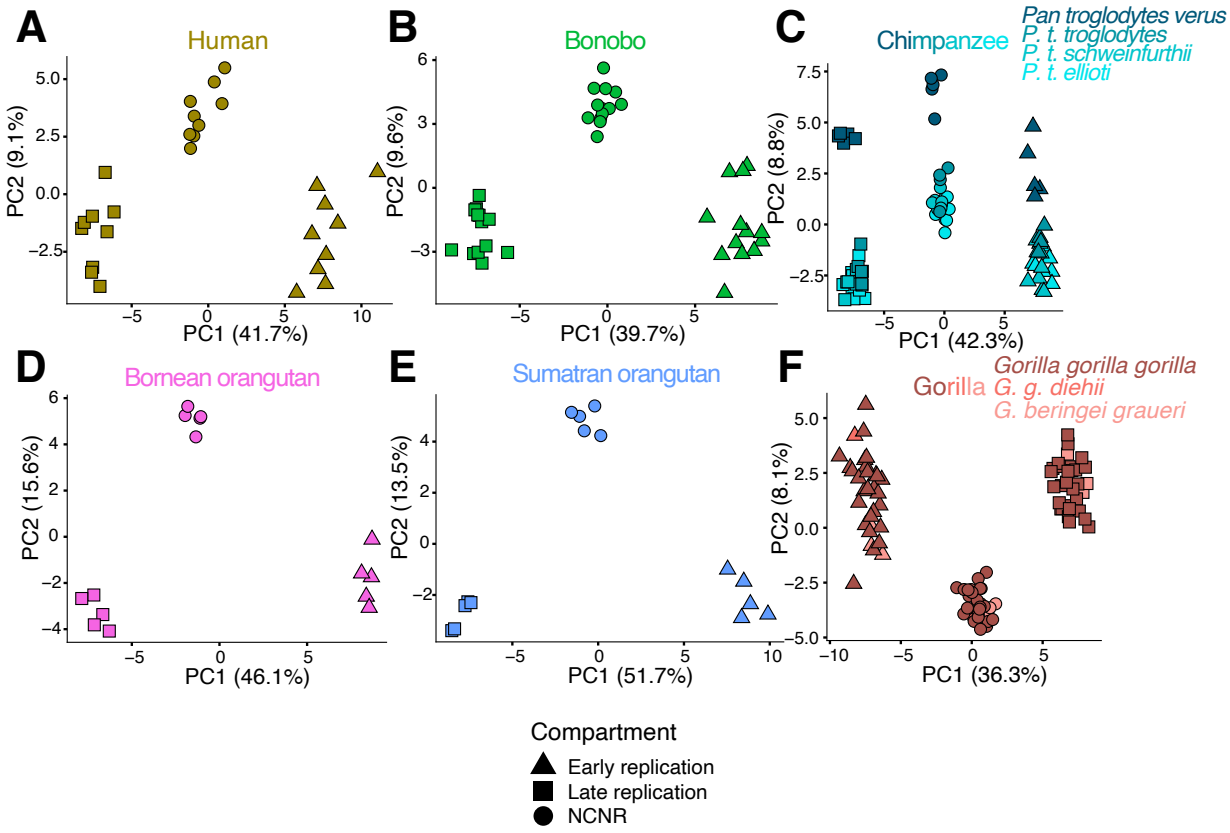


Figure S8

PCA of replication timing and NCNR compartments for individuals in each species. Each point in these PCA represents the mutation spectrum from a single individual's NCNR, early replicating, or late replicating compartment. Clustering by compartment implies differences in mutation spectra based on replication timing.

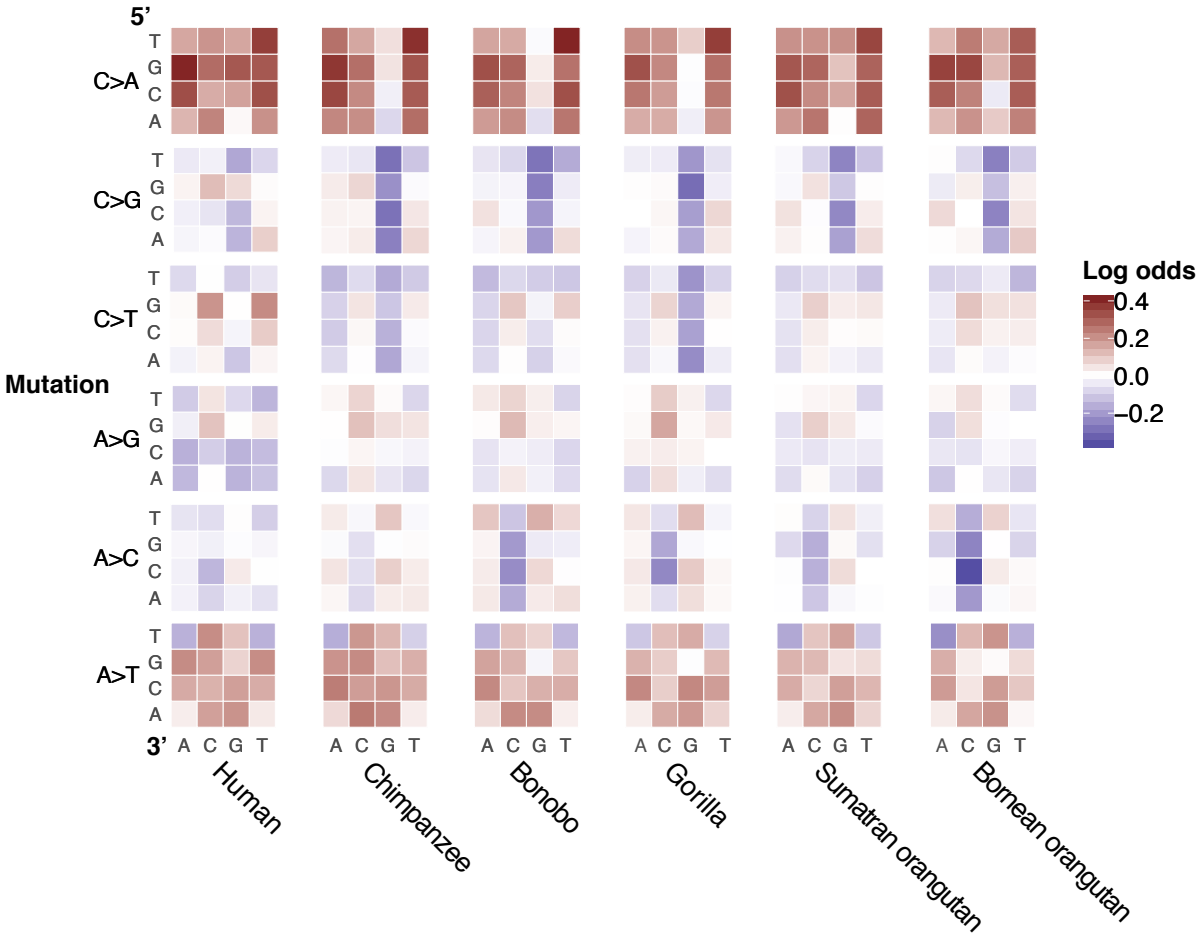


Figure S9

A heatmap of the log ratios of triplet mutation fractions in each species shows an enrichment for C>A and A>T mutations in late replicating compartment compared to early replicating compartment. This mutation signature recapitulates recently described late replication timing signature in humans. To generate the species mutation spectra, we counted the number of SNVs with triplet context segregating within a species that occurred in each compartment. The triplet mutation fractions were normalized by compartment nucleotide content. Statistical quantification of the correlation of these heatmaps can be found in Figure S10.

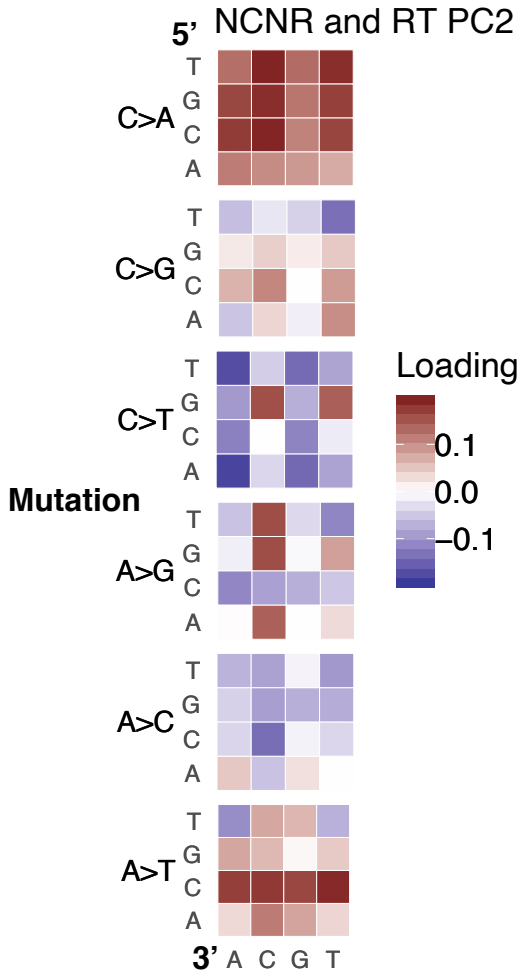


Figure S10

A heatmap showing the PC2 weights associated each triplet mutation type in Figure 2D. The C>A and A>T mutation types dominate PC2, which correlates with a late replication timing signature.

