Great ape mutation spectra vary across the phylogeny and the genome due to distinct mutational processes that evolve at different rates

Michael E. Goldberg[1] and Kelley Harris[1,2]

[1]University of Washington Department of Genome Sciences; [2]Fred Hutchinson Cancer Center Computational Biology Division

ABSTRACT

Recent studies of hominoid variation have shown that mutation rates and spectra can evolve rapidly, contradicting the fixed molecular clock model. The relative mutation rates of three-base-pair motifs differ significantly among great ape species, suggesting the action of unknown modifiers of DNA replication fidelity. To illuminate the footprints of these hypothetical mutators, we measured mutation spectra of several functional compartments (such as late-replicating regions) that are likely targeted by localized mutational processes. Using genetic diversity from 88 great apes, we find that compartment-specific mutational signatures appear largely conserved between species. These signatures layer with species-specific signatures to create rich mutational portraits: for example, late-replicating regions in gorillas contain an identifiable mixture of a replication timing signature and a gorilla-specific signature. Our results suggest that cis-acting mutational modifiers are highly conserved between species and trans-acting modifiers are driving rapid mutation spectrum evolution.

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 (Alexandrov et al., 2013; Sima and Gilbert, 2014). Some regions of the genome accumulate mutations faster than others, such as DNA that replicates late in the cell cycle and motifs that are modified epigenetically (Koren et al., 2012; Liu et al., 2013; Polak et al., 2015; Schuster-Böckler and Lehner, 2012; Sima and Gilbert, 2014; Stamatoyannopoulos et al., 2009). Such mutation rate differences can confound efforts to infer patterns of purifying selection and background selection, as regions with elevated mutation rates may be more evolutionarily constrained than their diversity levels might suggest. As a result, understanding how mutation rate varies across the genome is an important prerequisite for identifying modes and targets of natural selection (Hellmann et al., 2003; Keightley et al., 2011; Kulathinal et al., 2008). Understanding the mutational landscape is similarly essential for predicting rates of deleterious *de novo* mutations in clinically relevant disease genes (Michaelson et al., 2012; Veltman and Brunner, 2012).

Many functional features of the genome appear to influence its mutation rate, or at least have landscapes of variation that correlate with the landscape of mutation rate variability (Ananda et al., 2011; Li and Luscombe, 2018). However, a large component of the variance of the mutational landscape is cryptic, meaning not associated with any known sequence motifs or

functional genomic features (Hodgkinson and Eyre-Walker, 2011; Johnson and Hellmann, 2011; Terekhanova et al., 2017). This cryptic variation is conserved over relatively long timescales, being highly similar between human and macaque, which diverged ~25mya (Tyekucheva et al., 2008). This suggests that there exist unknown genetic modifiers of the mutational landscape that are both widespread and functionally important.

In some systems, it is possible to estimate the number of unknown mutation rate modifiers by studying the spectrum of sequence contexts where mutations most commonly occur. Mutation spectrum analysis was pioneered in the setting of cancer, where many highly penetrant mutagens cause mutations in specific sequence contexts (Alexandrov et al., 2013). For example, tumors that replicate their DNA with a defective polymerase epsilon accumulate high rates of TCT>TAT and TCG>TTG mutations (Alexandrov et al., 2013; Shinbrot et al., 2014). 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 (Agarwal and Przeworski, 2019). The consistent presence of this mutational signature across all late-replicating regions suggests that the effects of replication timing on mutation rate are mediated by a single, currently unknown, mechanism.

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 (Goodman, 1985; Scally and Durbin, 2012). 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 (Harris and Pritchard, 2017). Even more surprisingly, closely related human populations have distinctive mutation spectra that provide enough information to classify individuals into continental ancestry groups (Harris and Pritchard, 2017). Over a period of just 10,000 to 20,000 years, Europeans experienced a temporary pulse of mutagenic activity that more than doubled the rate of TCC>TTC mutations (Harris, 2015; Speidel et al., 2019).

A previous study observed that the cryptic component of mutation landscape variation appears to evolve faster over time than components that are correlated with functional features of the genome (Terekhanova et al., 2017). This motivated us to take a closer look at the differences between the mutation spectra of great ape species; in particular, to measure whether these differences were concentrated in specific genomic regions. Although this approach is not designed to reveal the root causes of any species-specific signatures, it is well powered to distinguish whether these causes are *cis*-acting or *trans*-acting mutators. A genetic variant in a DNA repair enzyme might modify the mutation spectrum in *trans* by broadly changing the efficacy of DNA repair across the entire genome. Since the rate and spectrum of germline mutations also depend on the age of the parents at conception (Jónsson et al., 2017; Wong et al., 2016), a variant affecting reproductive life history could also effectively act in *trans* as a mutator allele. In contrast, a genetic variant that alters local chromatin state might modify the mutation spectrum in *cis* by changing the accessibility of nearby DNA to damage as well as repair, but not on other chromosomes.

If *cis*-acting mutators have been evolving more quickly than *trans*-acting mutators, we should expect to see large differences between genomic regions in the differentiation of ape species. Conversely, if *trans*-acting mutators emerge more often, we should expect all regions of the genome to show similar pictures of species differentiation. By assessing the fits of *cis*- and *trans*-acting models to ape mutational data, we aim to narrow down the long list of potential genetic and environmental mutators to a smaller set of candidates that fit the spatial profile of the variability between ape genomes.

RESULTS

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 (Harris and Pritchard, 2017; Prado-Martinez et al., 2013). We first sought to recapitulate these results and measure for the first time how the differences between species compare to differences within species.

To minimize the effects of natural selection on our mutation spectrum ascertainment, we defined set of genomic regions, collectively called a "compartment", characterized as non-conserved and non-repetitive (NCNR). The NCNR compartment consisted of 1.28Gb of the non-repetitive (annotated by RepeatMasker), non-coding human genome annotated as significantly conserved ($p<0.05$ in the PhastCons 44-way primate alignment), and excluding CpG islands. In these compartments, we computed the fractions 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 (see Methods).

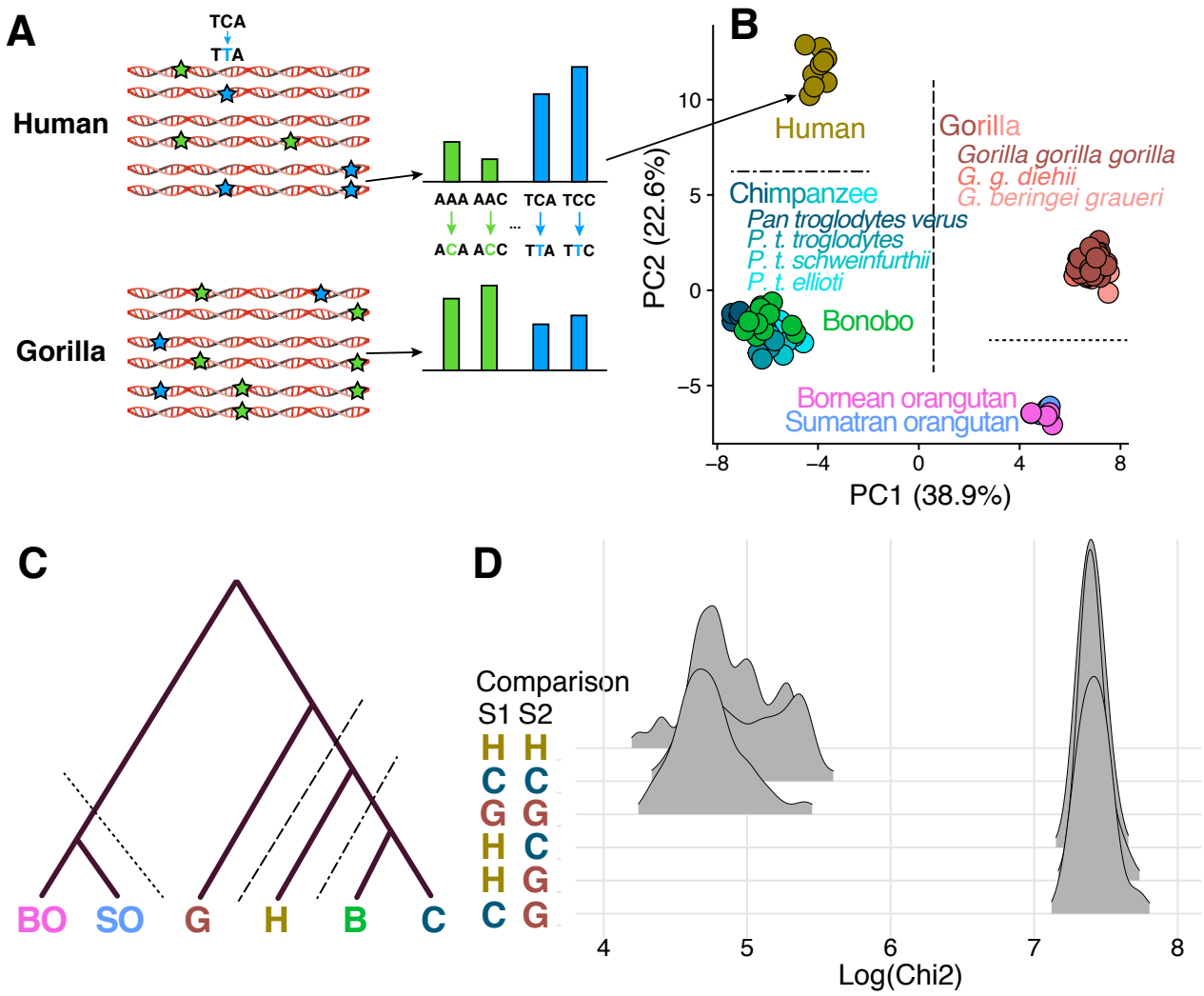

Figure 1: Covariance of species-specific and replication timing mutation spectra in great apes.

A. SNVs segregating within a species are 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, or 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. The positioning of the clusters recapitulates the great ape phylogeny. PC1 separates humans, chimpanzees, and bonobos from gorillas and orangutans, while PC2 separates humans from chimpanzees and bonobos and gorillas from orangutans.
D. Mutation spectra are more similar between individuals of the same species than between individuals from different species. We plotted the Chi-square distances between triplet mutation spectra of all possible pairs of individuals in the GAGP within and between species.

