Evolution of Local Mutation Rate and Its Determinants

Nadezhda V. Terekhanova,[*,†,1,2] Vladimir B. Seplyarskiy,[*,†,1] Ruslan A. Soldatov,[1,2] and Georgii A. Bazykin[1,2,3]

[1]Sector for Molecular Evolution, Institute for Information Transmission Problems of the RAS (Kharkevich Institute), Moscow, Russia

[2]M. V. Lomonosov Moscow State University, Moscow, Russia

[3]Skolkovo Institute of Science and Technology, Skolkovo, Russia

[†]These authors contributed equally to this work.

[*]**Corresponding authors:** E-mails: terekhanova@fbb.msu.ru; alicodendrochit@gmail.com.

Associate editor: Joel Dudley

Abstract

Mutation rate varies along the human genome, and part of this variation is explainable by measurable local properties of the DNA molecule. Moreover, mutation rates differ between orthologous genomic regions of different species, but the drivers of this change are unclear. Here, we use data on human divergence from chimpanzee, human rare polymorphism, and human de novo mutations to predict the substitution rate at orthologous regions of non-human mammals. We show that the local mutation rates are very similar between human and apes, implying that their variation has a strong underlying cryptic component not explainable by the known genomic features. Mutation rates become progressively less similar in more distant species, and these changes are partially explainable by changes in the local genomic features of orthologous regions, most importantly, in the recombination rate. However, they are much more rapid, implying that the cryptic component underlying the mutation rate is more ephemeral than the known genomic features. These findings shed light on the determinants of mutation rate evolution.

Key words: **local mutation rate, molecular evolution, recombination rate.**

Introduction

Germline mutation rate is known to vary substantially between chromosomal regions (Gaffney and Keightley 2005; Nusbaum et al. 2006; Hodgkinson and Eyre-Walker 2011), and this variation is medically relevant (Michaelson et al. 2012; Veltman and Brunner 2012; Wong et al. 2016). Mutagenesis is affected by a range of biochemical processes, most importantly, meiotic recombination, replication and transcription, as well as by chromatin structure. Consistently, both the somatic and the germline local mutation rate (LMR) is strongly dependent on genomic features such as replication timing (Woo and Li 2012), density of DNase-hypersensitive sites (DHSs) (Thurman et al. 2012), gene density, histone modifications, GC-content, etc. (Ananda et al. 2011; Schuster-Bockler and Lehner 2012; Kuruppumullage Don et al. 2013). Still, up to 70% of the human germline LMR variation at megabase scale cannot be explained by the known features (see supplementary fig. S4, in Schuster-Bockler and Lehner (2012)). Understanding the LMR variation and its causes is critical for inferring genomic functional elements and the genetic basis of heritable disease and cancer (Veltman and Brunner 2012; Lawrence et al. 2013; Eyre-Walker and Eyre-Walker 2014).

The LMR landscape is dynamic. Germline LMRs co-vary between closely related species (Tyekucheva et al. 2008), but are almost independent of each other in remote species (Imamura et al. 2009). Although the variation in LMR has been studied extensively, the dynamics and causes of LMR evolution are poorly understood (but see Tyekucheva et al. 2008; Ananda et al. 2011; Hodgkinson and Eyre-Walker 2011). LMR evolution may be driven by changes in known genomic features or by other factors.

Evolution and co-evolution of the LMR and genomic features can be studied by analyzing the correlations between LMRs and features of orthologous genomic regions in species at a range of phylogenetic distances from each other. Here, we make use of complete genome alignments of nine primate species, and of mouse, to study the evolution of the germline LMR between closely related vertebrates. We show that although only less than a half of the variance in LMR either in human or in apes can be explained by the known human genomic features, the LMRs in human and in apes are very strongly correlated, implying the existence of a strong "cryptic" component of the LMR variability. Furthermore, most of the genomic features are evolutionary stable and are good predictors of the LMR even in distantly related species; still, some changes in the LMR between species may be traced to changes in the underlying features, notably, in the recombination rate. In contrast, the "cryptic" fraction of the LMR variation not explainable by genomic features evolves very rapidly.

Results

LMRs Are Strongly Correlated between Humans and Apes

We study the multiple sequence alignment of eight primate genomes (chimpanzee, gorilla, orangutan, gibbon, rhesus macaque, green monkey, squirrel monkey, and marmoset) with human (Karolchik et al. 2014), split into 2,261 1 Mb non-