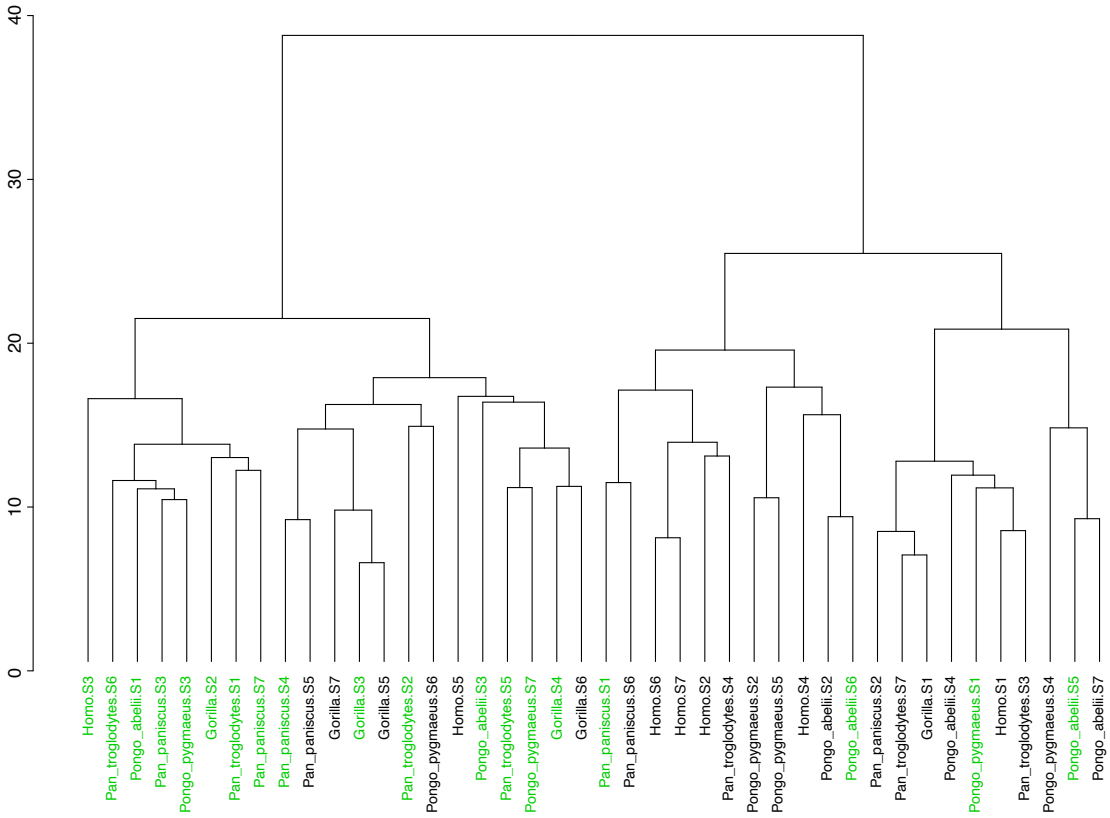


Figure S11. Mutation signatures associated with late-replication timing cluster together. NMF was run on early and late replication timing compartments separately for each species. Distances between these mutation signatures determine the dendrogram's structure; similar signatures are closer together in the tree. Signatures colored in green are enriched in the late replication timing compartment.

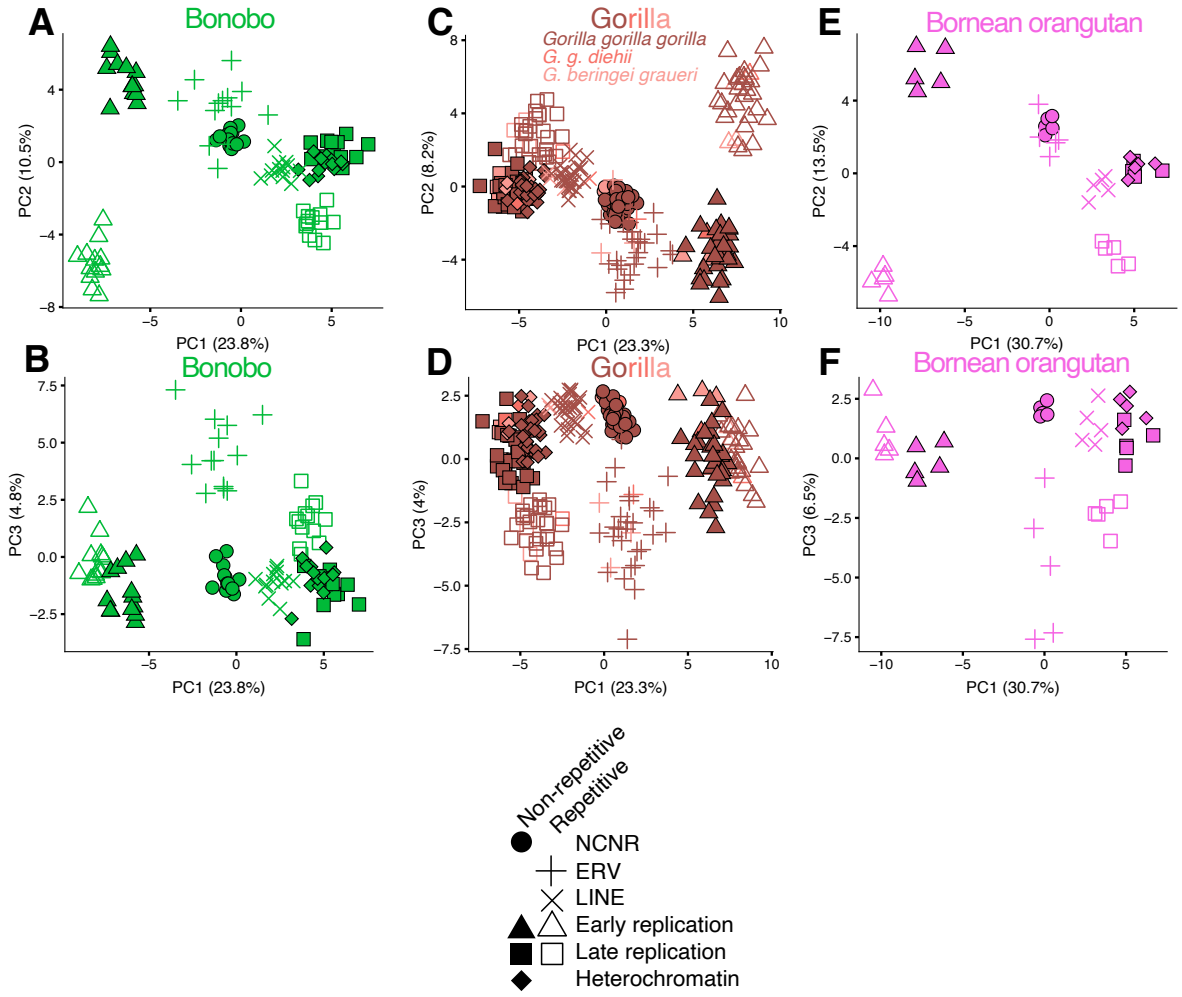


Figure S12

We defined eight overlapping functional compartments to test for evolution of mutation spectrum modifiers along axes of chromatin accessibility, replication timing, and repetitive content. We then ran a PCA on the individual mutation spectra for all eight compartments for each species separately (only bonobo, gorilla, and Bornean orangutan shown here; see Figure 3 for other species). For all species, PC1 and PC2 separate compartments along gradients that correspond to replication timing and repetitive content, respectively. PC3 separates ERVs from

other compartments. The similarities of these independent PCAs across all species implies conservation of *cis*-acting mutational signatures.

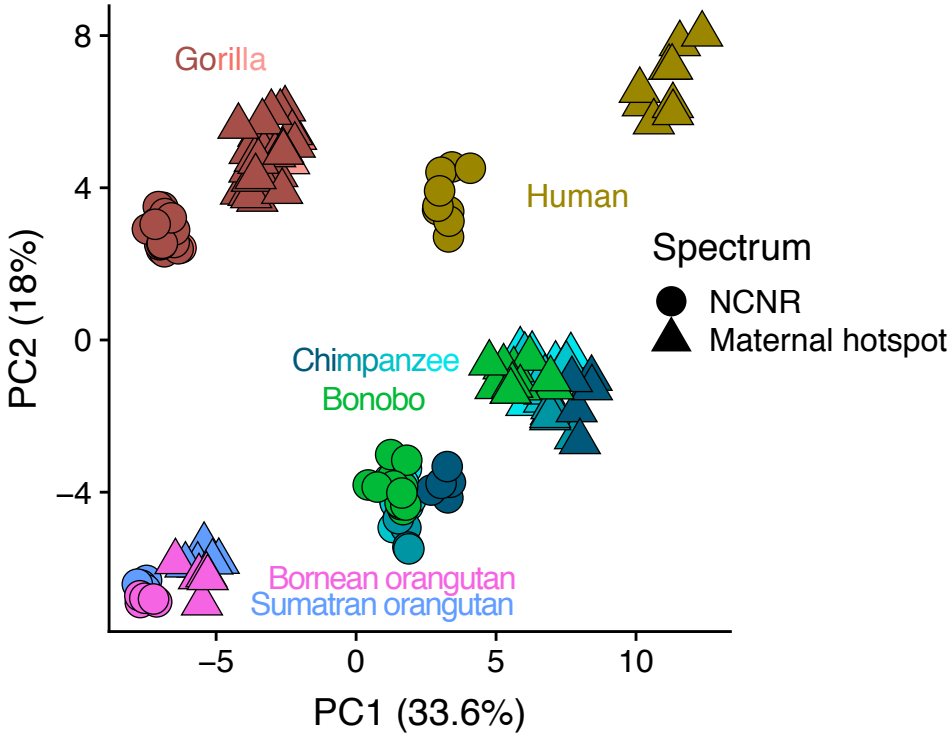


Figure S13

PCA of NCNR and maternal hotspot compartments. Individual mutation spectra for the NCNR and maternal hotspot compartments were calculated and normalized (Methods). The distance between the maternal hotspot and the NCNR mutation spectra is negatively correlated with phylogenetic distance from humans, implying evolution of a *cis*-acting mutational modifier largely absent from orangutans.

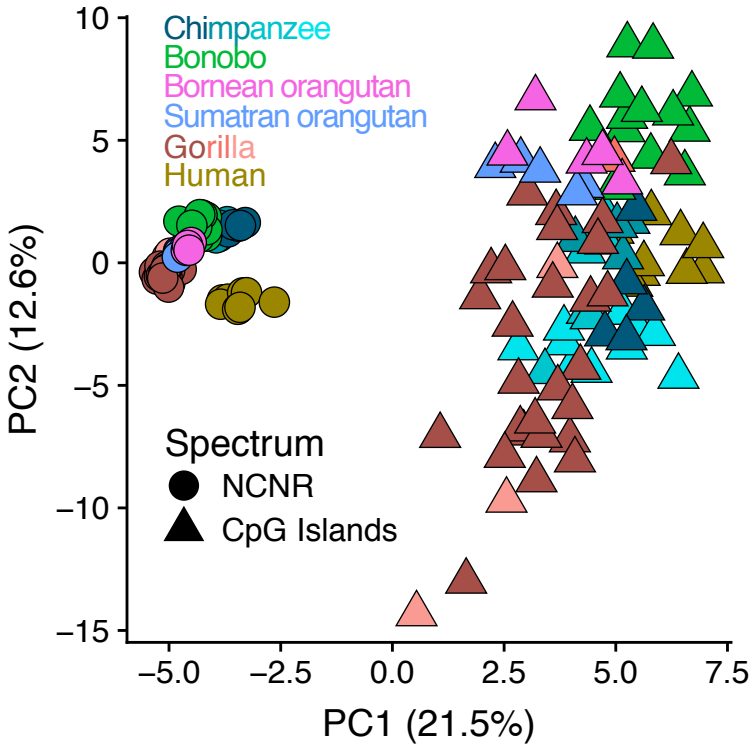


Figure S14

A PCA of NCNR and CpG island compartments demonstrates that differentiation of the CpG island compartment exceeds the magnitude of spectrum differentiation between species.