A principal component analysis (PCA) of the individual mutation fractions shows clustering of individuals by species in a manner that recapitulates phylogeny; this pattern reveals distinct and rapidly evolving species-specific mutation spectra (Figure 1B,C). Closely related humans,

135 chimpanzees (*Pan troglodytes*), and bonobos (*Pan paniscus*) separate from more distantly related
136 gorillas (*Gorilla gorilla*) and Sumatran and Bornean orangutans (*Pongo abelii* and *Pongo*
137 *pygmaeus*, respectively) along principal component 1 (PC1). Humans separate from chimpanzees
138 and bonobos along PC2; gorillas separate from the two orangutan species along the same axis.
139 These results are robust to repeated subsampling of the number of individuals to match among
140 species (Figure 1-supplemental figure 1).
141
142 To quantify these mutation spectrum differences further, we computed Chi-square distances
143 between the spectra of individual genomes and found that interspecific differences exceeded
144 conspecific differences (Figure 1D, Figure 1-supplemental figure 2). We even observed mutation
145 spectrum differences among chimpanzee subspecies: Western chimpanzees (*P. troglodytes*
146 *verus*) separate from bonobos along PC1, while other subspecies only separate along PC2. This
147 pattern, even more visible in a PCA of chimpanzees alone, implies an acceleration of mutation
148 spectrum divergence along the Western chimp lineage (Prado-Martinez et al., 2013; Sudmant et
149 al., 2013) (Figure 1- supplemental figure 3). Gorilla subspecies and human populations exhibit
150 more subtle mutation spectrum differences that are not visible when spectra are projected onto
151 the principal axes of ape variation.
152
153 **A mutational signature of DNA replication timing is conserved among great apes.**
154 Differences in replication timing explain a substantial portion of the variation in somatic and
155 germline mutation rate across the genome (Koren et al., 2012; Stamatoyannopoulos et al., 2009).
156 Compared to regions that replicate early in S phase, late replicating regions tend to have a higher
157 overall mutation rate, and in humans they particularly harbor a higher rate of A>T and C>A
158 mutations (Agarwal and Przeworski, 2019). The established correlation between late replication
159 timing and elevated mutation rate implies that replication timing QTLs (rtQTLs) may be examples
160 of *cis*-acting mutation spectrum modifiers. We analyzed late- and early-replicating compartments
161 of the genome to determine whether replication timing had a similar effect on the mutation
162 spectrum across great apes.
163

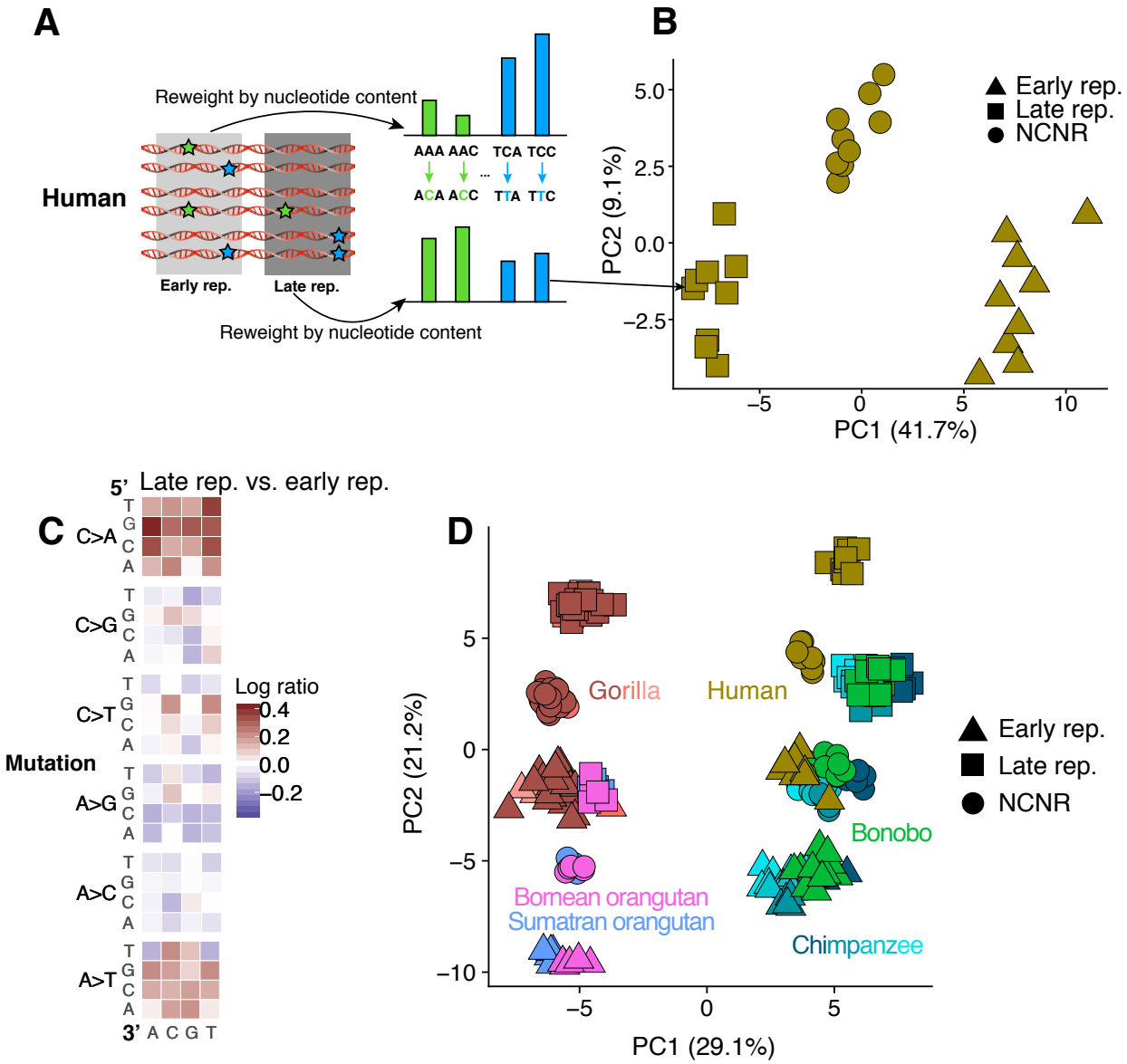

Figure 2

A. We calculated separate mutation spectra for each individual in compartments that replicate early or late in S1 phase and re-weighted spectra by the nucleotide content of the respective compartment.
B. Each point in this PCA represents the mutation spectrum from a single individual's NCNR, early replicating, or late replicating compartment. Clustering by compartment imply differences in mutation spectra based on replication timing.
C. A heatmap of the log ratios of triplet mutation fractions in humans 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.

D. Mutation spectra from late and early replication timing and NCNR compartments from each individual in the GAGP cluster and position along orthogonal "phylogeny" and "replication timing" axes. The mutation signature associated with late replication timing appears conserved among all great apes. Different shades of each species' color represents subspecies, as in Figure 1D.

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 (Figure 2A; Koren et al., 2012). We then generated two mutation spectra for each individual corresponding to the early and late replication timing compartments. We ran a PCA on a matrix including the individual mutation spectra of all humans in the GAGP for early and late replication timing compartments in addition to the NCNR compartment (Figure 2B, Figure 2-supplemental figure 1). 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 2-supplemental figure 2). 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.

To determine if the variation in mutation spectra due to differences in replication timing was conserved among species, we ran a PCA on a matrix of the early replicating, late replicating, and NCNR compartment mutation spectra of all individuals (Figure 2D). The separation along PC1 reflects phylogeny, while PC2 primarily separates early from late replication timing compartments. The direction of separation of early and late replication timing compartments is similar across all species, again implying distinct and conserved mutational mechanisms. 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 2 – supplemental figure 3). Our results show that late replication timing is associated with a conserved mutational signature across great apes. Moreover, the genomic landscape of replication timing is broadly conserved across species, biasing all genomes toward C>A and A>T mutations in orthologous late replicating compartments (Figure 2 – supplemental figure 4).

All 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 (Figure 2- supplemental figure 2). Late-replicating regions of chimp genomes have high loads of both a chimp-specific signature and the same late-replication signature that is found in human genomes. We see no evidence of any mutational signature unique to late-replicating chimp DNA that is not also found in early-replicating chimp

221 DNA or in late-replicating regions of other ape genomes. Furthermore, we see no evidence that
222 species-specific signatures have a rate or dosage that depends on replication timing.
223
224 Using published annotations of the human genome, we defined several more overlapping
225 functional compartments to test for the presence of *cis*-acting mutation spectrum modifiers. We
226 used RepeatMasker to delineate repetitive vs. non-repetitive DNA and used ENCODE chromHMM
227 output (the intersection of heterochromatic regions in nine cell types) to annotate several types
228 of heterochromatin (Ernst and Kellis, 2012). Another compartment we annotated consists of
229 ancient repeats, which have the potential to mutate differently from higher complexity DNA via
230 several mechanisms, including the formation of non-B-DNA secondary structures and the editing
231 activity of antiviral enzymes such as APOBECs (Bacolla et al., 2004; Guiblet et al., 2018; Harris et
232 al., 2002).
233
234 We ran a PCA for each species including the individual mutation spectra of all eight
235 compartments and observed similar pattern of compartment separation in each species. In all
236 cases, the first principal component separates compartments along a replication timing gradient,
237 with the percentage of variance explained ranging from 23.3% to 34.1%. PC2 separates by
238 repetitive content (PVE 8.2% to 13.5%), and PC3 shows an ERV-distinct signature (PVE 4.0% to
239 8.7%; see Figure 3A-F, Figure 3 supplement 1). The similarities of the independent PCAs across
240 all species of great apes imply conservation of the *cis*-regulated mutational signatures associated
241 with repetitive content, methylation, and replication timing. Each compartment shows a similar
242 degree of separation between species, with high degrees of correlation between the positioning
243 of compartments in different species' PCAs (Figure 3 supplement 2). This suggests that species-
244 specific signatures fit the profile of *trans*-acting mutation modifiers that act relatively uniformly
245 across the genome. These species-specific signatures tend to cause more variance in mutation
246 rate than compartment-specific signatures do—for example, the greatest axis of variance in the
247 PCA in Figure 1H captures species signatures, while replication timing separates on the second-
248 greatest axis (29.1% and 21.2% of variance explained, respectively).

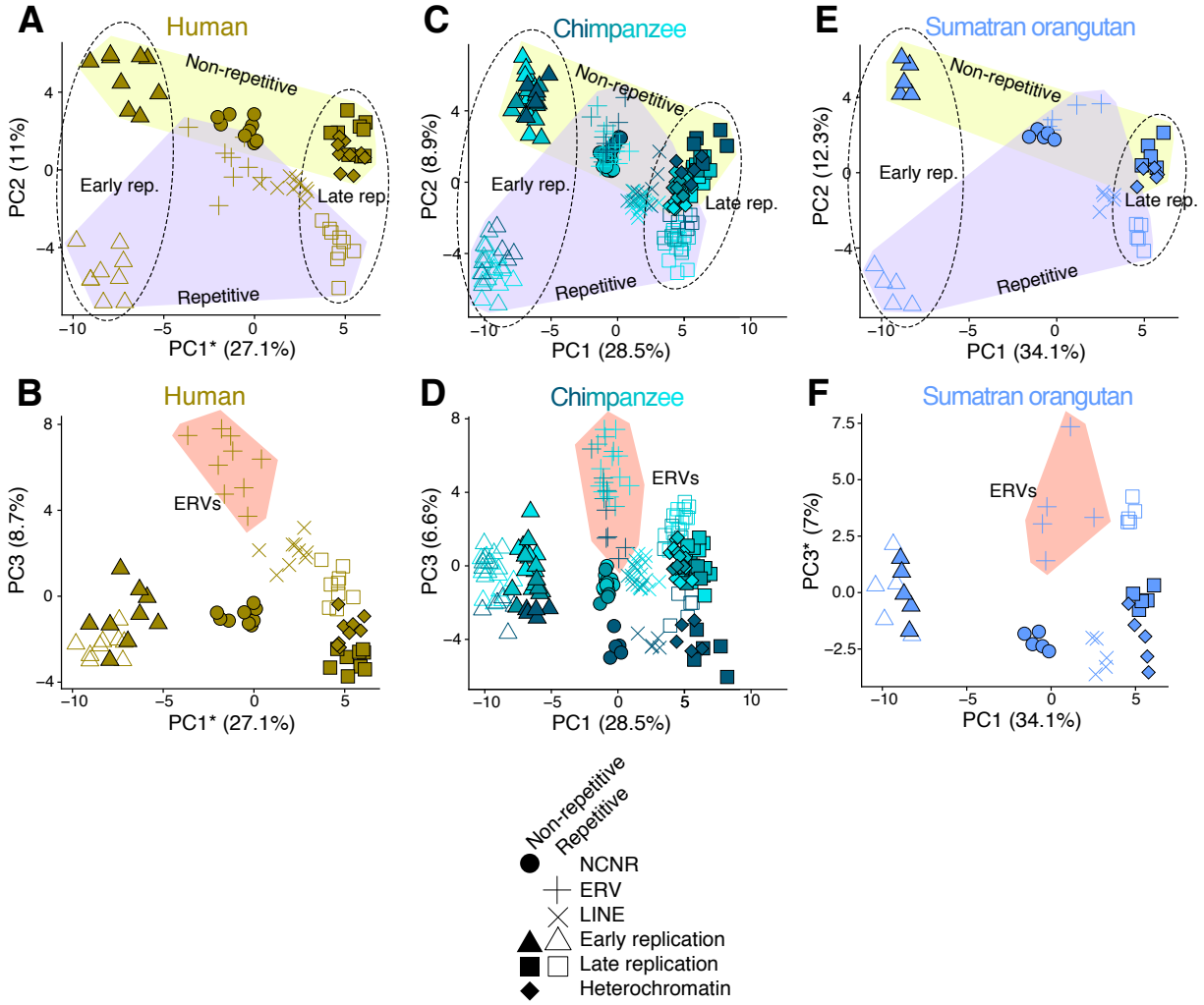

Figure 3: Conserved axes of mutation spectrum variance among great apes

A-F. We defined eight overlapping functional compartments to test for presence of 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.