Article

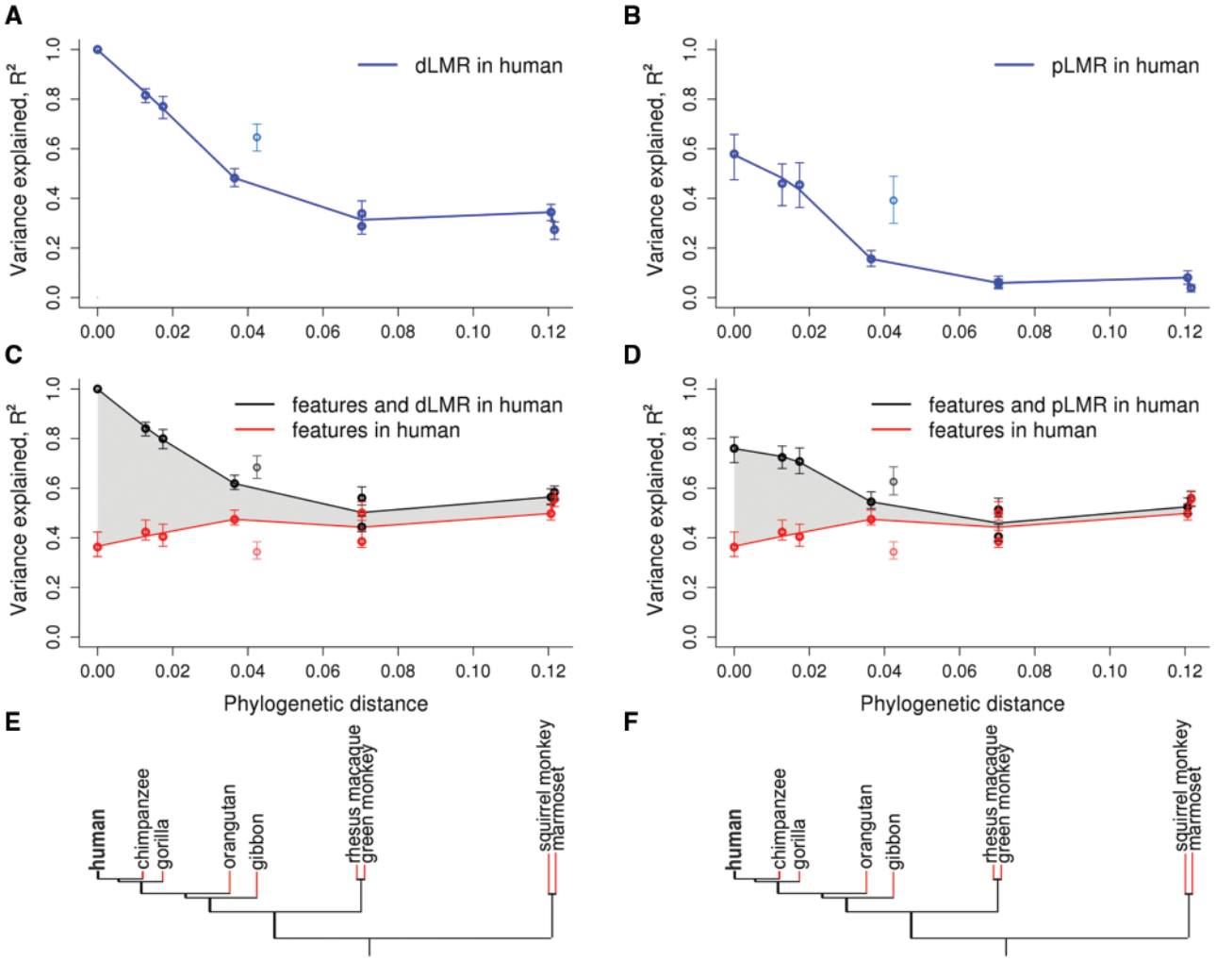

FIG. 1. LMR variation in primate species explained by genomic features and LMR in the human lineage. *(A–D)* For each primate species, the fraction of the explained variance in LMR (R^2, vertical axis) is plotted against the phylogenetic distance from human (horizontal axis). The values for the gibbon, which has a high rate of rearrangements (Carbone et al. 2014), are plotted, but were not included in the fit. Variance in dLMR explained by the human dLMR (*A*) or pLMR (*B*). Variance in dLMR explained by human genomic features alone (red) or in combination with the human dLMR (*C*) or pLMR (*D*; black). The following features were included in the model: GC-content, recombination rate, number of exonic nucleotides, replication timing, number of DHSs, densities of H3K27ac, H3K27me3, and H3K9me3 histone marks and MAF. The shaded area represents the inferred fraction of the variance in non-human LMR explainable by the human LMR independently of the genomic features. Error bars correspond to 95% CIs obtained by bootstrapping. *(E, F)* Phylogenetic tree of the considered species. Red color denotes branches for which the LMR was calculated.

overlapping windows (the results obtained for 100 Kb windows were generally similar; supplementary text S1, Supplementary Material online), together with the data on human polymorphism and de novo mutations in these same windows. To minimize the effect of selection on LMR estimates, we exclude exons and untranslated regions (UTRs), and include among the analyzed genomic features proxies for the strength of selection in non-coding windows: the mean frequency of minor allele (MAF), the fraction of exonic nucleotides, and the mean multispecies conservation (see below). Still, selection acting at non-coding regions may confound inference of mutation rate variability (see Discussion section).

For each species, we infer the nucleotide substitutions since its divergence from the last common ancestor with its closest relative (fig. 1*E* and *F*), and count the number of such substitutions in each window as a proxy for the LMR (divergence-based LMR [dLMR]). Additionally, for humans, we also estimate LMR using the numbers of rare SNPs (McVean 2012) (polymorphism-based LMR, pLMR) and the numbers of observed de novo mutations (Wong et al. 2016) (mLMR). Although mLMR is the gold standard for LMR measurements, the data on it are limited (supplementary text S2, Supplementary Material online), and we only use it for validation. pLMR is slightly better correlated with mLMR than dLMR, although both correlations are significant ($P < 0.01$; supplementary fig. S1, Supplementary Material online). We exclude substitutions prone to biased gene conversion from analyses (Materials and Methods section and supplementary text S3, Supplementary Material online).