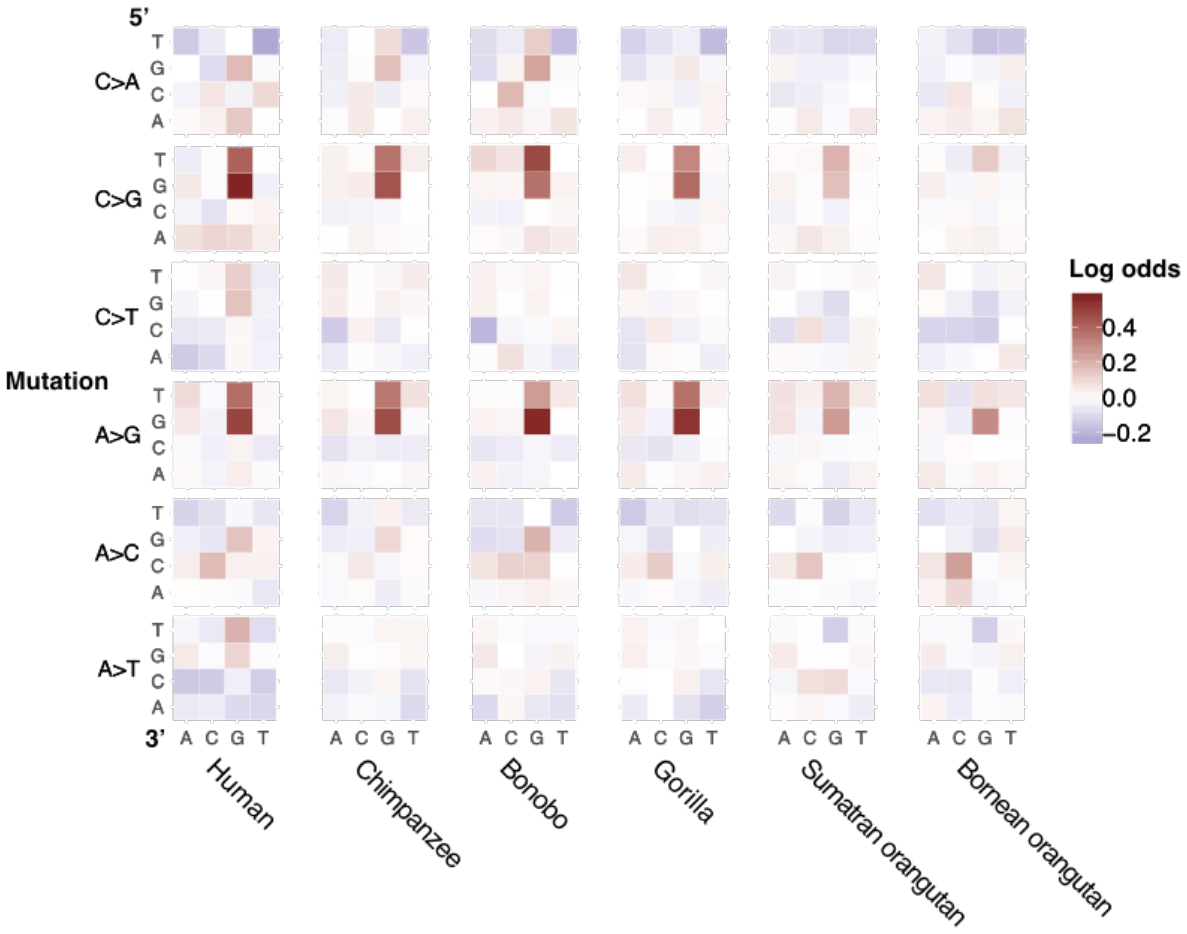


Figure S15

Differences in 7-mer content between the ERV and nonrepetitive heterochromatin compartments do not fully explain the enrichment of CG>GG mutation types in ERVs. We rescaled the counts of each 7-mer mutation type in ERVs by the ratio of the mutating 7-mer's nucleotide content between nonrepetitive heterochromatin and ERV compartments (see Methods). The heatmap shows the log ratio between the 7-mer-corrected 3-mer mutation fractions in ERVs to (uncorrected) 3-mer mutation fractions in nonrepetitive heterochromatin.

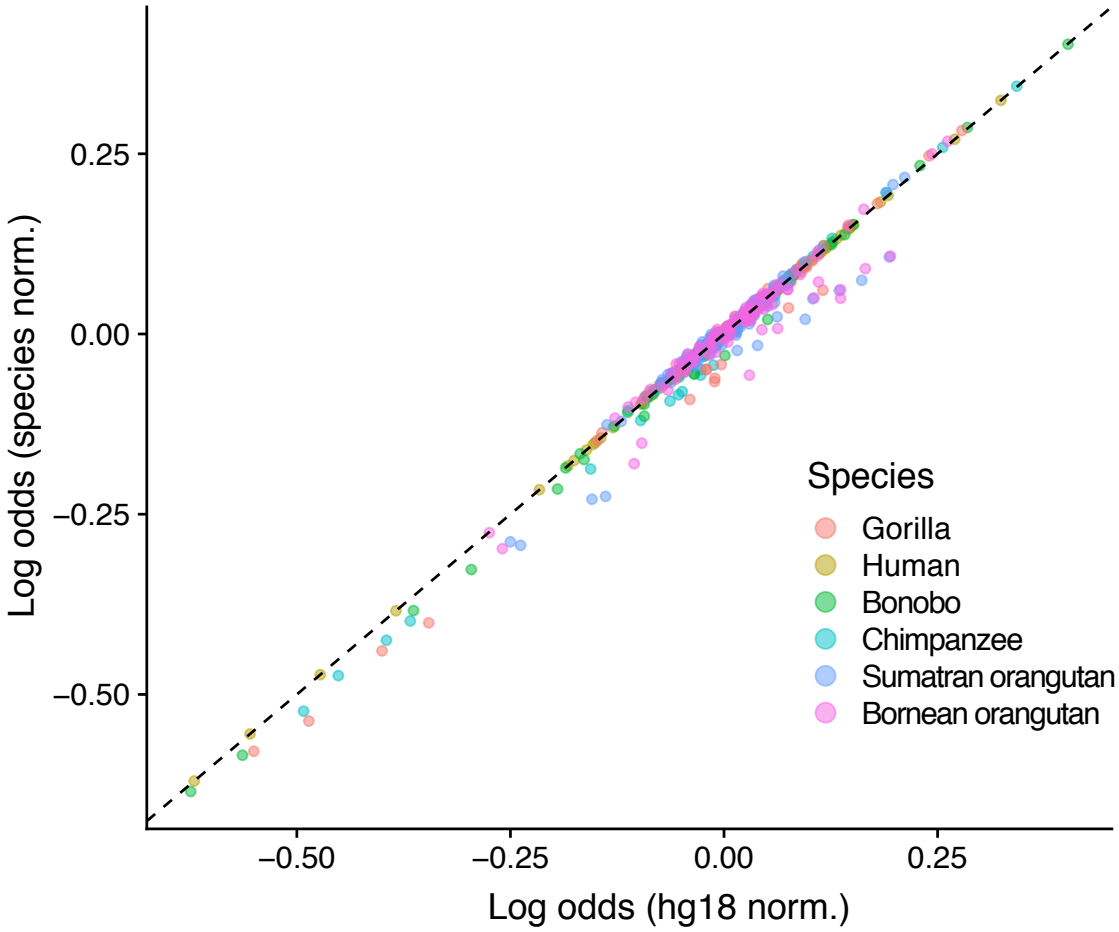


Figure S16

Normalization of mutation spectra is unaffected by variation in species liftovers. We normalized the ERV and nonrepetitive heterochromatin species mutation spectra by the nucleotide content of the genomic regions for each respective compartment the successfully lifted over to each species' reference genome. The log-odds of the ratio of ERV-nonrepetitive heterochromatin mutation spectra are plotted using this species-specific (Y axis) against our standard species-nonspecific normalization (X axis).

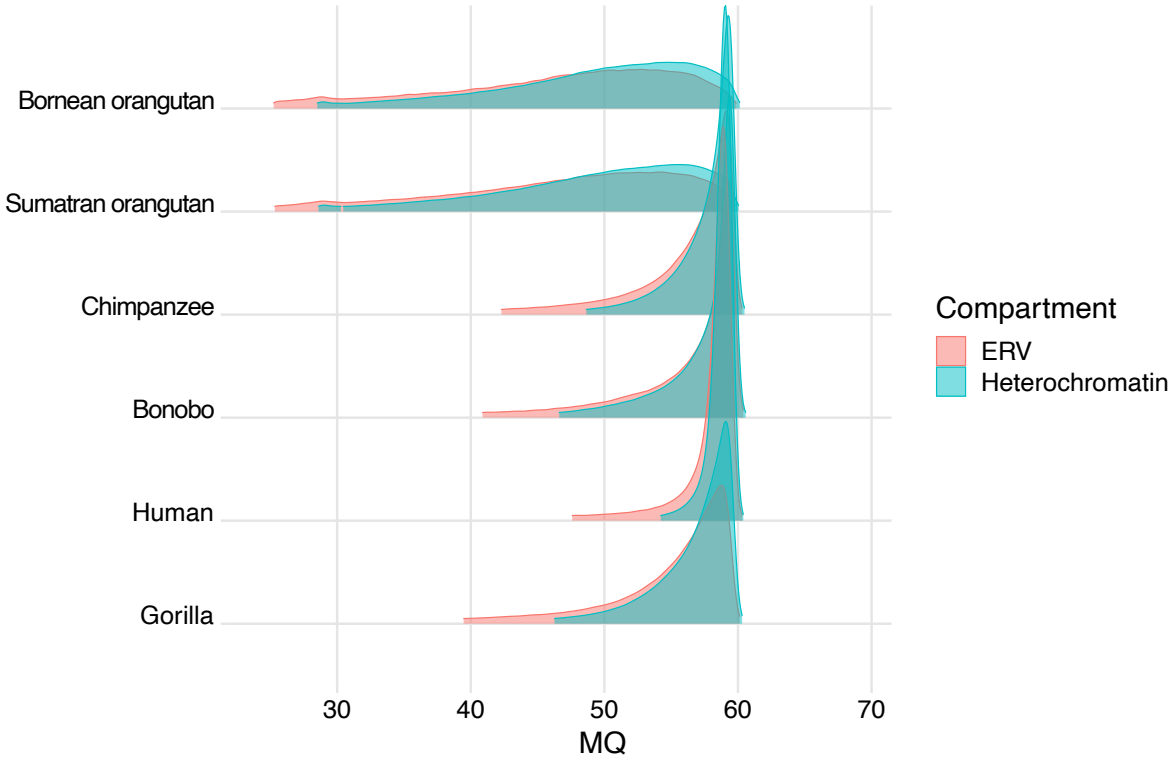


Figure S17

The distribution of SNV mapping quality does not differ substantially between the ERV and nonrepetitive heterochromatin compartments. The plot above shows the layered distributions of mapping quality (MQ in the GAGP VCF files) for SNVs included in the ERV and nonrepetitive heterochromatin compartments in red and blue, respectively. Although each ERV compartment contains a few more low quality SNPs than are present in the matched heterochromatin compartment, they are not numerous enough to explain the spectrum difference between the two regions.