We found only two examples of compartments that violate these general trends: maternal mutation hotspots and CpG islands. Maternal mutation hotspots are genomic regions that are enriched for *de novo* C>G mutations specifically from the maternal lineage; the rate of these mutations is also correlated with maternal age (Jónsson et al., 2017). These hotspots exist in chimpanzees and, to a lesser extent, gorillas, but their signal is largely absent from orangutans. Our mutation PCAs similarly show that human genomes have higher levels of a compartment-specific mutational signature in these regions (Figure 3 supplement 3). The separation of NCNR

268 to maternal mutation hotspot mutation spectra compartments decays with phylogenetic
269 distance from humans, recapitulating findings from Jonsson et al (2017). This compartment
270 demonstrates evolution of a *cis*-acting mutational mechanism.
271
272 CpG islands are 200-2000bp long genomic regions enriched for CpG dinucleotides (Gardiner-
273 Garden and Frommer, 1987; Larsen et al., 1992). Outside of these regions, CpGs are often
274 methylated to 5-methylcytosine, which mutate to TpG at a rate ten times higher than
275 unmethylated CpGs. In contrast, CpG islands are hypomethylated and are often situated in
276 conserved 5' promoters and genic regions. We hypothesized that, due to their lack of CpG>TpG
277 mutations and overall conservation, a compartment containing CpG islands would demonstrate
278 a contrasting mutation spectrum relative to that of the NCNR compartment. A PCA of individual
279 mutation spectra from the NCNR and a CpG compartment from all GAGP individuals
280 demonstrated that differentiation of the CpG island compartment exceeds the magnitude of
281 spectrum differentiation between species (Figure 3 supplement 4). This is unsurprising given the
282 unique mutational properties of CpG islands compared to the rest of the genome.
283
284 **Endogenous retroviruses carry a distinct mutational signature conserved across great apes.**
285 ERVs are a class of repetitive, transposable DNA elements that duplicate themselves in a copy
286 and paste manner. The act of duplication into new regions of the genome can disrupt function;
287 for example, integration into the coding region of a gene could result in a complete knock-out of
288 gene function. Therefore, ERV activity is restricted by a number of known mechanisms, including
289 hypermethylation, inhibition of integration, and hypermutation.
290
291 In light of these mechanisms that target ERVs, we were intrigued by the fact that ERVs separated
292 from other genomic compartments along the third principal component of our mutation
293 spectrum analysis (4-8% variance explained). ERVs bear an excess load of a unique mutational
294 signature that appears to be largely conserved among great apes (Figure 2B,D,F) and is previously
295 undescribed, to our knowledge.
296
297 To determine whether any component of the ERV signature could be caused by high rates of
298 methylation and heterochromatization, we directly compared the mutation spectrum of the ERV
299 compartment to that of nonrepetitive heterochromatin. We generated the log ratio of the 96
300 mutation types between the ERV and the nonrepetitive heterochromatin compartment for each
301 species and found an enrichment for CpG C>G mutations and a depletion of TAA>TTA mutations
302 in ERVs that appears conserved in all species other than *Pongo pygmaeus* (Figure 4A). This
303 comparison shows that ERVs' high rate of CpG>CpT transitions is likely caused by their
304 heterochromatic status, but that heterochromatinization cannot explain the other components
305 of the ERV signature. Furthermore, we determined that differences in 7-mer nucleotide content
306 between the two compartments explained some, but not all, of the ERV-specific enrichment for
307 CG>GG mutation types (Figure 4, supplemental figure 1).

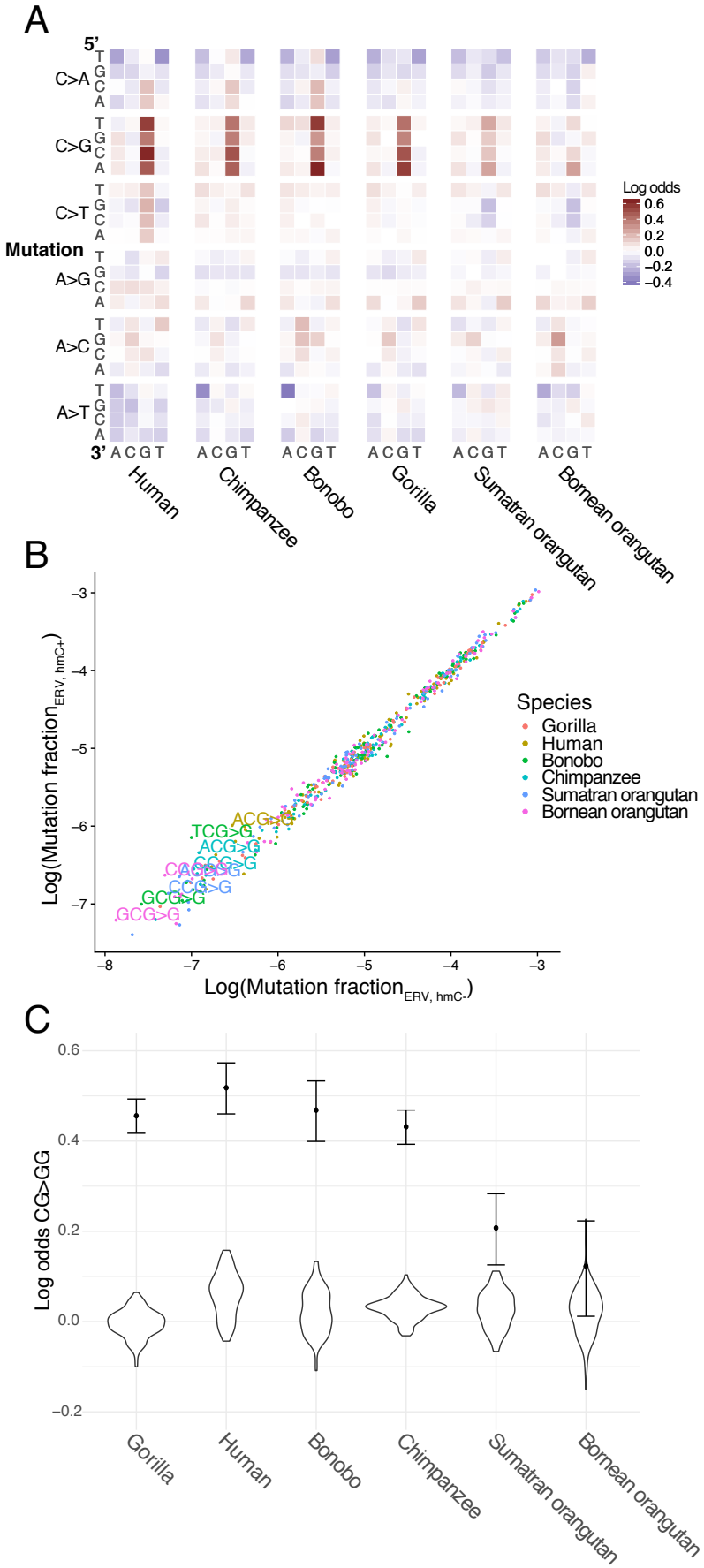
A
5'
3'
Mutation
C>A
C>G
C>T
A>G
A>C
A>T
T G C A
A C G T
Log odds
0.6
0.4
0.2
0.0
−0.2
−0.4
Human
Chimpanzee
Bonobo
Gorilla
Sumatran orangutan
Bornean orangutan
B
Log(Mutation fraction$_{ERV, hmC+}$)
Log(Mutation fraction$_{ERV, hmC-}$)
−3
−4
−5
−6
−7
−8
Species
Gorilla
Human
Bonobo
Chimpanzee
Sumatran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

Figure 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.

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 (Supek et al., 2014). 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 (Yu et al., 2012). 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, Figure 4 supplement 2,3). 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, Figure 4-supplemental 4).

DISCUSSION

Despite considerable documented evidence of mutation rate variation across genomic space and phylogenetic time, little was previously known about the covariation of mutation rates or spectra along these 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 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 chimps and gorillas and nearly nonexistent in orangutans. Nevertheless, or results show that mutation spectrum divergence between ape species is mostly driven by processes that act promiscuously across the genome.

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, which are a phylogenetic outgroup to other chimpanzees, are distinctive in several ways, with 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 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 (Aggarwala and Voight, 2016; Carlson et al., 2018). 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 replication timing quantitative trait loci (rtQTLs) (Koren et al., 2012), but our results suggest that the coarse shape of the replication timing landscape is extremely stable across the great ape clade. Since replication timing is highly correlated with topologically-associated domain (TAD) structure (Pope et al., 2014), our results imply that human TAD structure can largely be imputed into other apes, a useful insight given the expense of direct Hi-C measurements. The genomic distribution of the replication timing signature could even be leveraged to estimate the extent of TAD variation within and between species.

In the majority of the genome where the replication timing landscape appears to be constrained, stable, and consistent in its mutational bias, we predict that the dosage of the replication timing signature is likely to behave as a more reliable molecular clock than the full spectrum of mutations that contains fast-evolving components and is known to vary in rate between lineages (Moorjani et al., 2016b, 2016a; Scally and Durbin, 2012). Using the spectrum loadings of the compartment signatures and species-specific signatures inferred here, we have enough information to estimate the number of mutations in an ape genome that were caused specifically by replication-associated signature. Given the apparent stability of this signature across ape species, this subset of mutations could be leveraged to estimate divergence times, population size changes, and other demographic parameters more accurately than can be done using mutations that accumulate preferentially in certain species.

If any mutational signatures accumulate at stable rates outside the clade of great apes, they could also prove useful for phylogenetic inference in larger sets of species as well. This has the potential to help resolve disputes about evolution and phylogeny in large clades where evolutionary inference is challenging. Although spatial and temporal variation of the mutation rate and spectrum both serve to distort the molecular clock model and complicate population genetic inference, their pattern of covariation suggests that spatially varying mutational signatures may provide the reliable molecular clock that has eluded evolutionary biologists for years.

ACKNOWLEDGEMENTS

We are grateful for support from NIGMS Training Grant T-32 GM081062 as well as the following grants awarded to K.H: NIGMS grant 1R35GM133428-01, a Burroughs Wellcome Career Award at the Scientific Interface, a Pew Biomedical Scholarship, a Searle Scholarship, and a Sloan Research Fellowship. 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.