The LMRs are strongly correlated between human and chimpanzee ($R^2 = 0.82$ for dLMR, $P < 2.2 \times 10^{-16}$; $R^2 = 0.46$ for pLMR, $P < 2.2 \times 10^{-16}$; fig. 1*A* and *B*). When less related species are considered, this correlation decays with phylogenetic distance, reaching the minimal value among

Table 1. Correlations between Human and Mouse Genomic Features in 1 Mb Genomic Windows.

Feature	Pearson's *R*	*P* value
LMR	0.34 (0.29, 0.38)	$<2.2 \times 10^{-16}$
Recombination rate	0.01 (−0.04, 0.06)	0.67
DHSs	0.83 (0.82, 0.85)	$<2.2 \times 10^{-16}$
Replication timing	0.71 (0.69, 0.74)	$<2.2 \times 10^{-16}$
GC-content	0.94 (0.93, 0.94)	$<2.2 \times 10^{-16}$
Exonic nucleotide density	0.95 (0.94, 0.95)	$<2.2 \times 10^{-16}$
H3K9me3	0.32 (0.28, 0.37)	$<2.2 \times 10^{-16}$
H3K27ac	0.73 (0.70, 0.75)	$<2.2 \times 10^{-16}$
H3K27me3	0.66 (0.63, 0.69)	$<2.2 \times 10^{-16}$

NOTE.—The 95% CIs (asymptotic CIs estimated based on Fisher's Z transform) are in parentheses.

primates in marmoset ($R^2 = 0.27$ for dLMR, $P < 2.2 \times 10^{-16}$; $R^2 = 0.04$ for pLMR, $P < 2.2 \times 10^{-16}$; fig. 1*A* and *B*), and is even lower in mouse ($R^2 = 0.11$ for dLMR, $P < 2.2 \times 10^{-16}$; $R^2 = 0.02$ for pLMR, $P < 3.13 \times 10^{-7}$; supplementary fig. S2, Supplementary Material online). This decay is independent of the decrease in the fraction of alignable nucleotides with phylogenetic distance, as the correlation between LMRs remains similar when only the columns of the multiple alignments without gaps or ambiguous nucleotides in any of the species are considered (supplementary fig. S3, Supplementary Material online). The proportion of the variance in LMR explainable by the human LMR decays by half at phylogenetic distance of ~0.04 substitutions per site, or ~16 million years (dos Reis et al. 2012), roughly corresponding to the last common ancestor of human and orangutan. Human mLMR is also better correlated with the dLMRs of the more closely related species, compared with more distant ones (supplementary figs. S1 and S4, Supplementary Material online).

A Cryptic Component to the LMR

The LMR depends on DNA properties (Ananda et al. 2011; Schuster-Bockler and Lehner 2012; Kuruppumullage Don et al. 2013). The linear model that predicts the human dLMR from the measured genomic features of embryonic stem cells explains 36% of the variance in dLMR, which is slightly higher than the previous estimates (Schuster-Bockler and Lehner 2012) based on a feature annotation from a different tissue (supplementary text S4, Supplementary Material online). The remaining variance may be random, or associated with genomic features not picked up by our analyses. The fact that the human LMR is a good predictor for the LMR in apes, in particular, in chimpanzee and in gorilla (fig. 1*A* and *B*), suggests that the LMR variation not explainable by the measured genomic features still has a strong non-random component conserved between species.

To better understand this cryptic component, we first ask how well the human genomic features predict the LMR in non-human primates (supplementary fig. S5, Supplementary Material online). For this, we construct, for each non-human species, a linear model predicting the dLMR in this species from human features alone and in combination with the human dLMR or pLMR. The non-human dLMR can be predicted nearly as well as the human dLMR by the features of the human orthologous segments (fig. 1*C* and *D*, red line). This is consistent with the generally conservative nature of genomic features (Yue et al. 2014). Indeed, most of the features are very strongly correlated even between human and mouse (table 1).

In contrast, adding the human LMR to the linear model radically increases the fraction of explained variance (fig. 1*C* and *D*, black line): from $R^2 = 0.42$ to 0.72 (for pLMR) or to 0.84 (for dLMR) in chimpanzee, and from 0.40 to 0.71 (for pLMR) or to 0.80 (for dLMR) in gorilla. Therefore, in closely related apes, the linear model that includes both the data on genomic features and the human polymorphism or divergence data explain about twice as much variance in LMR as the model with the genomic features alone, implying that about a half of the explained variance in LMR is cryptic (fig. 1*C* and *D*, shaded area).

Furthermore, there is a striking difference in how the explanatory power of genomic features and LMR changes with phylogenetic distance. Although the human genomic features explain about as much variance in the LMR for distantly related as for closely related species, the human LMR predicts the LMR in closely related apes much better than that in less related primates. Therefore, unlike the measured genomic features which are relatively stable, the cryptic component of variation in LMR is short-lived. For the mouse LMR, human genomic features are much better predictors than the human LMR (supplementary fig. S2, Supplementary Material online), suggesting that conserved features are important predictors of the LMR at large phylogenetic distances, while the LMR of a remote species carries no additional information. The cryptic component decays uniformly with phylogenetic distance, with the exception of the gibbon genome which carries an unusually high number of rearrangements (Carbone et al. 2014; see supplementary text S5, Supplementary Material online). Again, the shape of this decay is independent of the differences in alignment quality between species (supplementary fig. S3, Supplementary Material online).