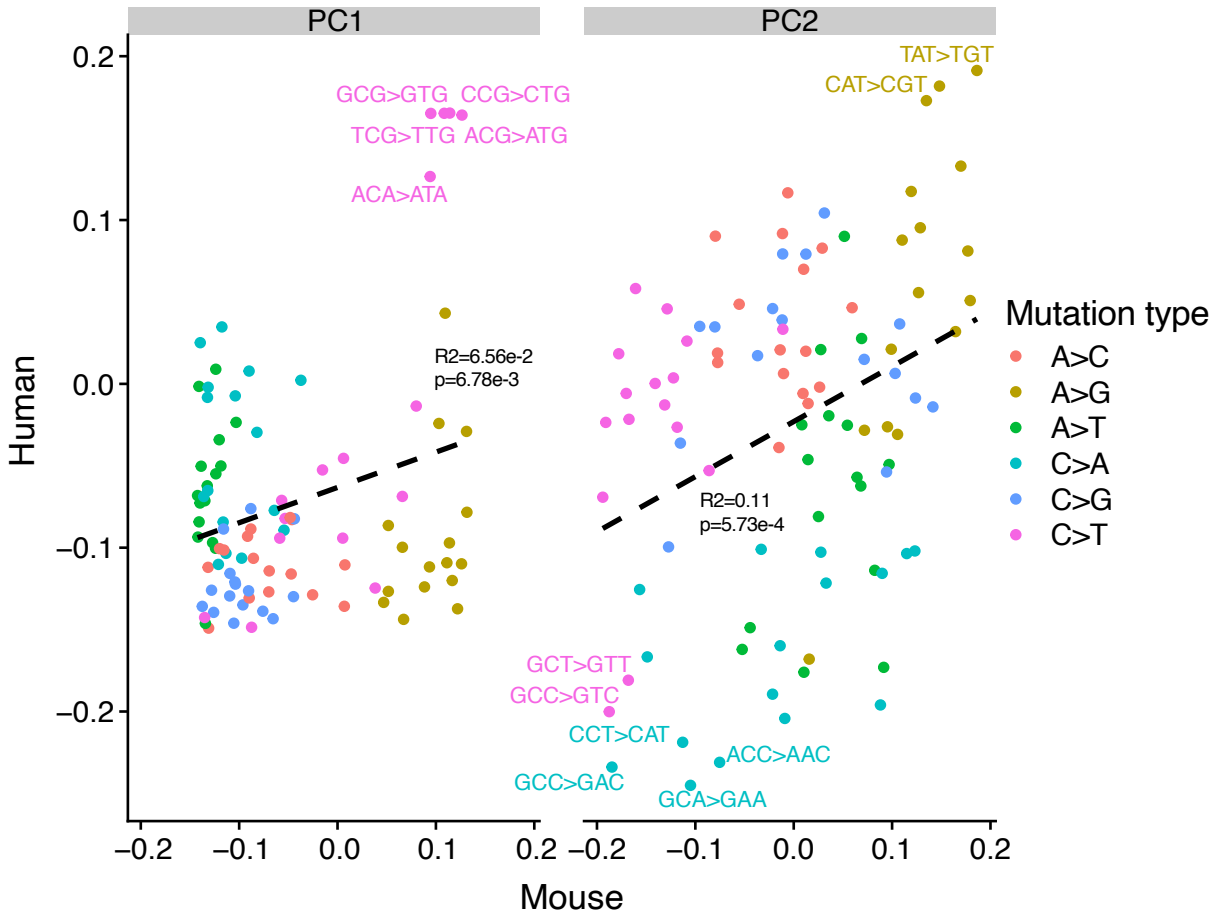


Figure S18

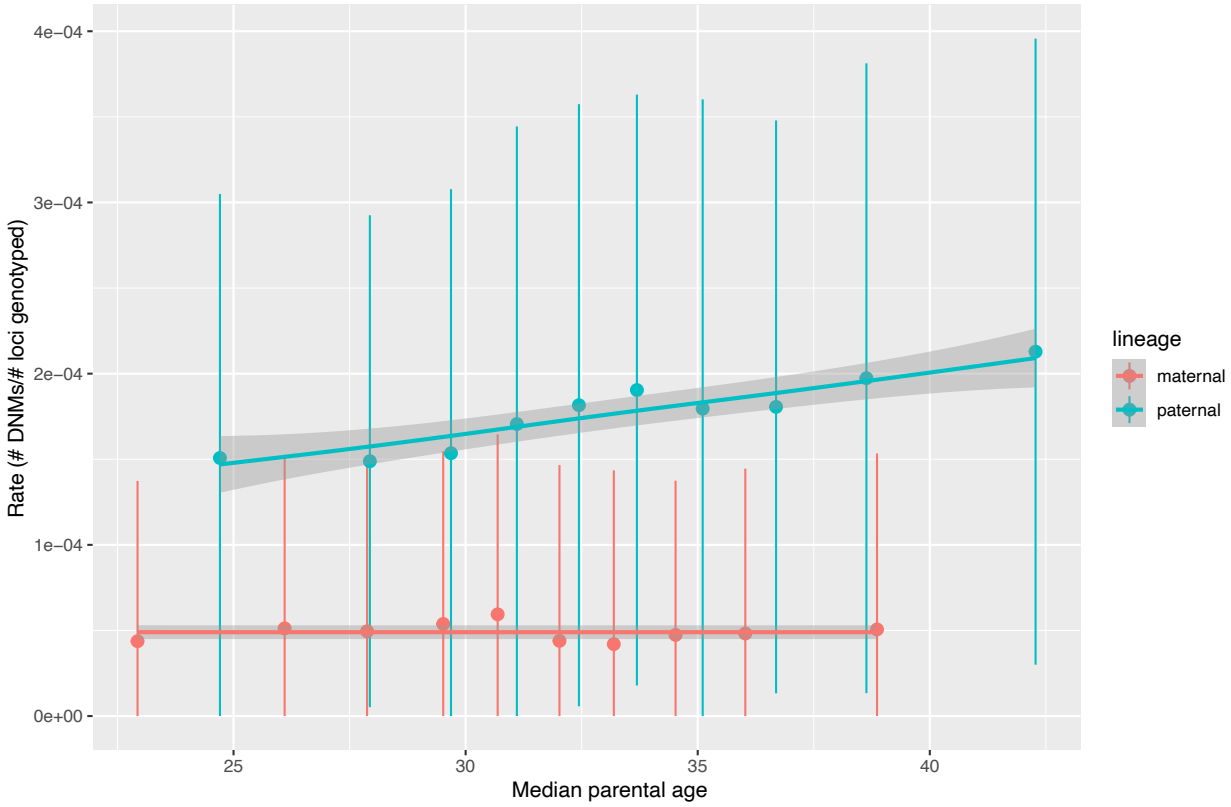
The PC loadings of human and mouse mutation spectra from different functional compartments are significantly correlated. Separate PCAs were run on matrices containing mutation spectra for a set of human and mouse individuals, whose genomes were split into compartments following chromHMM annotations. The annotations were generated for each species separately; thus, regions in the same functional compartment between the two species were not necessarily homologous. The loadings of PC1 and PC2 from the separate PCAs are significantly correlated. Similar mutation types drove the separation and clustering of mutation spectra along both PCs. The dashed line was generated by a linear regression, whose fit and slope significance are listed on the plots.