BIBLIOGRAPHY

Agarwal I, Przeworski M. 2019. 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* **116**:17916–17924. doi:10.1073/pnas.1900714116

Aggarwala V, Voight BF. 2016. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics* **48**:349–355. doi:10.1038/ng.3511

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Imielinsk M, Jäger N, Jones DTW, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague

441 JW, Totoki Y, Tutt ANJ, Valdés-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A,
442 Waddell N, Yates LR, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer
443 Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi J, Futreal PA,
444 McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T,
445 Pfister SM, Campbell PJ, Stratton MR. 2013. Signatures of mutational processes in
446 human cancer. *Nature* **500**:415–421. doi:10.1038/nature12477
447 Ananda G, Chiaromonte F, Makova KD. 2011. A genome-wide view of mutation rate co-
448 variation using multivariate analyses. *Genome Biol* **12**:R27. doi:10.1186/gb-2011-12-3-
449 r27
450 Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova N, Abeysinghe SS, O'Connell CD,
451 Cooper DN, Wells RD. 2004. Breakpoints of gross deletions coincide with non-B DNA
452 conformations. *Proceedings of the National Academy of Sciences* **101**:14162–14167.
453 doi:10.1073/pnas.0405974101
454 Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, The BRIDGES Consortium, Myers RM,
455 Boehnke M, Kang HM, Scott LJ, Li JZ, Zöllner S. 2018. Extremely rare variants reveal
456 patterns of germline mutation rate heterogeneity in humans. *Nature Communications* **9**.
457 doi:10.1038/s41467-018-05936-5
458 Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and
459 characterization. *Nat Methods* **9**:215–216. doi:10.1038/nmeth.1906
460 Gardiner-Garden M, Frommer M. 1987. CpG Islands in vertebrate genomes. *Journal of*
461 *Molecular Biology* **196**:261–282. doi:10.1016/0022-2836(87)90689-9
462 Goodman M. 1985. Rates of molecular evolution: The hominoid slowdown. *Bioessays* **3**:9–14.
463 doi:10.1002/bies.950030104
464 Guiblet WM, Cremona MA, Cechova M, Harris RS, Kejnovská I, Kejnovsky E, Eckert K,
465 Chiaromonte F, Makova KD. 2018. Long-read sequencing technology indicates genome-
466 wide effects of non-B DNA on polymerization speed and error rate. *Genome Res*
467 **28**:1767–1778. doi:10.1101/gr.241257.118
468 Harris K. 2015. Evidence for recent, population-specific evolution of the human mutation rate.
469 *Proc Natl Acad Sci USA* **112**:3439–3444. doi:10.1073/pnas.1418652112
470 Harris K, Pritchard JK. 2017. Rapid evolution of the human mutation spectrum. *eLife* **6**:e24284.
471 doi:10.7554/eLife.24284
472 Harris RS, Petersen-Mahrt SK, Neuberger MS. 2002. RNA Editing Enzyme APOBEC1 and Some of
473 Its Homologs Can Act as DNA Mutators. *Molecular Cell* **10**:1247–1253.
474 doi:10.1016/S1097-2765(02)00742-6
475 Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. 2003. A Neutral Explanation for the
476 Correlation of Diversity with Recombination Rates in Humans. *The American Journal of*
477 *Human Genetics* **72**:1527–1535. doi:10.1086/375657
478 Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian
479 genomes. *Nature Reviews Genetics* **12**:756–766. doi:10.1038/nrg3098
480 Johnson PLF, Hellmann I. 2011. Mutation Rate Distribution Inferred from Coincident SNPs and
481 Coincident Substitutions. *Genome Biology and Evolution* **3**:842–850.
482 doi:10.1093/gbe/evr044
483 Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson
484 KE, Eggertsson HP, Gudjonsson SA, Ward LD, Arnadottir GA, Helgason EA, Helgason H,

485 Gylfason A, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Rafnar T, Frigge M, Stacey SN, Th.
486 Magnusson O, Thorsteinsdottir U, Masson G, Kong A, Halldorsson BV, Helgason A,
487 Gudbjartsson DF, Stefansson K. 2017. Parental influence on human germline de novo
488 mutations in 1,548 trios from Iceland. *Nature* **549**:519–522. doi:10.1038/nature24018
489 Keightley PD, Eöry L, Halligan DL, Kirkpatrick M. 2011. Inference of Mutation Parameters and
490 Selective Constraint in Mammalian Coding Sequences by Approximate Bayesian
491 Computation. *Genetics* **187**:1153–1161. doi:10.1534/genetics.110.124073
492 Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012. Differential
493 Relationship of DNA Replication Timing to Different Forms of Human Mutation and
494 Variation. *The American Journal of Human Genetics* **91**:1033–1040.
495 doi:10.1016/j.ajhg.2012.10.018
496 Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MAF. 2008. Fine-scale mapping of
497 recombination rate in Drosophila refines its correlation to diversity and divergence.
498 *Proceedings of the National Academy of Sciences* **105**:10051–10056.
499 doi:10.1073/pnas.0801848105
500 Larsen F, Gundersen G, Lopez R, Prydz H. 1992. CpG islands as gene markers in the human
501 genome. *Genomics* **13**:1095–1107. doi:10.1016/0888-7543(92)90024-M
502 Li C, Luscombe NM. 2018. Nucleosome positioning stability is a significant modulator of
503 germline mutation rate variation across the human genome (preprint). Genomics.
504 doi:10.1101/494914
505 Liu L, De S, Michor F. 2013. DNA replication timing and higher-order nuclear organization
506 determine single-nucleotide substitution patterns in cancer genomes. *Nature*
507 *Communications* **4**. doi:10.1038/ncomms2502
508 Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A,
509 Wu W, Corominas R, Peoples Á, Koren A, Gore A, Kang S, Lin GN, Estabillo J, Gadomski T,
510 Singh B, Zhang K, Akshoomoff N, Corsello C, McCarroll S, Iakoucheva LM, Li Y, Wang J,
511 Sebat J. 2012. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo
512 Germline Mutation. *Cell* **151**:1431–1442. doi:10.1016/j.cell.2012.11.019
513 Moorjani P, Amorim CEG, Arndt PF, Przeworski M. 2016a. Variation in the molecular clock of
514 primates. *Proceedings of the National Academy of Sciences* **113**:10607–10612.
515 doi:10.1073/pnas.1600374113
516 Moorjani P, Gao Z, Przeworski M. 2016b. Human Germline Mutation and the Erratic
517 Evolutionary Clock. *PLOS Biology* **14**:e2000744. doi:10.1371/journal.pbio.2000744
518 Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E,
519 Vlahoviček K, Stamatoyannopoulos JA, Sunyaev SR. 2015. Cell-of-origin chromatin
520 organization shapes the mutational landscape of cancer. *Nature* **518**:360–364.
521 doi:10.1038/nature14221
522 Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK,
523 Thurman RE, Cheng Y, Gülsoy G, Dennis JH, Snyder MP, Stamatoyannopoulos JA, Taylor
524 J, Hardison RC, Kahveci T, Ren B, Gilbert DM. 2014. Topologically associating domains
525 are stable units of replication-timing regulation. *Nature* **515**:402–405.
526 doi:10.1038/nature13986
527 Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR,
528 Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H,

529 Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I,
530 Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C,
531 Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C,
532 Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A,
533 Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D,
534 Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC,
535 Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J,
536 Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD,
537 Hammer MF, Eichler EE, Marques-Bonet T. 2013. Great ape genetic diversity and
538 population history. *Nature* **499**:471–475. doi:10.1038/nature12228
539 Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding
540 human evolution. *Nat Rev Genet* **13**:745–753. doi:10.1038/nrg3295
541 Schuster-Böckler B, Lehner B. 2012. Chromatin organization is a major influence on regional
542 mutation rates in human cancer cells. *Nature* **488**:504–507. doi:10.1038/nature11273
543 Shinbrot E, Henninger EE, Weinhold N, Covington KR, Göksenin AY, Schultz N, Chao H,
544 Doddapaneni H, Muzny DM, Gibbs RA, Sander C, Pursell ZF, Wheeler DA. 2014.
545 Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific
546 mutation patterns and human origins of replication. *Genome Res* **24**:1740–1750.
547 doi:10.1101/gr.174789.114
548 Sima J, Gilbert DM. 2014. Complex correlations: replication timing and mutational landscapes
549 during cancer and genome evolution. *Current Opinion in Genetics & Development*
550 **25**:93–100. doi:10.1016/j.gde.2013.11.022
551 Speidel L, Forest M, Shi S, Myers S. 2019. A method for genome-wide genealogy estimation for
552 thousands of samples (preprint). Genetics. doi:10.1101/550558
553 Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009.
554 Human mutation rate associated with DNA replication timing. *Nat Genet* **41**:393–395.
555 doi:10.1038/ng.363
556 Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I,
557 Bontrop RE, Persengiev S, Antonacci F, Ventura M, Prado-Martinez J, Great Ape Genome
558 Project, Marques-Bonet T, Eichler EE. 2013. Evolution and diversity of copy number
559 variation in the great ape lineage. *Genome Research* **23**:1373–1382.
560 doi:10.1101/gr.158543.113
561 Supek F, Lehner B, Hajkova P, Warnecke T. 2014. Hydroxymethylated Cytosines Are Associated
562 with Elevated C to G Transversion Rates. *PLoS Genetics* **10**:e1004585.
563 doi:10.1371/journal.pgen.1004585
564 Terekhanova NV, Seplyarskiy VB, Soldatov RA, Bazykin GA. 2017. Evolution of local mutation
565 rate and its determinants. *Mol Biol Evol* msx060. doi:10.1093/molbev/msx060
566 Tyekucheva S, Makova KD, Karro JE, Hardison RC, Miller W, Chiaromonte F. 2008. Human-
567 macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol*
568 **9**:R76. doi:10.1186/gb-2008-9-4-r76
569 Veltman JA, Brunner HG. 2012. De novo mutations in human genetic disease. *Nat Rev Genet*
570 **13**:565–575. doi:10.1038/nrg3241

571 Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, Baker R, Thach DC, Iyer
572 RK, Vockley JG, Niederhuber JE. 2016. New observations on maternal age effect on
573 germline de novo mutations. *Nat Commun* **7**:1–10. doi:10.1038/ncomms10486
574 Yu M, Hon GC, Szulwach KE, Song C-X, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, Min J-H, Jin P,
575 Ren B, He C. 2012. Base-Resolution Analysis of 5-Hydroxymethylcytosine in the
576 Mammalian Genome. *Cell* **149**:1368–1380. doi:10.1016/j.cell.2012.04.027
577
578
579 METHODS
580 **SNV filtering**
581 We ascertained mutation spectra from a set of high-quality great ape SNVs that were
582 previously called and filtered by Prado-Martinez et al. (2013). For each species and compartment,
583 we collated the set of biallelic SNVs falling within the genomic segments that comprise that
584 compartment. Ancestral states were assigned using a parsimony approach. Briefly, a biallelic site
585 segregating within a genus (*Homo*, *Pan*, *Gorilla*, or *Pongo*) was polarized to the allele fixed in all
586 other genera. Sites segregating in multiple genera, sites with multiple fixed alleles, and sites with
587 more than two alleles in a single genus were excluded due to their inconsistency with the
588 assumptions of no balancing selection and only one mutation event per site. Singletons were also
589 excluded due to their higher likelihood of sequencing error. We used the inferred ancestral base
590 to classify 3' and 5' neighboring nucleotides. For 3-mer mutational analyses, we excluded SNVs
591 whose 3' and 5' neighboring nucleotides were an 'N' in the hg18 reference and SNVs
592 demonstrating evidence of recurrent mutation in the great ape lineage; we expanded this filter
593 to include the three 3' and 5' neighboring nucleotides for 7-mer mutational analyses. Finally, we
594 removed SNVs out of Hardy-Weinberg Equilibrium with excess heterozygosity (using an exact
595 test, $p<0.05$). Excess heterozygosity at a locus could indicate a cryptic segmental duplication with
596 a single, fixed mutation in a copy. We also excluded SNVs with ≥0.5 derived allele frequency to
597 avoid mutational classes with an elevated risk of ancestral state misidentification.
598
599 **Computing the mutation spectra of individuals and species**
600 The PCA analyses in this paper require the computation of mutation spectra from
601 individual genomes, whereas the complementary heat map analyses involve calculating
602 aggregate mutation spectra from larger samples. Each analysis employs the filtering system
603 described above and ultimately involves counting the number of filtered derived alleles, classified
604 by 3-mer context. However, slightly different calculation details are involved in the two cases.
605 The aggregate mutation spectrum of a species S is obtained from a set of counts
606 $C(m_1,S),\ldots,C(m_{96},S)$ where $m_1,\ldots,m_{96}$ are the 96 3-mer mutation type categories AAA>C, …, TCT>T.
607 The count $C(m_1,S)$ is the total number of SNVs segregating in species S that fall into the mutational
608 equivalence class m_1. To compare spectra across samples with different amounts of variation,
609 these mutation type counts are normalized to obtain a 96-dimensional histogram with frequency
610 categories summing to 1.
611 A mutation spectrum can be similarly calculated from a particular individual I as the
612 distribution of 3-mer mutation types across the derived alleles present in I's genome.
613 Homozygous derived alleles are given twice the weight of heterozygous alleles such that the