Stability of Genomic Features in Determination of the LMRs

The LMRs of orthologous genomic regions evolve with time (fig. 1). The fraction of the variance in LMR explained by the human genomic features is similar in closely related and in distantly related primate species. Still, it is possible that individual features, and thus their power to predict the LMR in another species, change at different rates. We asked to what extent changes in LMR are determined by the evolution of individual features. As the data on genomic feature landscapes of primates are limited (Cain et al. 2011; Zhou et al. 2014), we addressed this question indirectly.

For this, we estimated the fraction of the variance in dLMR explained by individual human features. Because features are correlated with each other (supplementary fig. S6, Supplementary Material online), we performed the ANOVA type III analysis to single out the independent contribution of each feature accounting for the contributions of other features (fig. 2*A*). The estimated relative contributions of different features to the human dLMR are in line with previous work (Tyekucheva et al. 2008; Schuster-Bockler and Lehner

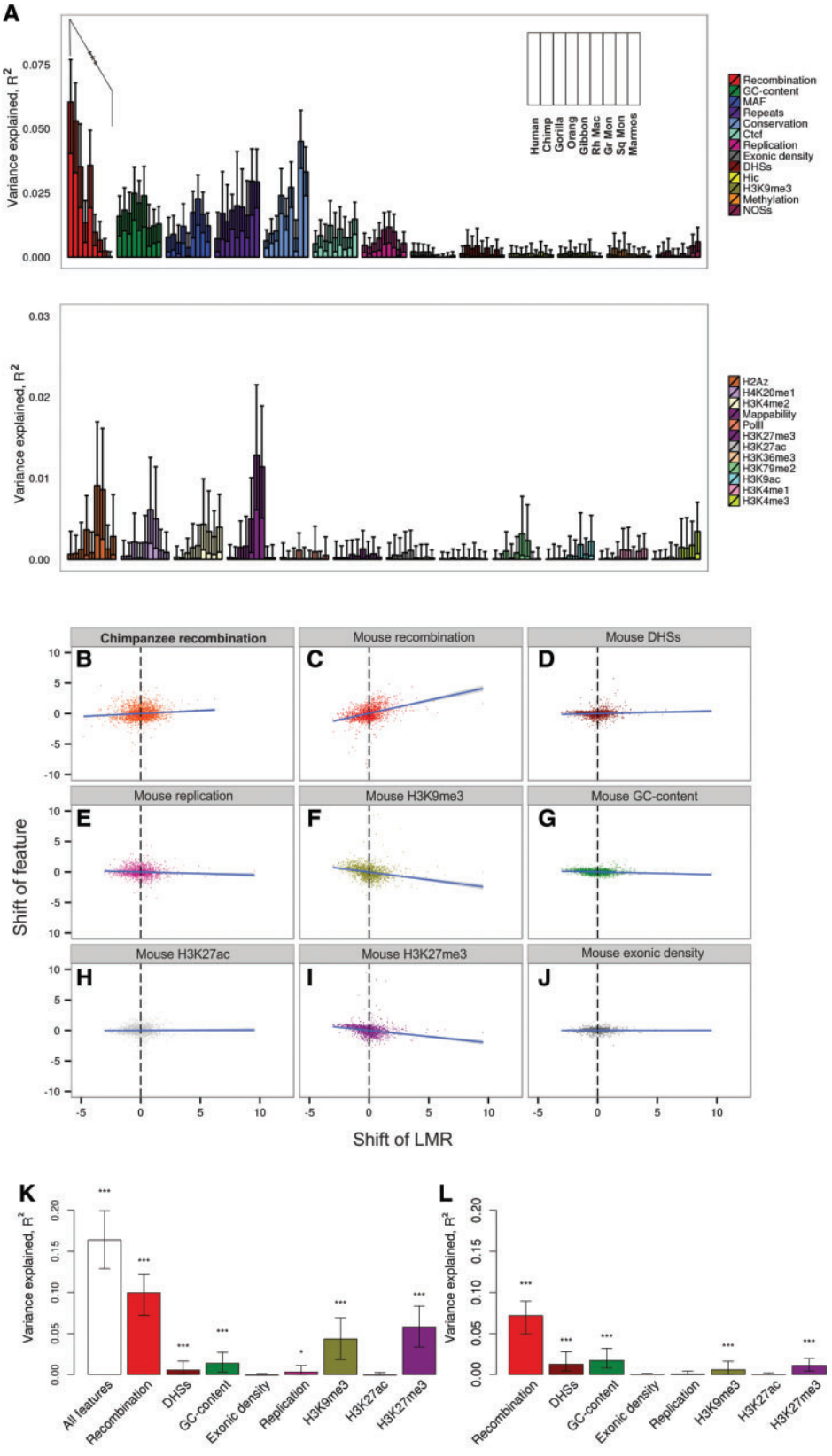

Fig. 2. LMR explained by individual genomic features. *(A)* Variance of the LMR explained by the 25 genomic features. For each genomic feature, the fraction of the variance explained by this feature in ANOVA type III analysis is shown for each primate at increasing phylogenetic distances from human. The inset shows the order of the species along the horizontal axis (same as in fig. 1). The features are ordered by the fraction of the variance explained in human (note the different scale of the vertical axis). The asterisks indicate the significance of the negative correlation between the R^2 and the phylogenetic distance. *(B–J)* Scatterplots for raw correlations of the changes in the LMR and in select genomic features between human and mouse lineages. Each dot corresponds to a 1 Mb window. *(K–L)* Changes in the LMR explained by the changes in select genomic features