APPENDIX B

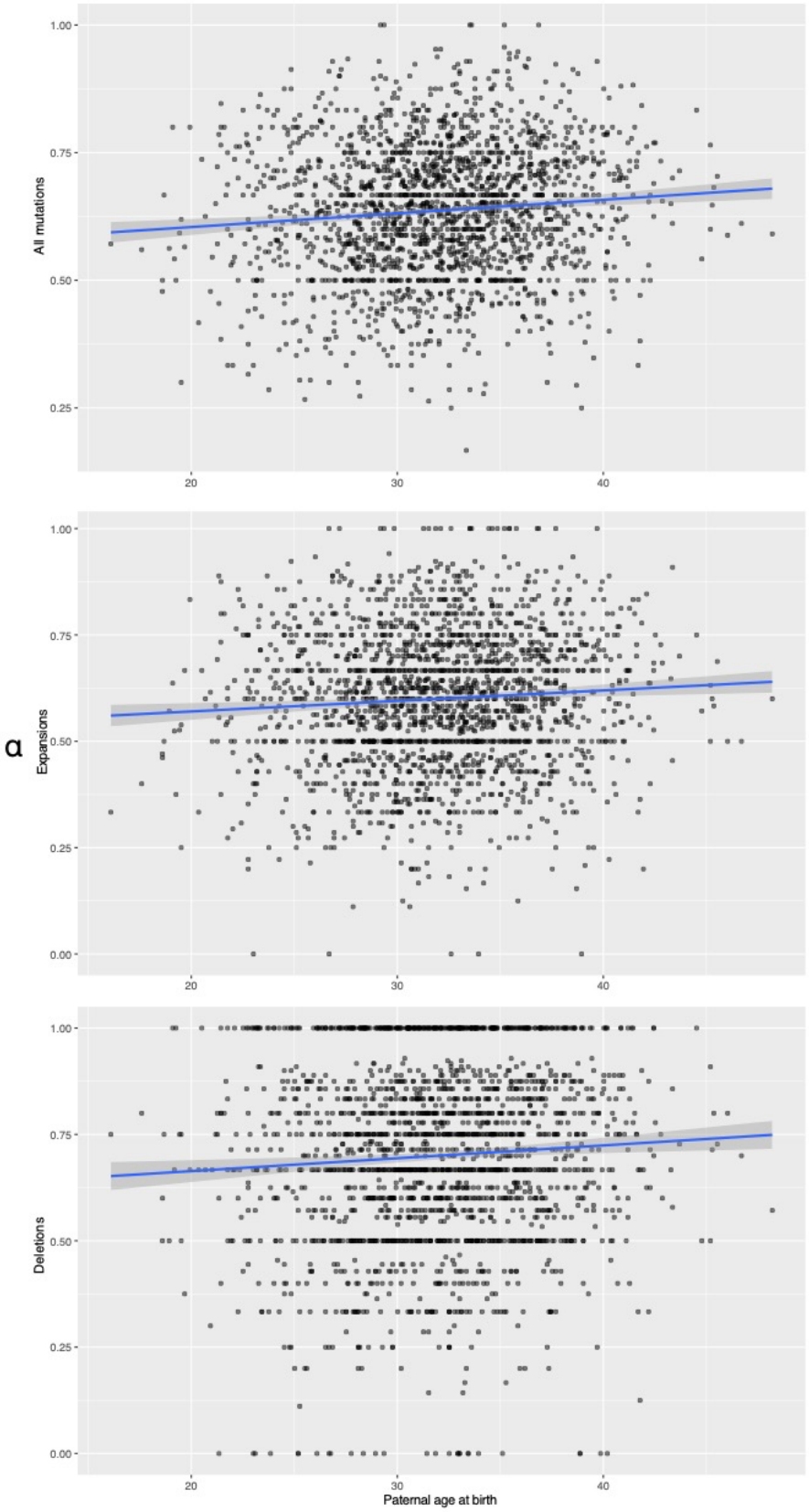
SUPPLEMENT FOR CHAPTER 3

	Inferred ancestry				
Self-reported race	African	Native American	East Asian	European	South Asian
African American	79	0	0	3	1
Asian	0	0	58	1	37
More than one race	33	2	33	113	3
Native American	0	2	0	3	0
Native Hawaiian	0	0	1	1	0
Not specified	1	5	0	11	0
Other	9	30	3	58	6
White	2	13	0	1807	1

Supplementary Table 1: Concordance of self-reported ancestry of children in the SSC and their genetic ancestry inferred with ADMIXTURE.

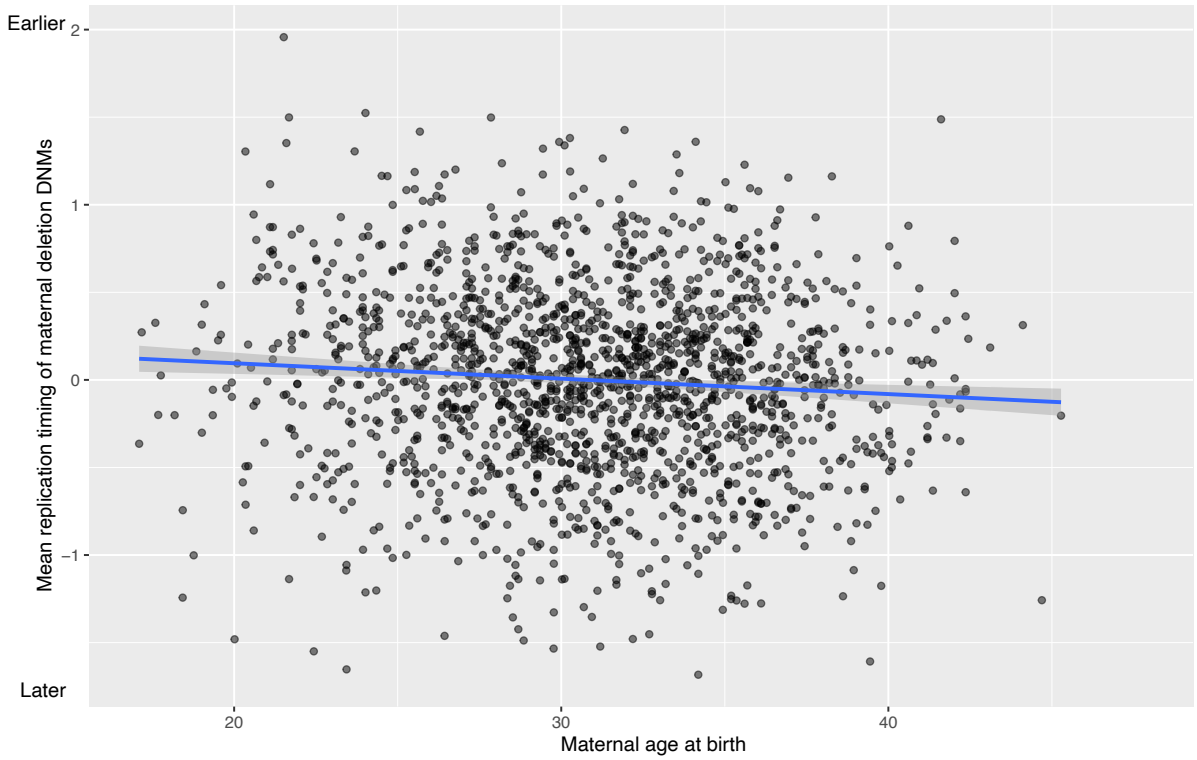


Supplementary figure 1: Association of parental age with mutagenesis at the STR loci genotyped in a previous study [104]. In concordance with the prior study, we calculated the mutation rate for each individual using a subset of the markers genotyped in [124]. Rate was calculated as the number of mutations observed divided by the total number of markers genotyped. Individuals were binned by parental age at birth into age bins of equal size. We observed a significant paternal age effect and no significant maternal age effect using this subset of loci ($P = 5.16*10^{-7}$, 0.91 for paternal, maternal age effects; linear regression).

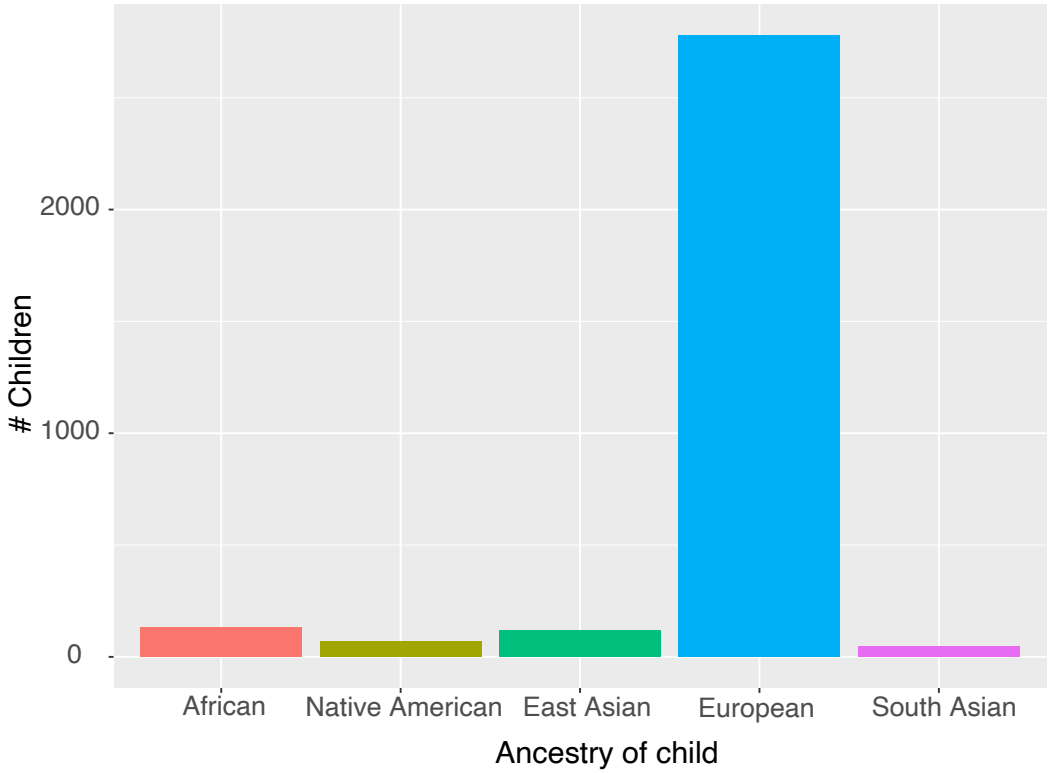

All mutations
Expansions
Deletions
Paternal age at birth
1.00
0.75
0.50
0.25
0.00
20
30
40