614 spectrum is the average of the spectra one would compute from the two phased haplotypes
615 making up *I*'s diploid genome.
616 When individual mutation spectra are computed in this way, two types of derived alleles
617 can contribute to spectrum covariance between individuals *I* and *J*. The first type are pairs of
618 derived alleles that occur at separate loci in *I* and *J* but belong to the same mutation equivalence
619 class. The second type are derived alleles inherited by both *I* and *J* from a common ancestor. To
620 maximize our power to detect mutation spectrum evolution and distinguish it from shared
621 genetic drift, we devised a randomization strategy to eliminate the second source of signal while
622 preserving the first.
623 This randomization strategy involves computing the mutation spectrum of individual *I*
624 from only a subset of the derived alleles present in *I*'s genome. If one copy of a particular derived
625 allele is present in *I*'s genome and has frequency $k/2N$ in the GAGP panel, the allele will be
626 counted toward *I*'s mutation spectrum with probability $1/k$. Conversely, this derived allele will be
627 counted toward the mutation spectrum of exactly one ape haplotype that carries it, with the
628 identity of that haplotype chosen uniformly at random.
629
630 **Comparing mutation spectra across genomic compartments**
631 Comparing mutation spectra between regions of the genome required accounting for
632 differences in compartment size and nucleotide content. Larger compartments naturally had
633 more mutations than smaller ones; it was therefore necessary to compare mutation fractions
634 rather than raw counts. Furthermore, differences in nucleotide content between compartments
635 could bias our comparison and calculation of local mutation rates. For example, a particular
636 compartment could have a relatively high count of AAA>ACA SNVs, but this high count might be
637 caused by the compartment having many occurrences of the triplet AAA, and therefore more
638 opportunities for an AAA>ACA to occur. Thus, we rescaled the number of mutations for each
639 compartment by the nucleotide content of the NRNC compartment before calculating fractions.
640 To calculate the rescaled rate $r(m)$ of mutation m_i: $\{m_1, \ldots, m_{96}\}$ corresponding to triplet t_t: $\{t_1, \ldots,$
641 $t_{32}\}$ and compartment C:

642 $$R_{i,C} = \#m_{i,C} * \frac{\#t_{t,NCNR}}{\#t_{t,C}}$$

643 $$r(m_i) = \frac{R_{i,C}}{\sum_j R_{j,C}}$$

644
645 We calculated triplet content of each compartment by sliding a 3bp window, 1bp at a time, across
646 each compartment. Edges of compartment segments and triplets with Ns were excluded.
647 Several statistical analyses comparing two different mutation spectra required count data
648 rather than frequency data (Chi-square tests, e.g.); we devised a slightly different rescaling
649 strategy to avoid artificially inflating the mutation counts. To calculate the rescaled count of
650 mutation m_i: $\{m_1, \ldots, m_{96}\}$ corresponding to triplet t_t: $\{t_1, \ldots, t_{32}\}$ and compartment C_1 in
651 preparation for comparison to compartment C_2, we scaled down the raw count of mutation m in
652 the compartment where m is more abundant, rather than scaling up the count of m in the
653 compartment where it is more abundant:

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

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

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 **chi-square 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 distributions chi-square value 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>AAACAC, … 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, AAAAAC, … TTTCTT} to that of compartment *C'*:

$$R_{7,i,C} = \sum_{l} \#m_{7,l,C} * \frac{\#s_{s,C'}}{\#s_{s,C}}$$

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

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 4,

690 supplemental figure 1 presents the log ratio of the rescaled ERV mutation and the non-rescaled
691 nonrepetitive heterochromatin mutation spectra.
692 We ran **correlation analyses** to quantify the similarities between the mutation spectrum
693 heatmaps between species. Each mutation spectrum heatmap comprised the log ratio between
694 each of the 96 mutation types between two compartments for a single species. To calculate the
695 similarity between heatmaps for two different species, we ran a Pearson correlation test on the
696 paired vectors of log odds.
697
698 **Compartments**
699 We defined compartments based on published annotations of genomic features. Each
700 compartment was a list of genomic segments in a bed file format. The following is a list of the
701 compartments used in our analyses and a short description of how we generated them.
702 **NCNR**: (non-conserved, non-repetitive) the entire hg18 genome, excluding repetitive
703 elements defined by repeatMasker, conserved regions in primates based on the phastCons 44-
704 way multi-species alignment, CpG islands from the UCSC genome browser, and coding exons
705 from refGene.
706 **ERVs**: all ERVs in repeatMasker run on hg18, excluding those classified specifically as
707 mammalian long-terminal repeats (MaLRs). MaLRs are believed to be largely inactive in great
708 apes, unlike ERVs which are still active in several species.
709 **LINEs**: all LINEs in repeatMasker run on hg18
710 **Heterochromatin**: the intersection of the heterochromatin domain called in the hg18
711 chromHMM run on 9 different cell types from ENCODE (Gm12878, H1 HESCs, HepG, Hmec,
712 Hsmm, Huvec, K562, Nhek, and Nhlf), minus repetitive elements defined in repeatMasker.
713 **Early/late replicating regions**: the genomic quartiles that replicate earliest and latest
714 during S phase were ascertained using replication timing data from Koren et al. 2012. In that
715 manuscript, the fine-scale replication timing of regions in the genome was determined by
716 sequencing human lymphoblastoid cell lines at S1/G phase; read depth over region in the genome
717 corresponded to its average relative replication timing. We calculated average replication timing
718 for non-overlapping 20kb window that included at least one measure of replication timing and
719 calculated replication timing quartiles. The earliest and latest quartiles were used as
720 compartments. **Early/late replicating, repetitive/non-repetitive regions** were the subsets of the
721 replication timing compartments that overlapped with or excluded repeatMasker-annotated
722 repeats, respectively.
723 **Human maternal mutation hotspots**: defined in Jonsson et al. (2017) as regions whose
724 *de novo* mutation rate strongly associates with maternal age, lifted over to hg18.
725 **ERVs ± 5hmC**: ERV compartment, split into segments that had or lacked evidence of ≥ 1
726 5hmC site, based on Tet-assisted bisulfite sequencing of human ESCs (Yu et al., 2012).
727
728 **Quality control analyses**
729 We ran a number of analyses to test the robustness of our methods and findings.
730 *Testing the robustness of PCA clustering to species representation*
731 We tested the robustness of the clustering of individuals by species in the NRNC PCA to
732 differences in number of individuals sequenced per species. We down-sampled the number of
733 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.

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 4, supplemental figure 4).

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 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 enrichment for CG>GG mutations comparing ERVs to nonrepetitive heterochromatin is significantly stronger than the enrichment in any of the bootstrapped, 'ERV-like' compartments.

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 4, supplemental figure 3).

SUPPLEMENTARY FIGURES

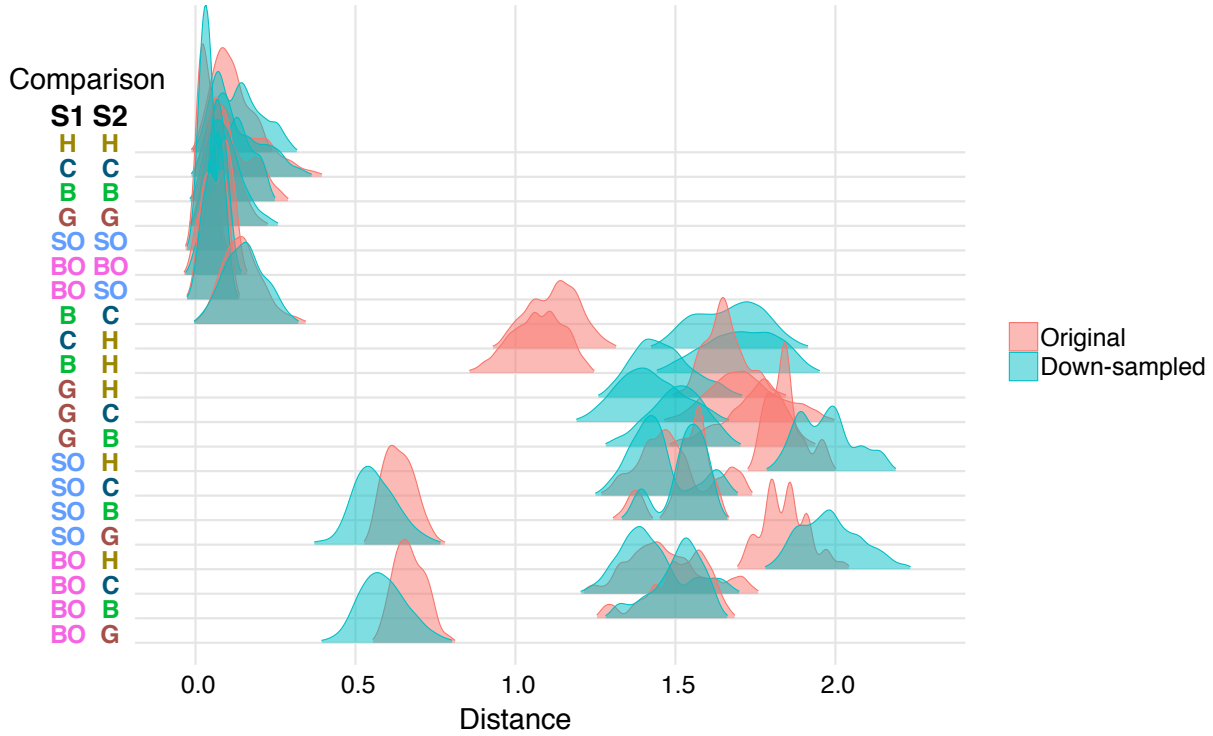

Figure 1 – Supplemental figure 1
NCNR PCA clustering is robust to differences in species representation. For each species (or both orangutan species, grouped together), we included a random down-sampling of nine individuals and generated a PCA. This method was bootstrapped 100 times. For each of the resulting PCAs, we calculated the Euclidean distance between each pair of individuals within and between species; the same distances were calculated for the original, non-bootstrapped PCA. The distributions of these distances are shown in the plot above; the blue and red distributions represent the bootstrapped and non-bootstrapped distances, respectively.