2012; Kuruppumullage Don et al. 2013). For the dLMRs in non-human primates, they are also mostly similar to those for the human dLMR, and for most features are independent of the phylogenetic distance to the analyzed species (supplementary fig. S7, Supplementary Material online). The only feature with contribution declining with the phylogenetic distance is the recombination rate ($P = 0.009$ for the correlation between R^2 and the phylogenetic distance, supplementary fig. S7, Supplementary Material online). The variance explained by it is high initially, in line with the evidence for its major effect on LMR (Lercher and Hurst 2002; Webster and Hurst 2012; Yang et al. 2015); but decreases rapidly with phylogenetic distance, from 6.05% for humans to 0.01% for marmoset (fig. 2A). This decay is linked to recombination per se, rather than to the associated process of gene conversion, as our analysis excludes the substitutions prone to biased gene conversion. Instead, they are in agreement with the GC-biased gene conversion (gBGC)-independent mutagenic role of recombination (Arbeithuber et al. 2015; Yang et al. 2015). The observed decay suggests that changes in recombination are rapid, in line with its known high evolvability (Auton et al. 2012; Pratto et al. 2014; Glemin et al. 2015). Such changes may contribute to changes in the LMR.

Changes in Recombination Rate Are Associated with Changes in LMRs

The importance of the local recombination rate for the LMR evolution is further supported by comparisons with chimpanzee and mouse. For these two species, recombination maps are available (Auton et al. 2012; Brunschwig et al. 2012). Recombination is poorly conserved between species: the human recombination rate is only weakly correlated even with that of chimpanzee ($R^2 = 0.24$, $P < 2.2 \times 10^{-16}$), and not correlated with that of mouse (table 1).

For each genomic window, we compared the interspecies differences in dLMR with differences in recombination rates. In both human–chimpanzee and human–mouse comparisons, they were weakly positively correlated ($R^2 = 0.01$, $P < 7 \times 10^{-6}$ for chimpanzee, fig. 2B; and $R^2 = 0.1$, $P < 2.2 \times 10^{-16}$ for mouse; fig. 2C), implying that an increase in the recombination rate of a genomic region between species is associated with an increase in LMR, and vice versa.

For the human–mouse comparison, we also analyzed several other genomic features, asking whether their changes are correlated with changes in the LMR. In total, ~16.4% of the variance in dLMR differences between branches could be explained by differences in the feature landscapes (fig. 2K). When contributions from individual variables were considered, differences in recombination rate alone explained ~10% of the variance (fig. 2C and K), while other features explained substantially less (fig. 2D–J and K). To single out the genomic features in which changes between mouse and human lineages independently contribute to changes in the LMR between these two species, we performed the ANOVA (type III) analysis (fig. 2L). Changes in only a few of the genomic features significantly contributed to changes in the dLMR. The most substantial contributor was recombination; its contribution was much higher than that of the second-best contributor (GC-content). Although the recombination landscape changes rapidly, and human recombination hotspots are not informative about the positions of hotspots in mouse (table 1), changes in recombination rate landscape explain changes in dLMR more than those of any other features.

To better understand the link between changes in recombination and mutation, we studied the genomic windows in which the LMR has been substantially accelerated or decelerated in the human lineage, or in the chimpanzee lineage, since divergence from the human-chimpanzee common ancestor. In the human-accelerated regions (HARs), the human recombination rate is substantially higher than the genome average (one-sided Wilcoxon rank sum test $P = 2.9 \times 10^{-9}$, fig. 3A), implying that the HARs frequently carry recombination hotspots. In contrast, in the chimpanzee-accelerated regions (CARs), the human recombination rate is only slightly higher than the genome average ($P = 8.0 \times 10^{-3}$, fig. 3B). Together, these data imply that the HARs are frequently associated with recombination hotspots that are short-lived, so that they increase the mutation rate in human more than in chimpanzee. Reciprocally, the chimpanzee recombination rate is elevated in CARs ($P = 3.7 \times 10^{-7}$, fig. 3D), slightly more than in HARs ($P = 7.4 \times 10^{-6}$, fig. 3C).

We also analyzed the genomic regions that were substantially decelerated in human (human decelerated regions, HDR) or chimpanzee (CDR). The human recombination rate at HDRs as well as the chimpanzee recombination rate at CDRs were slightly reduced compared with the genome average (one-sided Wilcoxon rank sum test $P = 7.7 \times 10^{-4}$ and $P = 3.9 \times 10^{-4}$, respectively; supplementary fig. S8, Supplementary Material online). This implies that the LMR deceleration is also partially caused by recombination-related factors, although likely to a smaller extent than LMR acceleration. All these patterns were associated with the recombination per se rather than biased gene conversion (see supplementary text S6, Supplementary Material online).