α

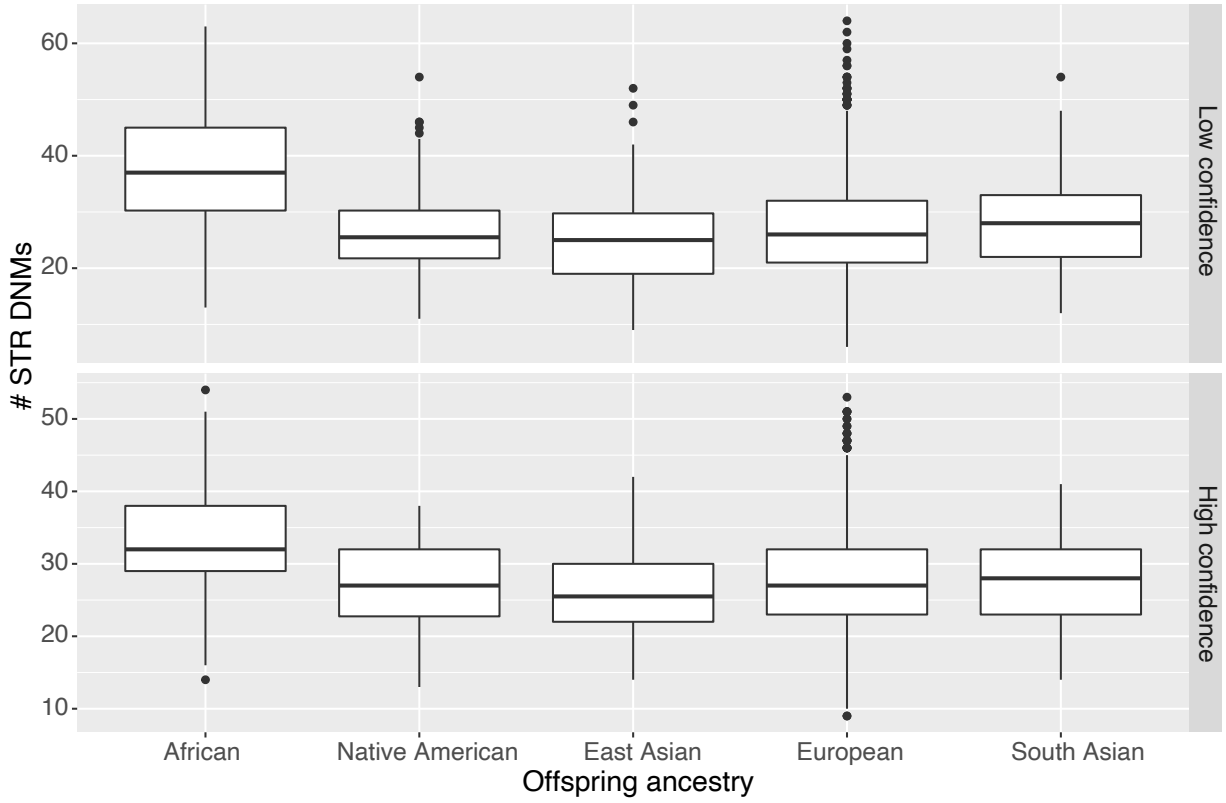
Supplementary figure 2: Paternal fraction of phased mutations is associated with paternal age. We observe a significant positive association of alpha with paternal age when considering all mutations and restricting to only expansions or only deletions (P = 1.47e-05, $1.42*10^{-3}$, $2.7*10^{-3}$, respectively; quasibinomial regressions with identity link functions, uncorrected for multiple testing). Following [29], we limited families to those in which the maternal and paternal ages at birth differed by less than 10%. The estimated slopes were $2.67 * 10^{-3}$, $2.47*10^{-3}$, and $3.02 * 10^{-3}$, respectively.



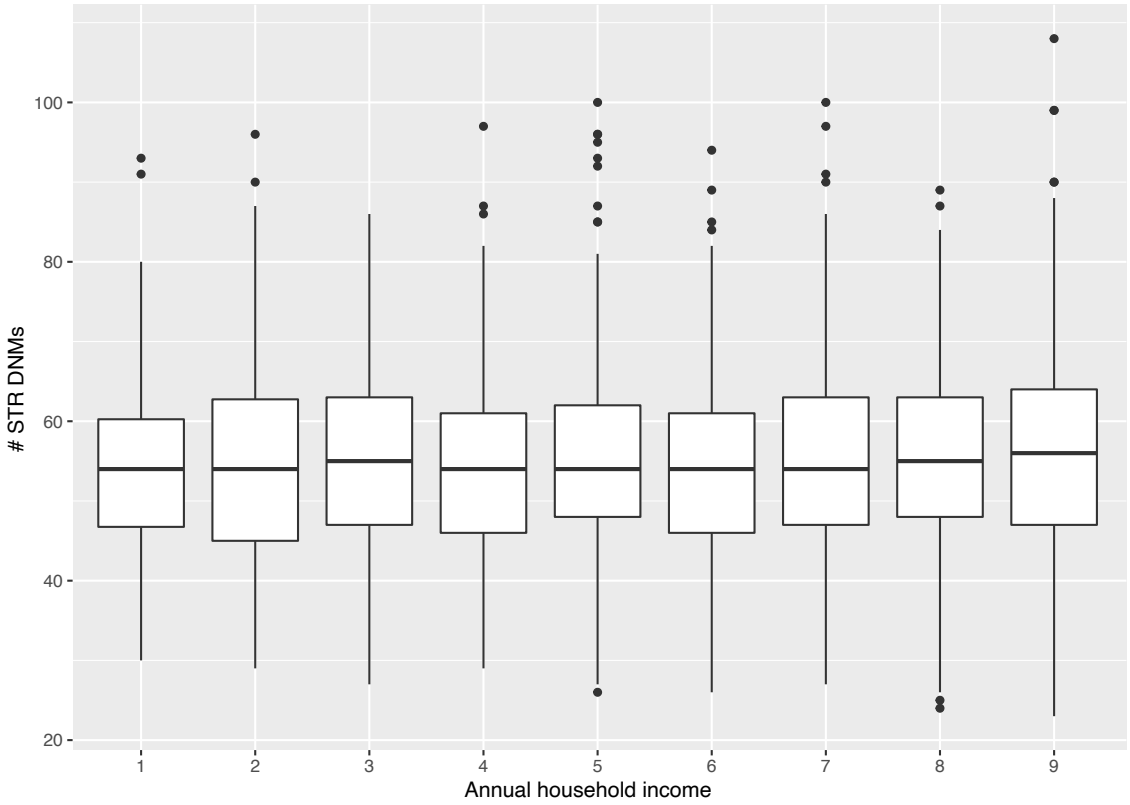
Supplementary figure 3: Maternally derived deletion DNMs occur in later replicating regions in older mothers. Line represents a generalized linear model, confidence interval calculated by stat_smooth in R.



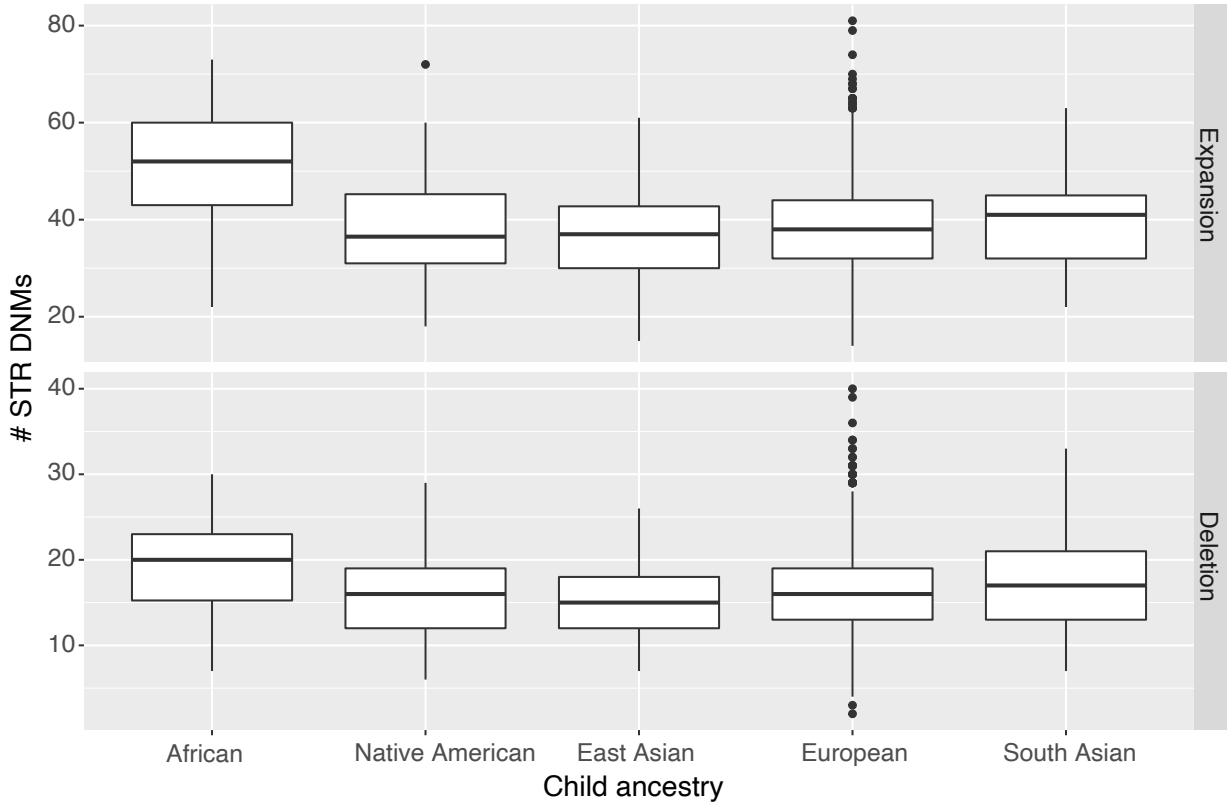
Supplementary figure 4: Children in the SSC, grouped by predominant inferred genetic ancestry.



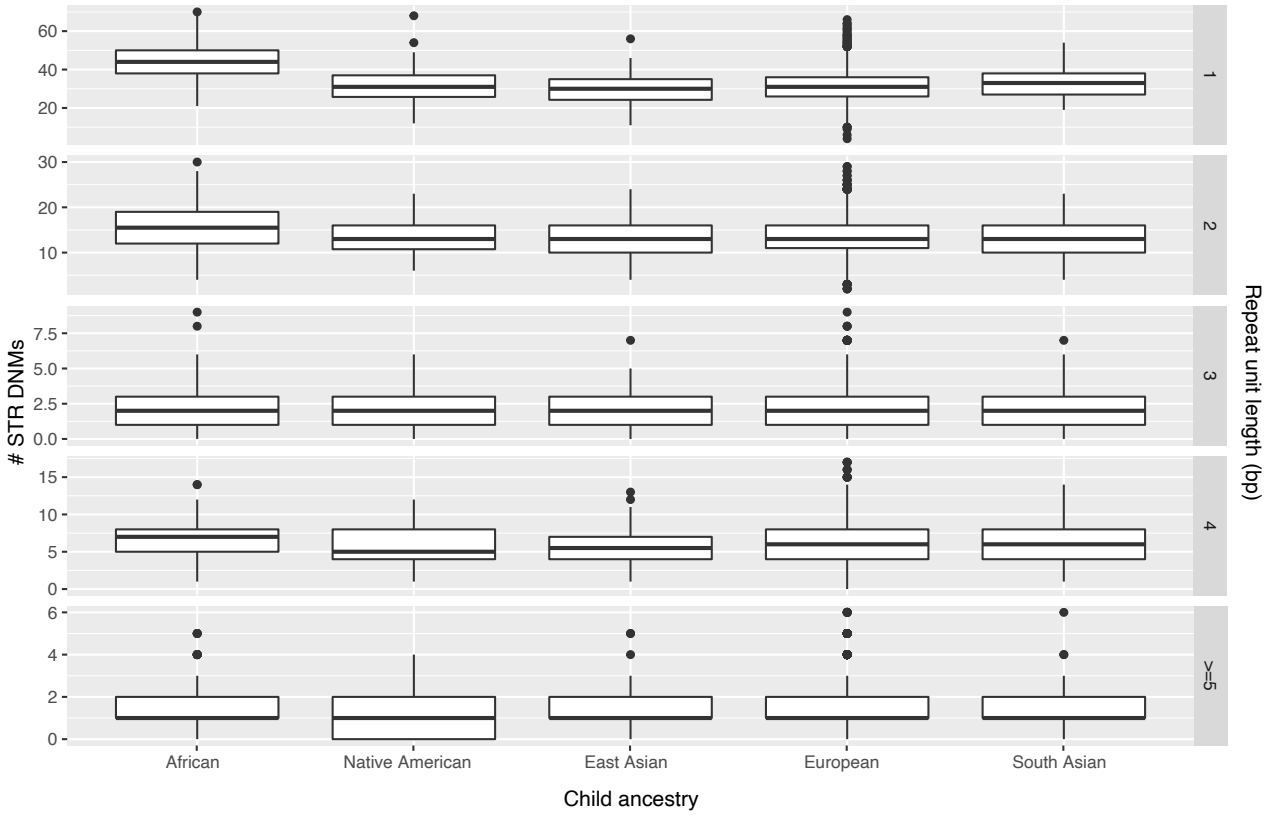
Supplementary figure 5: African ancestry is more strongly associated with the number of STR DNMs in lower confidence sites, but is still significantly associated with mutation rate at higher confidence sites.