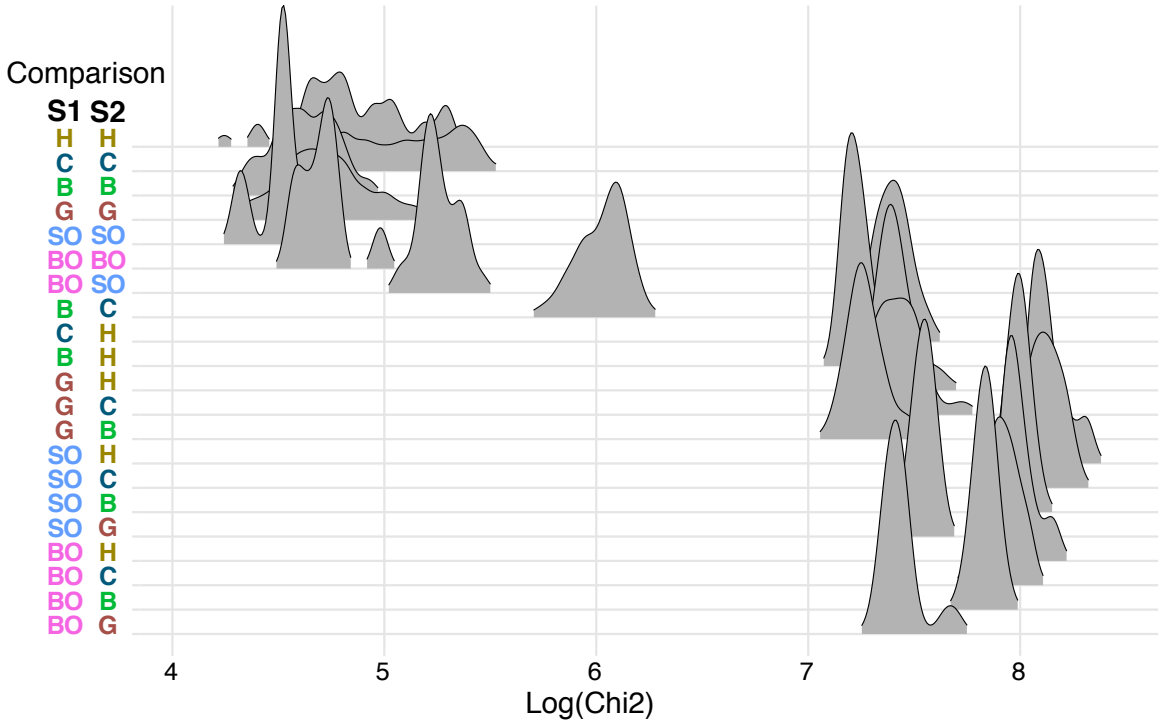

Figure 1, Supplemental figure 2

Mutation spectra are more similar between individuals of the same species than between individuals from different species. We plotted the Chi-square distances between triplet mutation spectra of all possible pairs of individuals in the GAGP within and between species.

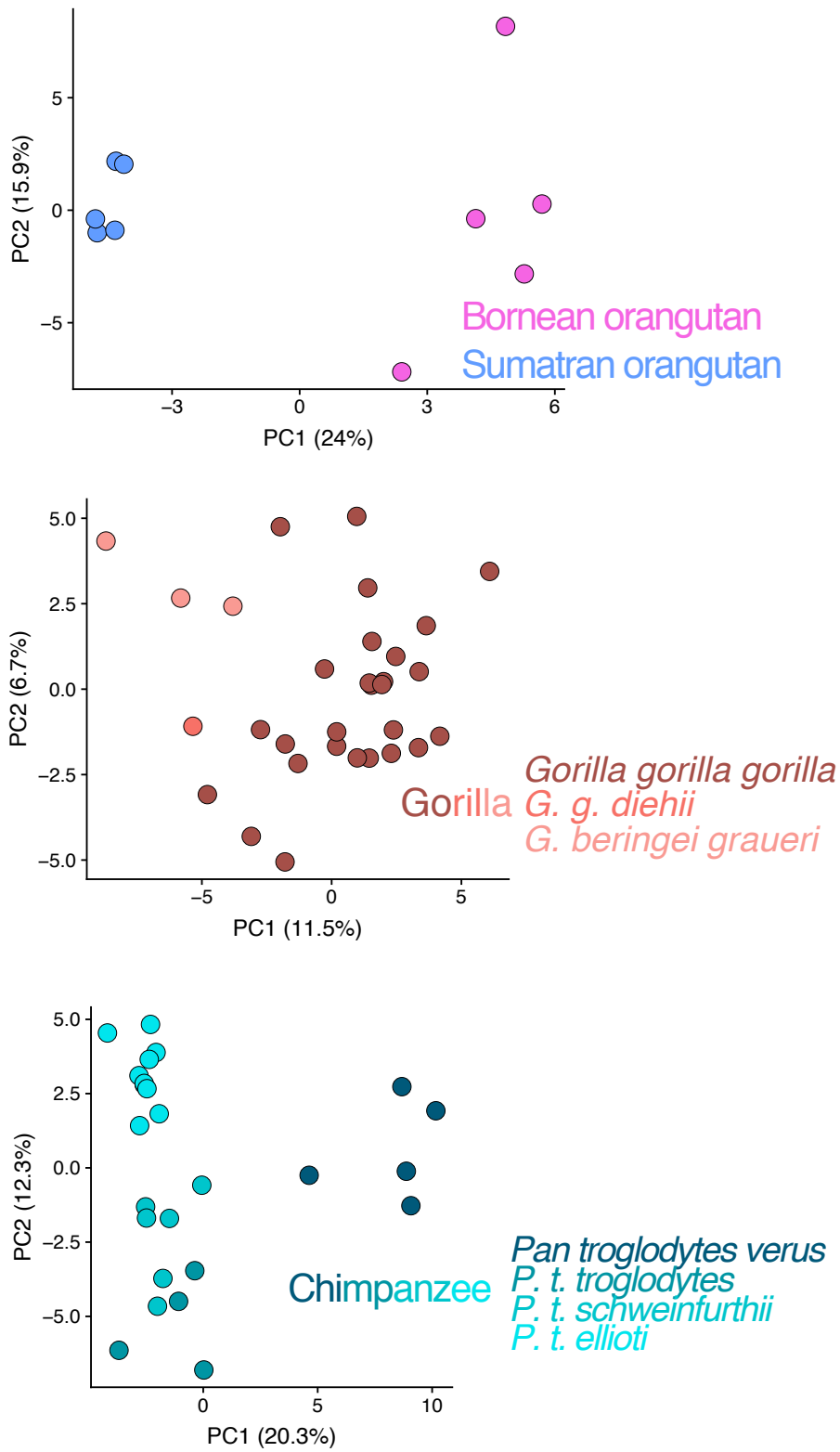

Figure 1 – Supplemental figure 3
PCAs of the NCNR compartment for the orangutan clade, gorillas, and chimpanzees demonstrate finer-scale separation and clustering of individuals by subspecies.

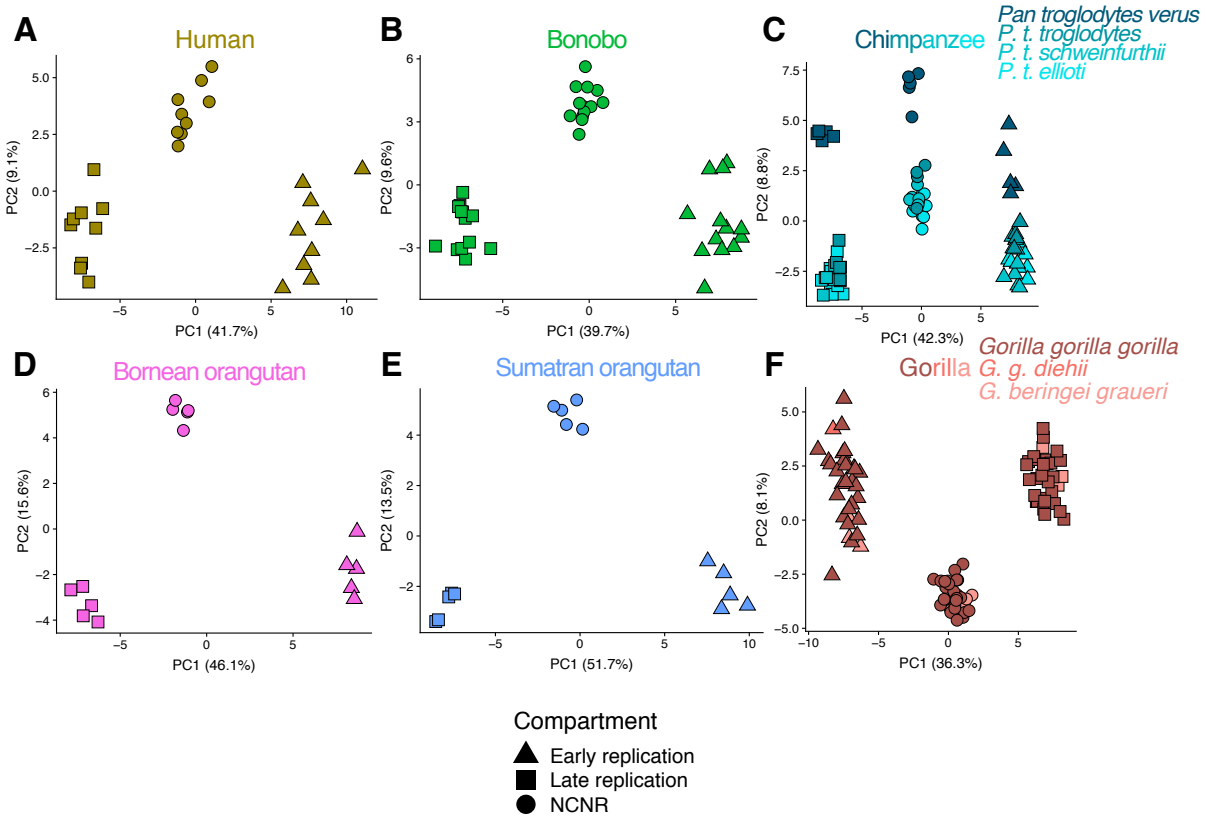

Figure 2 – supplemental figure 1

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.