Discussion

Although the germline LMR is known to differ between species (Hodgkinson and Eyre-Walker 2011; Lynch et al. 2016), the rate and the driving forces of the LMR evolution are obscure. To our knowledge, this study is the first quantitative analysis to this end.

FIG. 2 Continued

between human and mouse. Vertical axis, fraction of variance in differences in LMRs explainable by differences in genomic features between human and mouse. Columns correspond to the variance explained by features measured by R^2 (K) or ANOVA type III analysis (L). Error bars correspond to 95% CIs obtained by bootstrapping. Asterisks indicate the significance of the deviation of the regression line from 0 (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$).

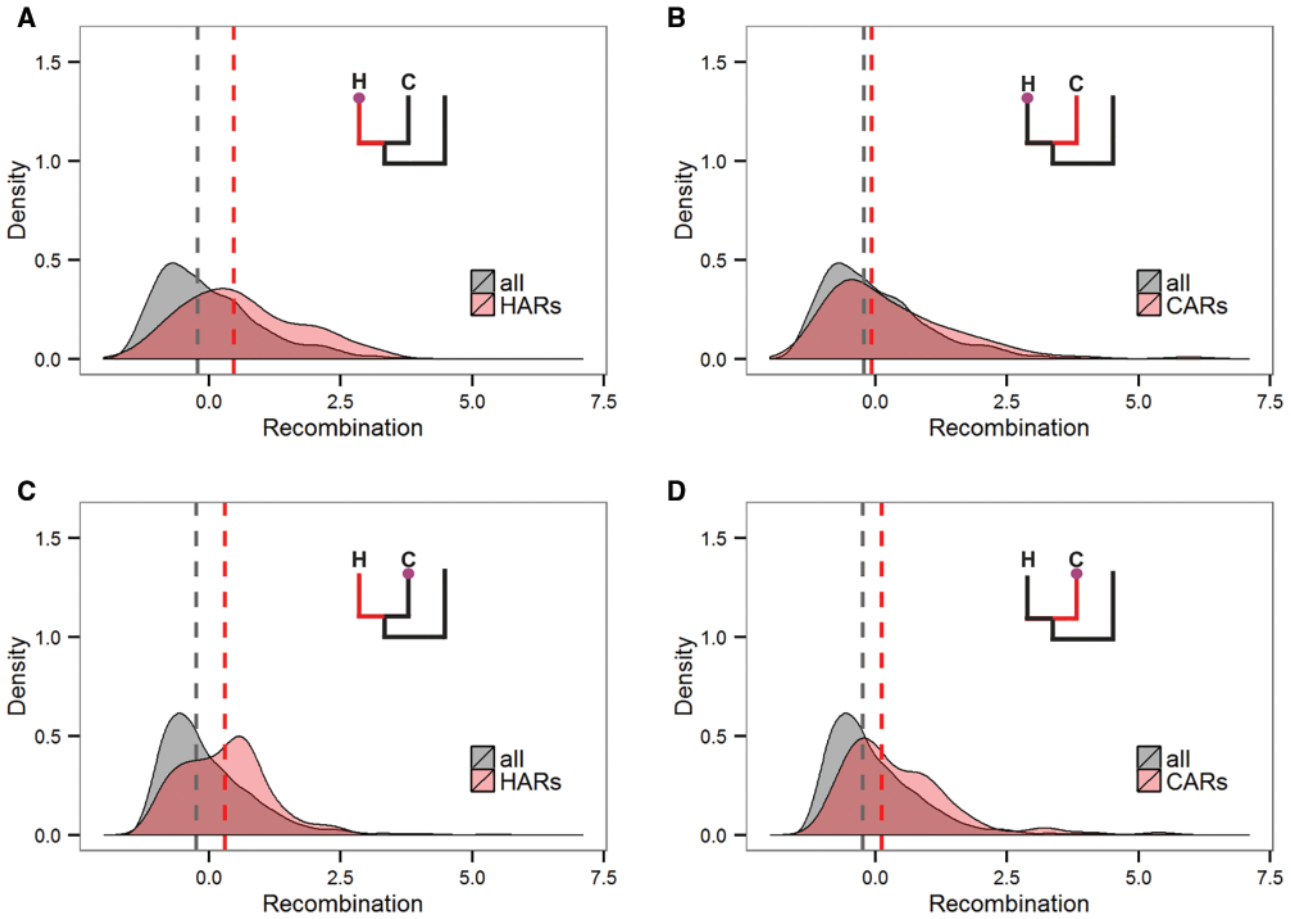

FIG. 3. Distributions of human recombination scores (*A*,*B*) in HARs (*A*) and CARs (*B*), and of chimpanzee recombination scores (*C*,*D*) in HARs (*C*) and CARs (*D*). The schematic phylogenies show the lineage in which the LMR was increased in red (H, human or C, chimpanzee), and the species in which recombination was measured, as a circle. The dashed line corresponds to the median recombination rate in the ARs and all genomic regions.