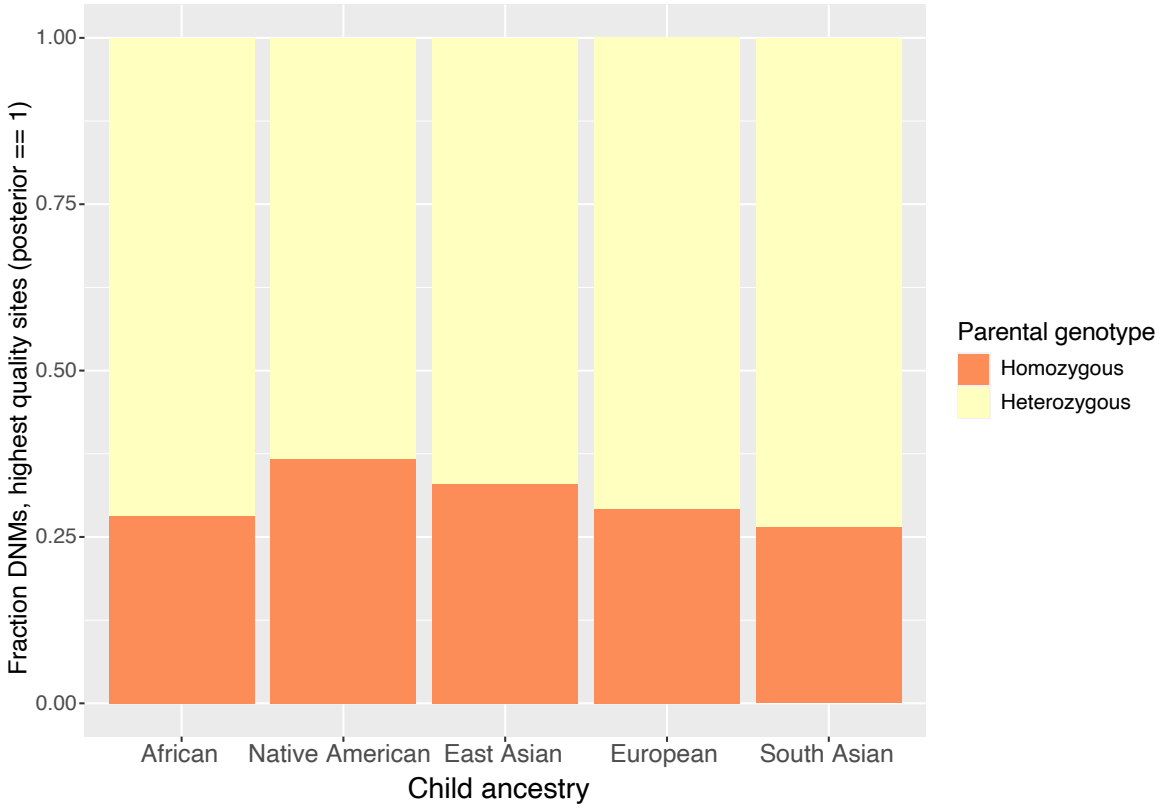
Supplementary figure 6: No significant association of socioeconomic status on number of DNMs. Socioeconomic status was estimated based on annual household income in US dollars. 1=less than $20k; 2=$21-35k; 3=$36-50k; 4=$51-65k; 5=$66-80k; 6=$81-100k; 7=$101-130k; 8=$131-160k; 9=over $161k.



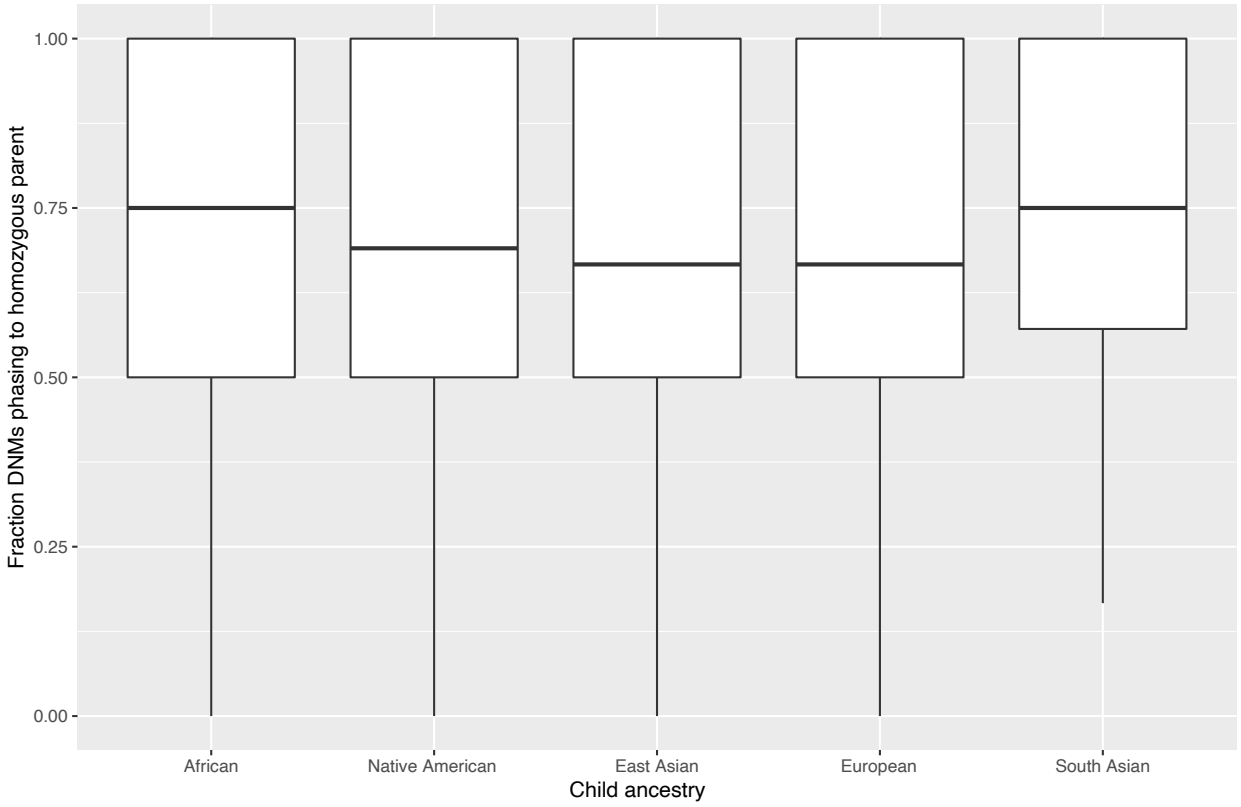
Supplementary figure 7: African ancestry is associated with significantly more expansion and deletion DNMs than other ancestries ($P = 1.04*10^{-36}$, $1.33*10^{-14}$, respectively; Wilcoxon rank sum test).



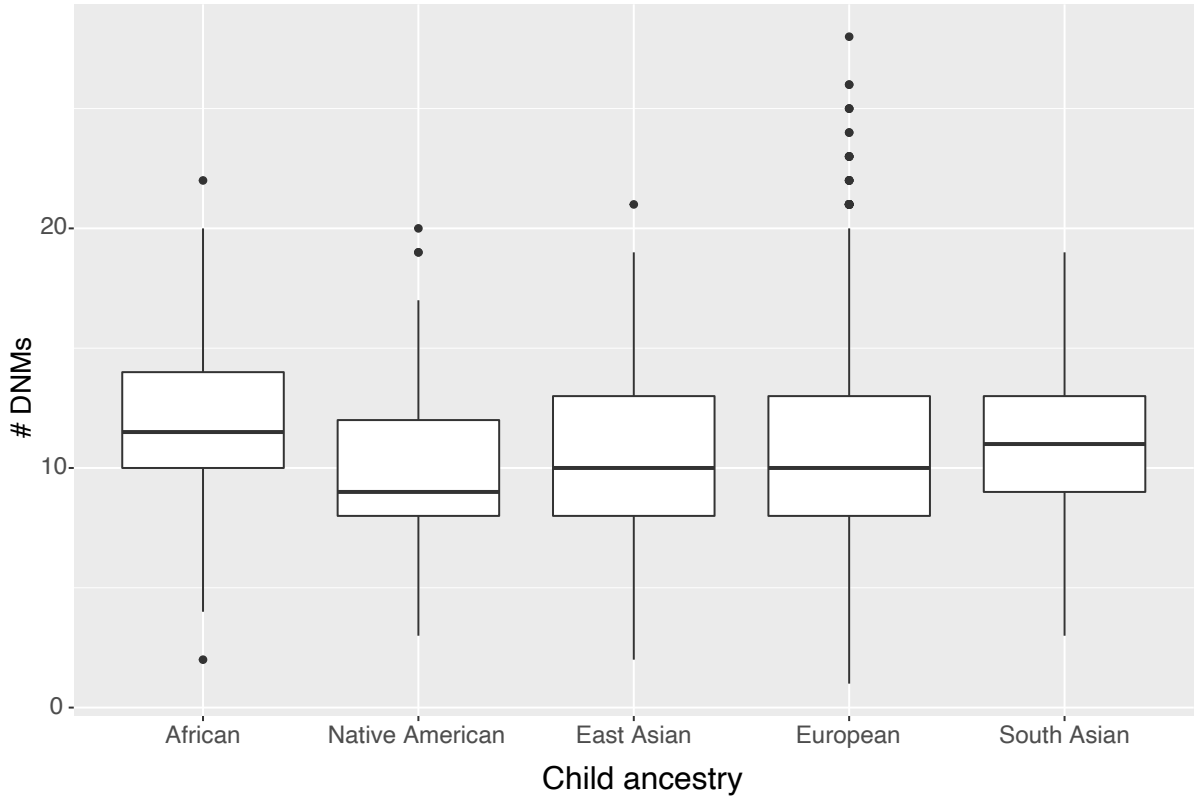
Supplementary figure 8: African ancestry is associated with significantly more DNMs at homopolymer, dinucleotide, and trinucleotide STRs ($P = 3.87 * 10^{-38}$, $1.88*10^{-7}$, $1.35*10^{-2}$, 0.051, and 1 for repeat units of 1, 2, 3, 4, and ≥ 5, respectively; Wilcoxon Rank Sum test with Bonferroni correction).