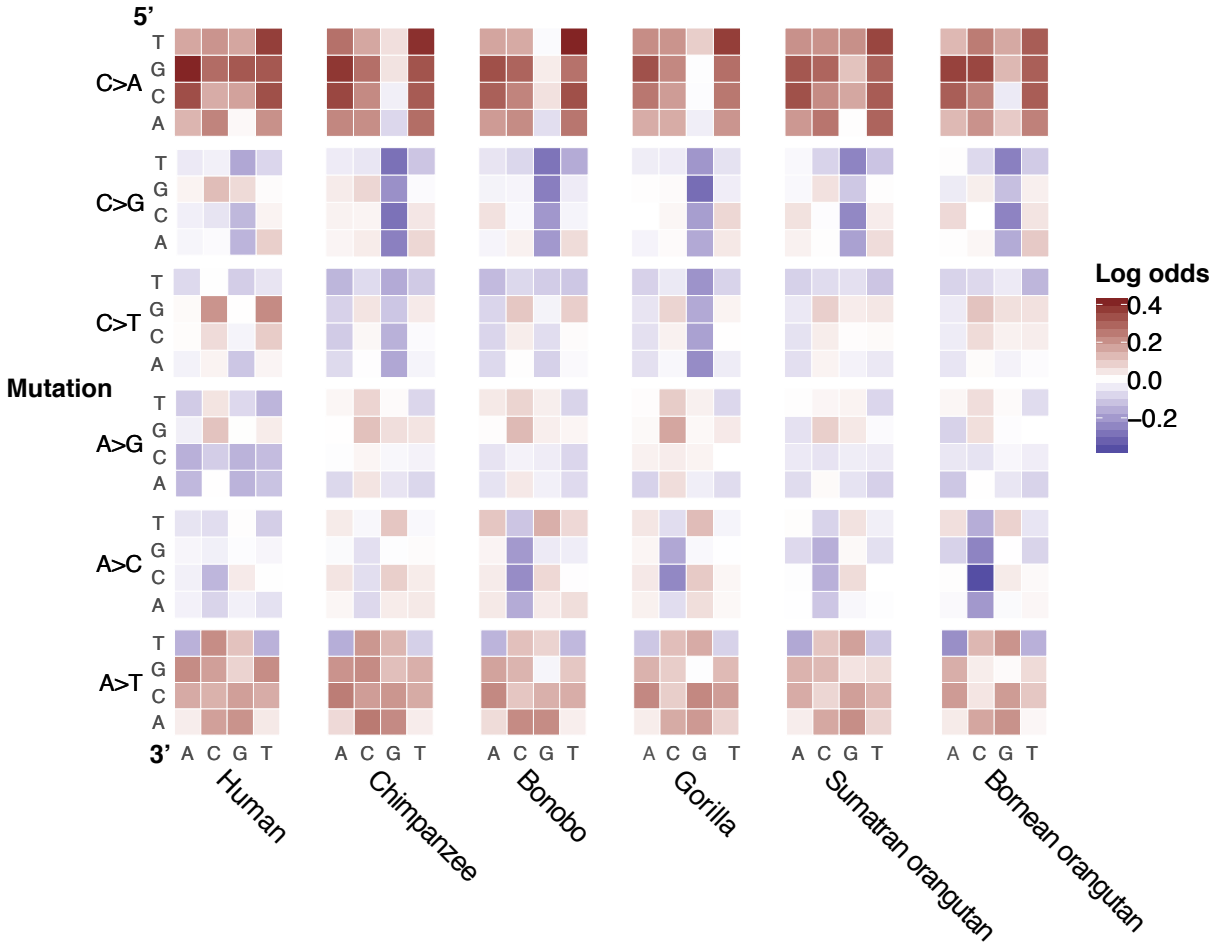

Figure 2, Supplemental figure 2

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 2, supplemental figure 3.

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

Figure 2, Supplemental figure 3
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.

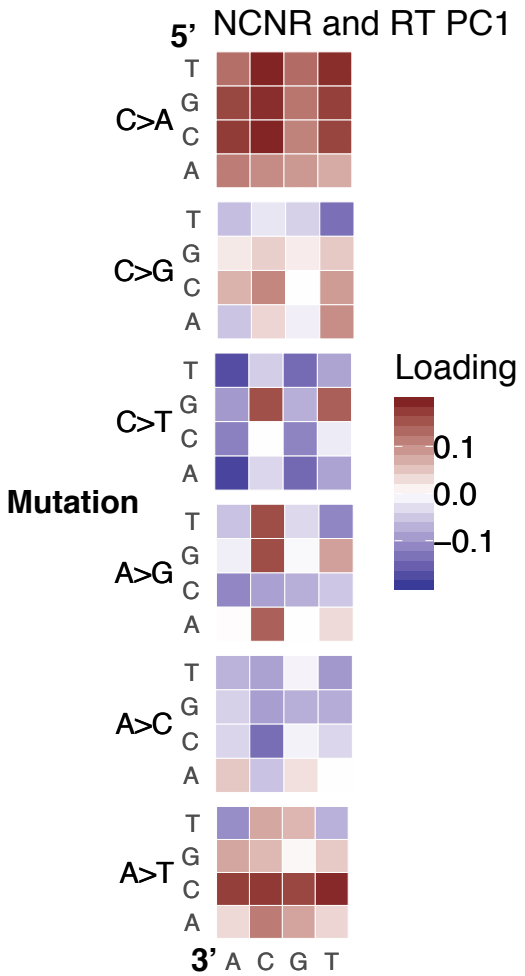

Figure 2 – supplemental figure 4

A heatmap showing the PC2 weights associated each triplet mutation type in Figure 1G. The C>A and A>T mutation types dominate PC2, which correlates with a late replication timing signature.

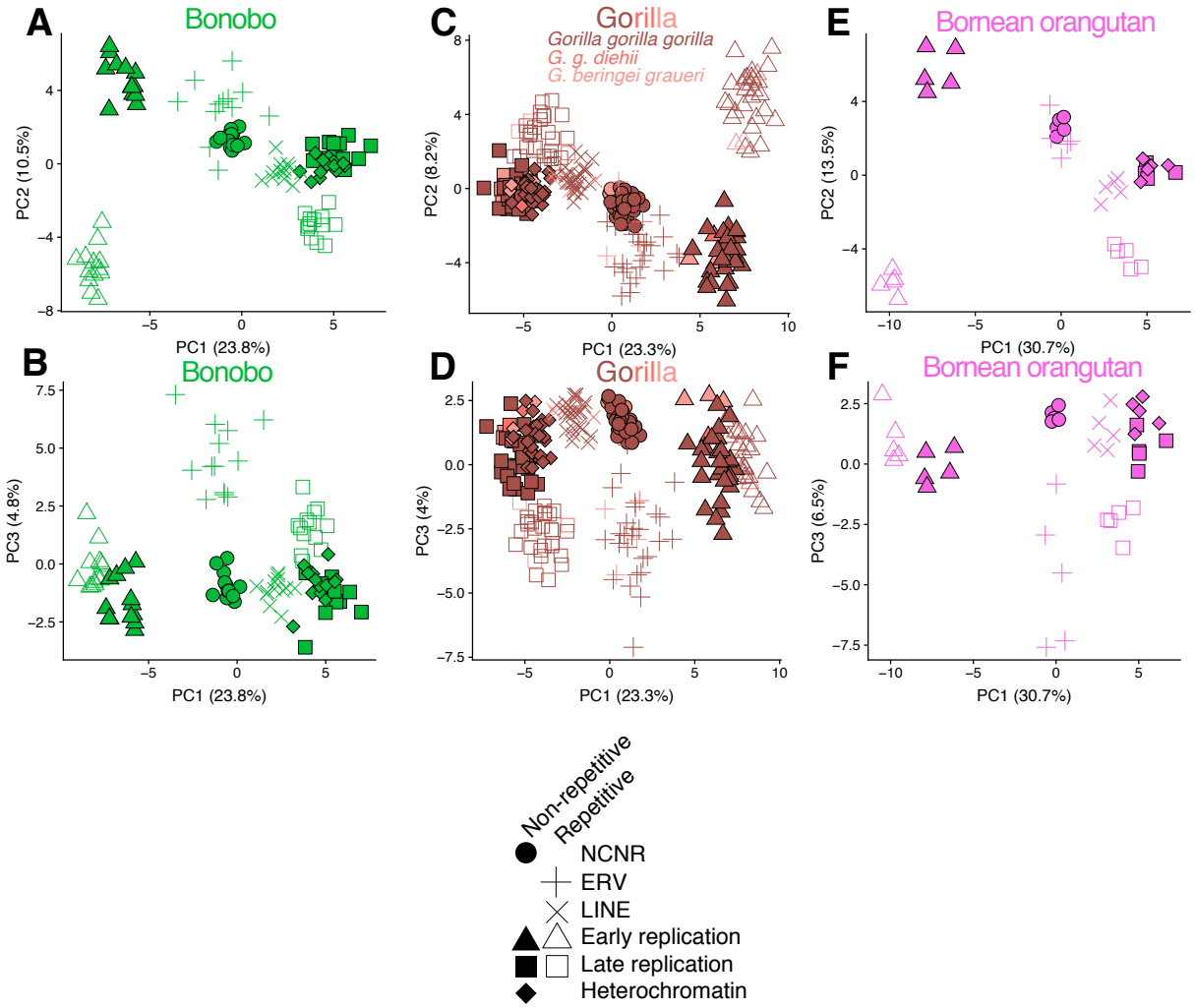

Figure 3 – figure supplement 1
PCAs of the remaining three species.

842

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

843
844 Figure 3 – figure supplement 2
845 PCAs run on individual mutation spectra from 8 different compartments are highly correlated
846 among species, demonstrating uniform dosages of species mutational signatures. We
847 calculated the midpoint of all individuals' mutation spectra for each compartment in PC1-3 and
848 for each species. We then calculated the distance between the midpoints for each pair of
849 compartments, computing a total of $\binom{8}{2} = 24$ distances for each species. We ran Pearson
850 correlation tests between the paired vectors of distances between each combination of species,
851 performing $\binom{6}{2} = 15$ tests; uncorrected P values and estimates of ρ are listed above
852
853

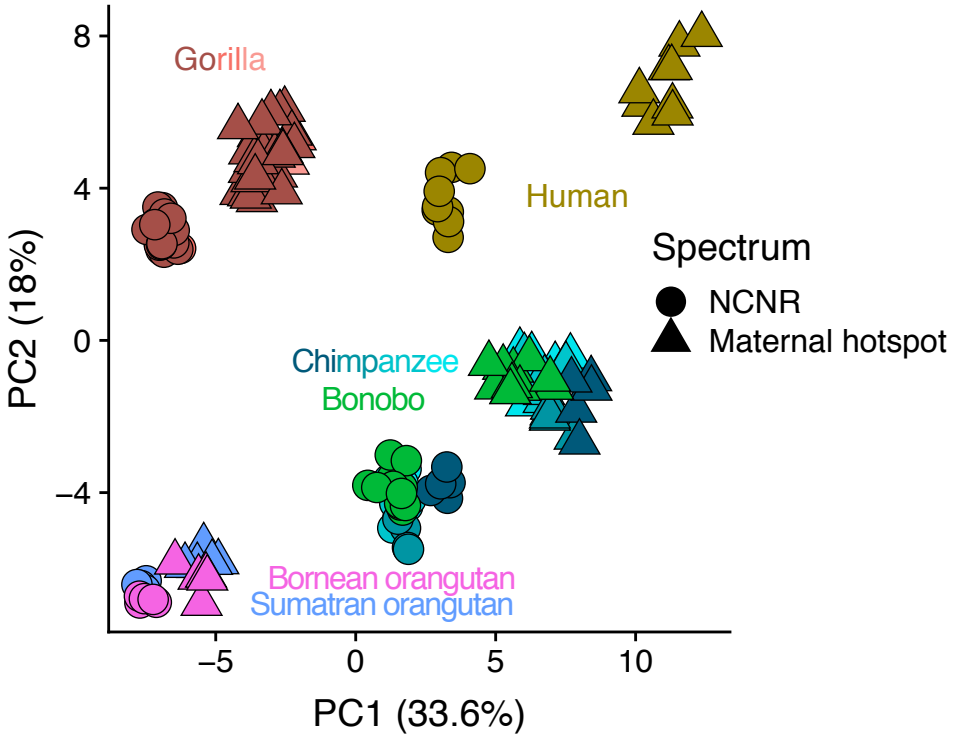

Figure 3 – figure supplement 3

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.