Our approach to estimation of the germline mutation rate from divergence and polymorphism data has three important caveats. First, aside from mutation, selection and gBGC can affect both divergence and polymorphism. To limit the effect of selection, we only analyzed intronic and intergenic regions, as only ~8% of mutations at these regions are affected by selection (Rands et al. 2014). The contribution of MAF to the explained variance is significant (fig. 2A), implying that the effect of selection on LMR is still high even within the noncoding regions. However, the contribution of MAF is much smaller than the unexplained component of the LMR variance, implying that the high correlation between LMRs of closely related species is not due to common selection pressures. Moreover, the contribution of MAF is roughly constant between species at different phylogenetic distances (fig. 2A), implying that selection pressures are rather stable, and their changes contribute little to the LMR evolution. gBGC is indeed associated with divergence (Glemin *et al.* 2015) and polymorphism; however, its effect is weak or absent in the considered subset of substitutions (supplementary texts S3 and S6, Supplementary Material online). Second, the estimates of the LMR can be affected by the polymorphism in the ancestral population. In particular, the differences in coalescence times between genomic regions may inflate the correlation between the dLMRs estimated from two descendant sister branches (McVicker et al. 2009; Charlesworth 2010; Gossmann et al. 2011). Similarly, such differences may contribute to the association between the dLMR and the recombination rate, as genomic regions with higher recombination rates have larger local effective population size, and therefore longer coalescence times (Gossmann et al. 2011; Hobolth et al. 2011; Francioli et al. 2015). Such phenomena do not affect correlations with de novo mutations (supplementary fig. S1, Supplementary Material online). They are also unlikely to contribute to pLMRs, as transspecies polymorphisms are rare (Leffler et al. 2013; Asthana et al. 2005), and can only contribute to the correlations between dLMRs, and the correlation between LMR and recombination, of human, chimpanzee and gorilla—the species in which the contribution of the ancestral polymorphism to divergence is reported (Scally et al. 2012). In all other comparisons, a relatively long period of independent divergence (black edges in the phylogeny of fig. 1) separates the branches at which the dLMR is measured (red edges in fig. 1). Furthermore, the time to coalescence of a genomic region is highly correlated with its exonic density (Hobolth et al. 2011), and the correlation between the exonic density and the dLMR is low ($R^2 = 0.01$, fig. 2A), suggesting that the ancestral polymorphism also does not contribute

much to correlations between dLMRs, or to correlations between the LMR and the recombination, even in the most closely related species. More generally, our model includes multiple proxies for the strength of selection, both direct and background (MAF, fraction of exonic nucleotides, multispecies conservation and recombination rate) among the genomic features that predict the LMR, and therefore, these features are unlikely to contribute to the unexplained variation ("gray zone" in fig. 1). Still, such confounders may affect dLMR estimates to some degree; in particular, they may be the reason why the correlations between dLMRs are higher and more long-lived than the correlations between the pLMR in human and dLMRs in other species. Third, the mutation rate estimation is dependent on the window size, with small windows giving unreliable estimates (Imamura et al. 2009). The window sizes we use, 100 Kb and 1 Mb, are a compromise between resolution and robustness.

Given these caveats, we show that the correlation between the LMRs of closely related species is surprisingly high. As a result, the LMR of a species can be much better predicted by the LMR of a closely related species than by its own genomic features. Still, the LMR is plastic, being rather poorly conserved even between moderately related species of primates, and evolves much faster than the known genomic features. These results imply the existence of a strong but transient cryptic component in LMR variation; the unknown genomic features that underlie it must undergo a rapid turnover, changing at the timescale of a few tens of millions of years or less. These yet-unknown features may include characteristics of efficiency of DNA repair or DNA susceptibility to damaging agents or replication errors. The rapid evolution of the LMR may be driven by the same processes that cause changes in the mutation spectrum (Harris and Pritchard 2016; Mathieson and Reich 2016) and mutation rate between the human populations (Mallick et al. 2016).

In contrast to the human LMR, the predictive power of known human genomic features for the non-human LMR changes little with phylogenetic distance, probably due to the conservative nature of these DNA properties. As a result, in the human-mouse comparison, human genetic features are better predictors of the LMR in mice than the human LMR.

Our findings therefore imply that changes in the known genomic features may not be responsible for most of the changes in LMR in the course of evolution, at least at short timescales. Nevertheless, we can still measure the correlation between the genomic features and the LMR defined in another species, and trace how this correlation changes with distance between species, as a proxy for the rate of evolution of DNA landscape. A decrease in correlation between the LMR of one species and DNA properties of another species mirrors changes in genomic features with time.

Using this approach, we show that features differ in their contributions to the mutation rate dynamics. The recombination rate is known to be one of the key features that influence the mutation rate variation in humans (Duret and Arndt 2008; Duret and Galtier 2009; Pratto et al. 2014; Williams et al. 2015). Furthermore, local recombination rate is very plastic, and its hotspots vary dramatically even between closely related species (Coop and Myers 2007; Pessia et al. 2012); the correlation of the recombination rates is low even between human and chimpanzee. Our results show that changes in the recombination rate are among the biggest contributors to the LMR evolution among the studied genomic features.

The high similarity of LMRs between closely related species implies that the estimates of the neutral mutation rate at a genomic region could be substantially improved by considering the mutation rate at homologous regions from closely related species. In contrast, the functional annotation may be more informative about the mutation rate in a more distantly related species.