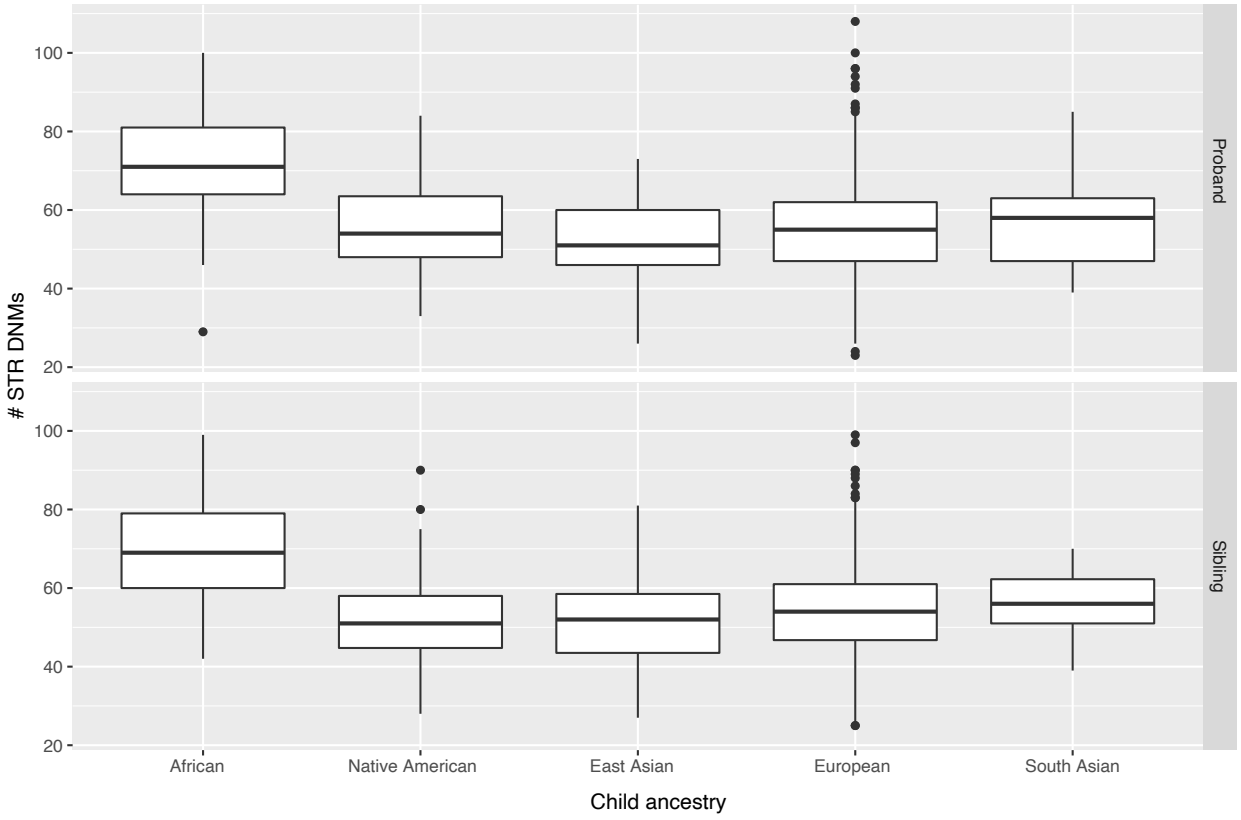
Supplementary figure 9: Sites of the highest confidence (posterior probability greater than 0.9999) phase more frequently to the heterozygous than to the homozygous parent. We grouped together mutations with a posterior probability of 1 at sites with one homozygous and one heterozygous parent across children of a given ancestry, then calculated the fraction of these mutations inherited from either parent as a function of genotype.



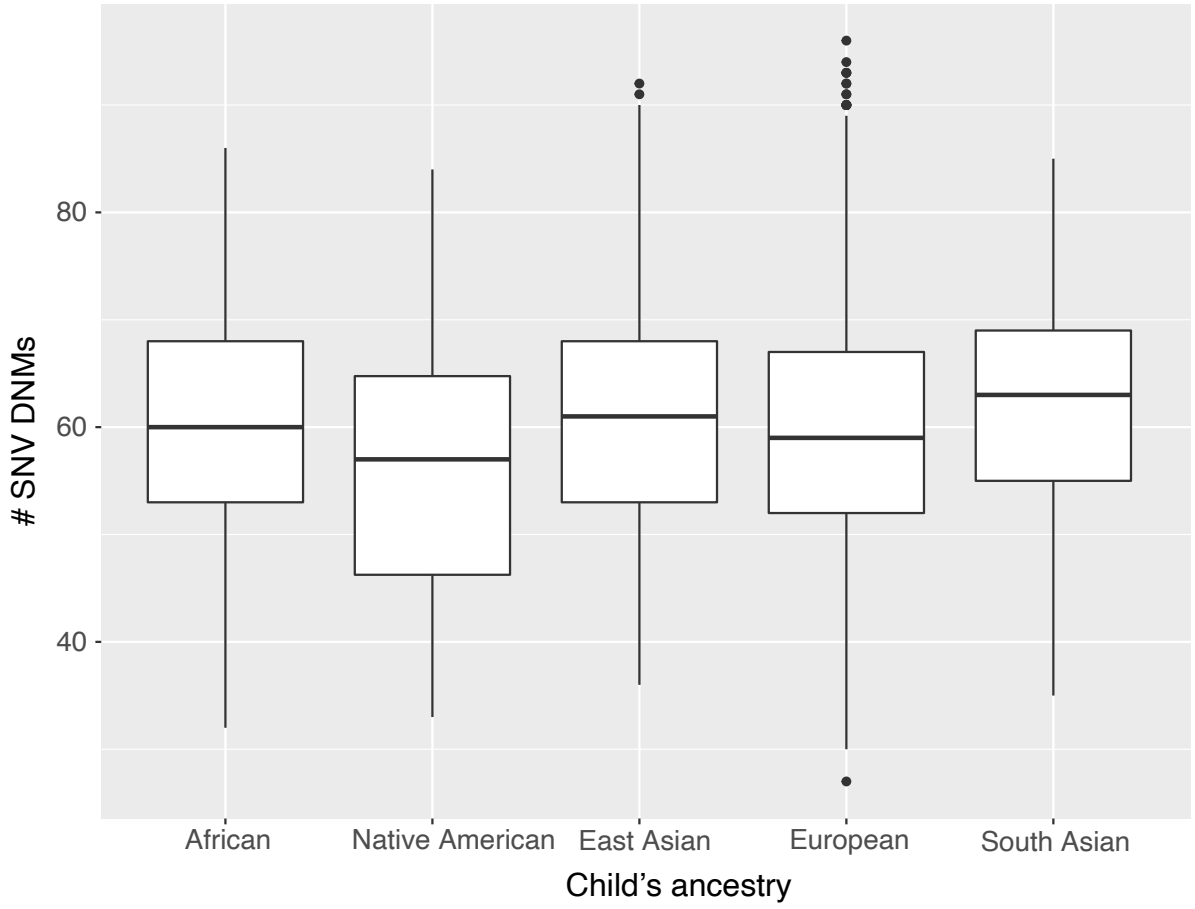
Supplementary figure 10: Filtering for DNMs that change allele length by a single repeat unit and for which the *de novo* allele is enclosed by 10 or more reads helps reduces the tendency of DNMs to phase more often to the chromosome inherited from a homozygous parent when considering mutations with one homozygous and one heterozygous parent.



Supplementary figure 11: Significantly more STR DNMs associated with African ancestry after filtering for high quality mutations with a maximum change in number of repeat units of 1 ($P = 8.00 * 10^{-5}$).



Supplementary figure 12: ASD diagnosis and African ancestry are both associated with significantly higher STR DNM rates, but these effects appear to be independent and additive with no significant interaction ($P > 0.28$ for all ancestries, Poisson linear model).



Supplementary figure 13: Number of SNV DNMs by ancestry. There is no significant difference in the number of SNV DNMs observed in children of African ancestry (P = 0.63).

VITA

Michael Goldberg was born in 1991 in Boston, MA. While in high school, he completed a senior thesis developing digital tools to study variance in shell morphometry between populations of *Homarus americanus* with Dr. Jelle Atema. He graduated from Brown University in 2013 with a ScB, *magna cum laude*, with a double concentration in biology (physical sciences track) and music. While at Brown, he completed an honors thesis co-advised by Drs. Sohini Ramachandran and Kate Smith entitled *Building and Implementing a Global Disease Outbreak Geodatabase*. After graduation, he spent a year in Montpellier, France working with Dr. Jean-François Guégan with a Fulbright Advanced Student fellowship, studying the biogeography of neglected tropical diseases and parasitology. From 2014-2017, he worked in Cambridge, MA for Foundation Medicine, Inc., a cancer diagnostics company for which he researched genomic biomarkers predictive of response to different cancer therapies. In 2017, he joined the Department of Genome Sciences at the University of Washington in pursuit of a PhD, ultimately joining the lab of Dr. Kelley Harris to study evolution of germline mutagenesis. After graduation, he will join the lab of Dr. Aaron Quinlan at the University of Utah's Department of Human Genetics as a postdoctoral fellow.