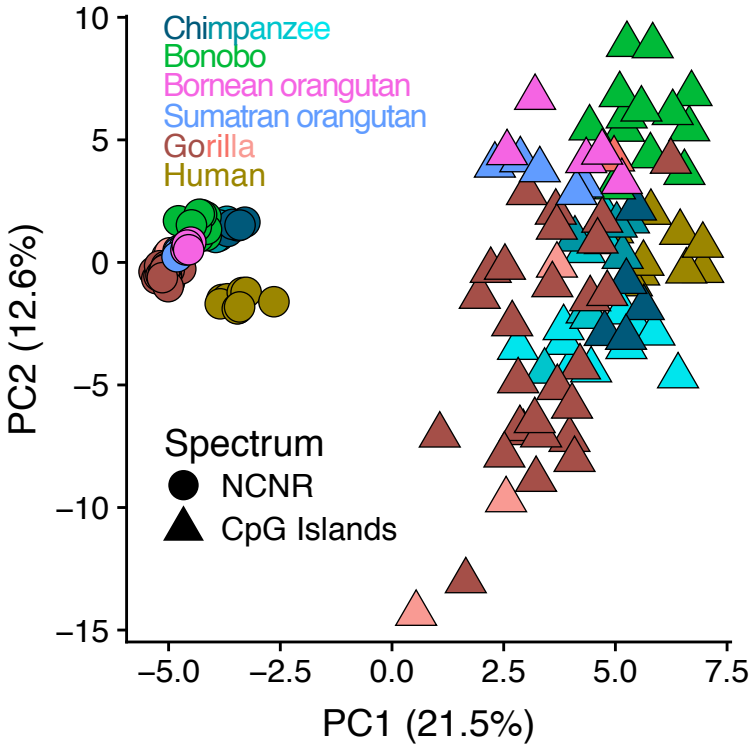

Figure 3 – supplemental figure 4
A PCA of NCNR and CpG island compartments demonstrates that differentiation of the CpG island compartment exceeds the magnitude of spectrum differentiation between species.

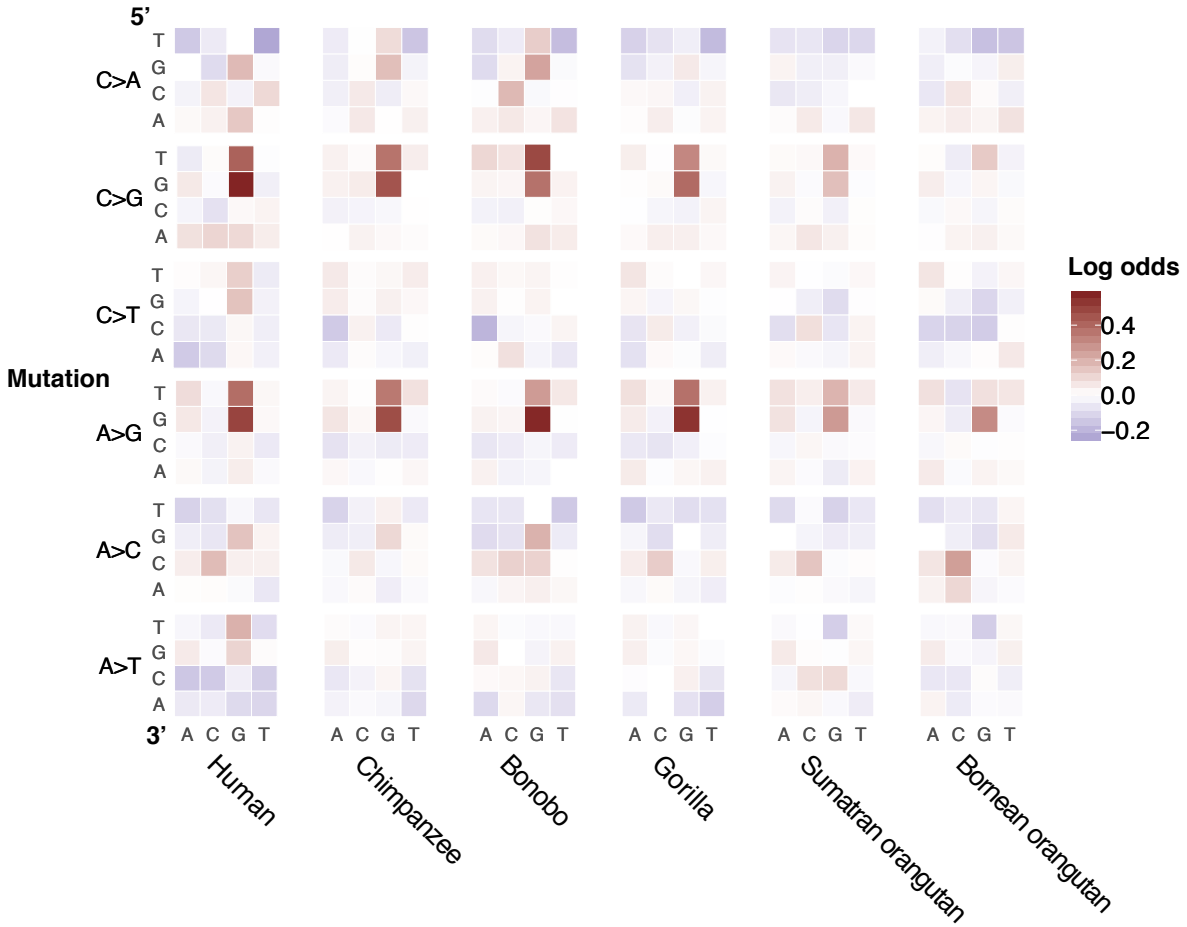

Figure 4 – supplemental figure 1

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.

878

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

879

880 Figure 4 – supplemental figure 2

881 Table of the P values from a Chi-square test for different rates of CG>GG mutation types

882 comparing the ERV hmC+ to hmC- compartments. P-values are uncorrected.

883

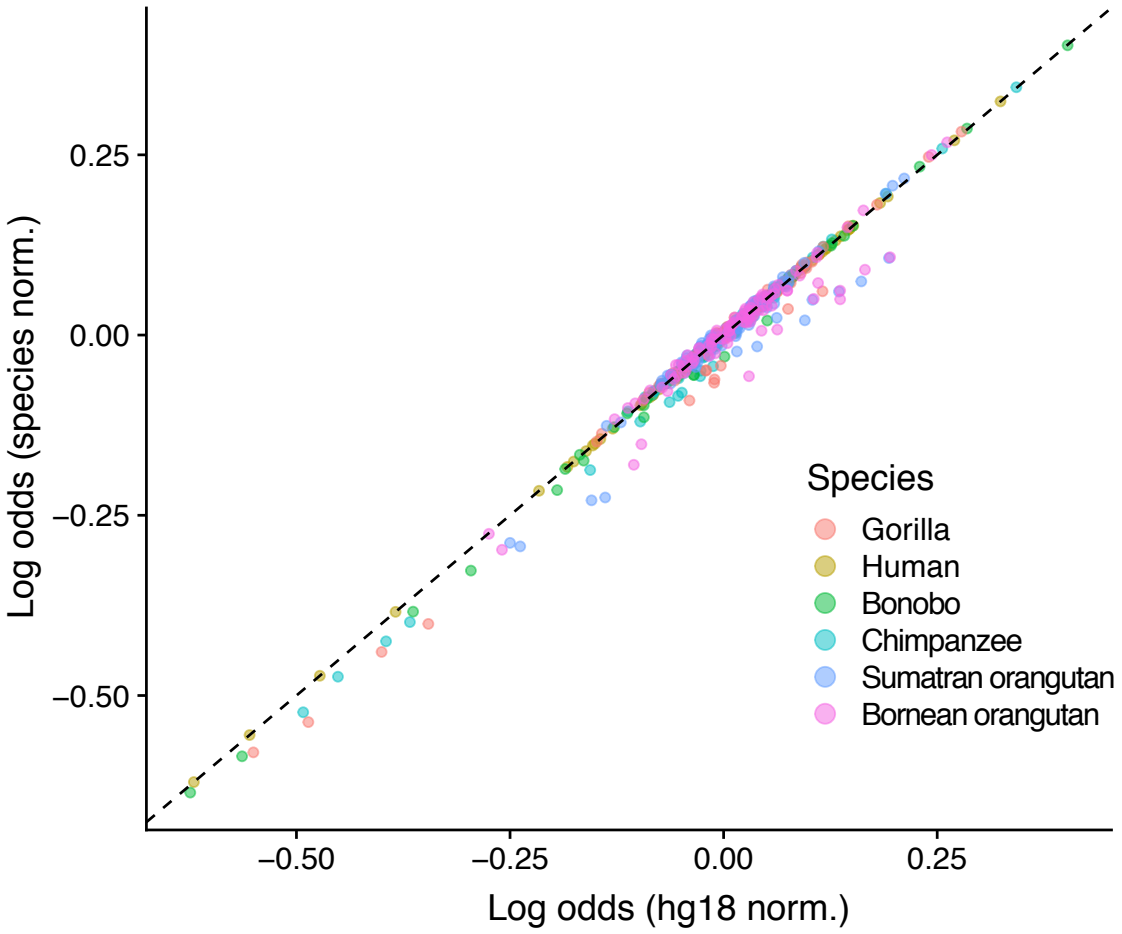

Figure 4 – supplemental figure 3

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).

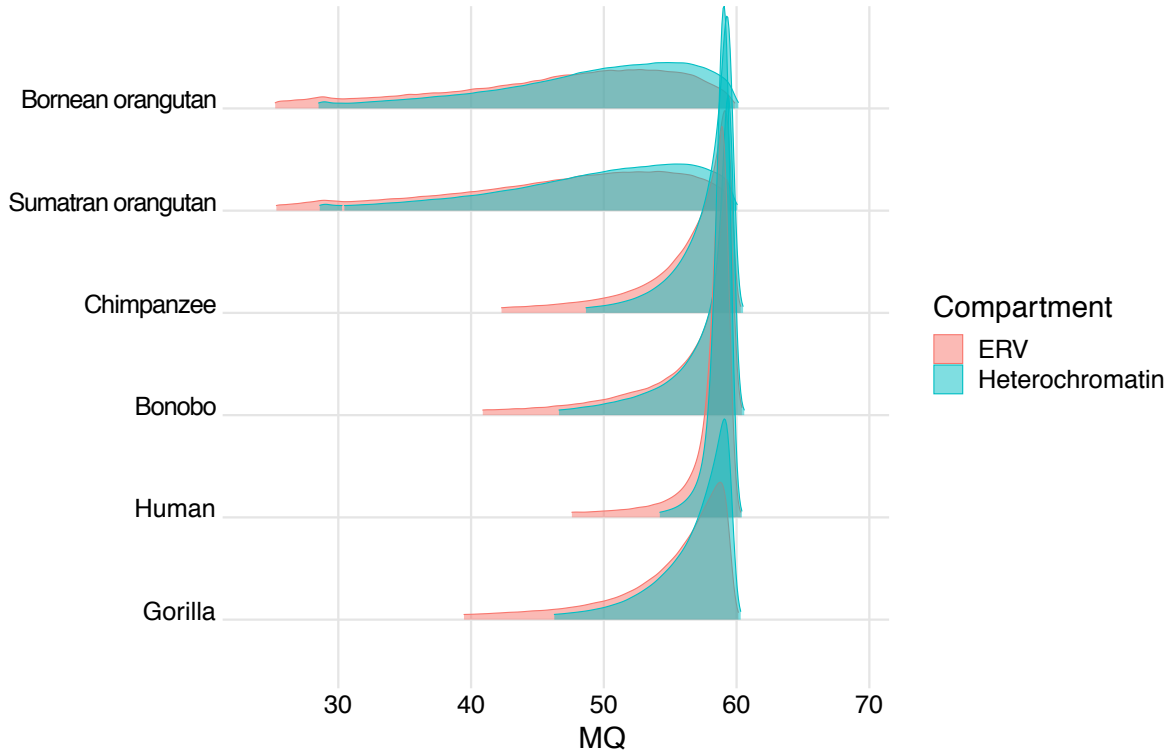

Figure 4 – supplemental figure 4

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.