More generally, existing approaches to inference of functional genomic elements often use interspecies conservation as a proxy for function. This involves two assumptions about the mutation process: first, that the mutation rate is uniform along the genome; second, that it is constant between species. Violations of these assumptions can lead to false inferences. The first assumption is now being relaxed (Hodgkinson and Eyre-Walker 2011; Martincorena et al. 2012; Lawrence et al. 2013); however, the second largely remains in place. Appreciation of the importance of the variation in the genomic mutation rates in the course of evolution has revolutionized the field of molecular dating (Hodgkinson and Eyre-Walker 2011; Bouckaert et al. 2014). Analogously, we propose that understanding the LMR variation between species and its causes may help predict the likelihood of mutations and infer their functional importance.

Materials and Methods

Alignments

The multiple sequence alignment of eight primate genomes (chimpanzee, gorilla, orangutan, gibbon, rhesus, green monkey, marmoset, and squirrel monkey) with the hg19 version of the human genome assembly was obtained from the multiple alignment of 100 vertebrate species downloaded from the UCSC Genome Browser (https://genome.ucsc.edu/; last accessed February 3, 2017) (Karolchik et al. 2014) using multiz-tba (Blanchette et al. 2004). The alignment was then split into non-overlapping 1 Mb or 100 Kb windows. Exonic nucleotides, UTRs, repeats, ambiguous nucleotides, and CpG dinucleotides were masked, and windows with >20% gaps and masked nucleotides in any of the nine species were excluded from further analysis. This procedure resulted in 2,261 1 Mb windows, or in 23,551 100 Kb windows. Independently, multiple alignments of the same nine genomes with the mouse genome were obtained in the same way. Since there are more gaps in the primates–mouse alignment than in the primates-only alignment we excluded windows with >10% gaps and masked nucleotides in any of the species. This procedure resulted in fewer windows: 1,454 1 Mb windows, and 16,449 100 Kb windows.

To make sure that our results are not affected by differences in the fraction of excluded (unaligned or masked) nucleotides between genomic regions, we repeated the analyses using only those alignment columns where all species had an aligned and unmasked nucleotide, and only those windows

where there were >10% in the primates-mouse alignment. This resulted in 1,091 1 Mb windows for the primates-and-mouse alignment.

Genomic Features Mapping

Replication time, DHSs, methylation, mappability, Ctcf, PolII, and histone modifications as measured by the ENCODE project (Consortium TEP 2012) were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/ENCODE/; last accessed February 3, 2017). The analysis in the main text uses these maps obtained for embryonic stem cells. Additionally, we used the maps for five other tissues: GM12878, HUVEC, NHEK, Hela-S3, K562, and supplementary figure S9 and table S1, Supplementary Material online. Recombination rates were obtained from the HapMap project (ftp://ftp.ncbi.nlm.nih.gov/hapmap/; last accessed February 3, 2017). gBGC tracts obtained using phastBias with the parameter B = 3 were taken from ref. (Capra et al. 2013). Nucleosome occupancy scores were obtained from (Yazdi et al. 2015). Data of evolutionary conservation were taken from (Lindblad-Toh et al. 2011). Map of human topological domains of ES cell were taken from the Hi-C experiment of (Dixon et al. 2012). Repeats refer to the map of simple repeats from the UCSC Genome browser. The value of a feature for a genomic window was calculated as the weighted average of this feature, excluding masked nucleotides and gaps. MAFs and polymorphism data were obtained for all human SNPs except those that were W↔S from the 1,000 genomes project (McVean 2012). Mean values of MAFs for each window were calculated. pLMRs were calculated as the polymorphism' frequencies by utilizing only 50% of the rarest SNPs. The list of de novo mutations was obtained from Wong et al. (2016).

Chimpanzee and mouse recombination rates were obtained from Auton et al. (2012) and Brunschwig et al. (2012), respectively. Mouse DHSs, replication timing, and histone modifications H3K9me3, H3K27ac, and H3K27me3 measured by the Mouse ENCODE project for ES cells (mouse genome version mm9) were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/ENCODE/), and mouse genomic coordinates were converted into hg19 coordinates using liftOver (Karolchik et al. 2014).

The fraction of exonic nucleotides was calculated for all nucleotides (including masked ones) at each genomic window. To compare changes between features in the course of evolution, the values of each feature for each species were normalized to mean = 0 and variance = 1.

Inference of LMR

As a proxy for the LMR in a genomic window, we inferred the number of single-nucleotide substitutions that occurred in this window at a given phylogenetic branch (red in fig. 1*E* and *F*). For this, we inferred the ancestral states using maximum parsimony, and measured the fraction of substituted nucleotides among all unmasked non-gapped nucleotides. These fractions for all genomic windows were defined as LMRs normalized to mean = 0 and variance = 1, and LMR was defined as this normalized value. To avoid the confounding effect of GC-biased gene conversion (gBGC), we excluded W↔S substitutions (supplementary text S3, Supplementary Material online). The analyses including the W↔S substitutions are presented in the supplementary figures S10–S13, Supplementary Material online. As an alternative approach, we also utilized the maximum likelihood as implemented in the baseml program of the PAML package (Yang 2007), using the REV model with no molecular clock; here, all substitutions, including W↔S, were analyzed, and the results were similar (supplementary fig. S10, Supplementary Material online).

We assessed the accuracy and power of LMR inference by bootstrapping nucleotide sites within each window. The observed dLMRs and pLMRs were very strongly correlated with the bootstrapped samples (dLMR: $R = 0.97$, $P < 2.2 \times 10^{-16}$, pLMR: $R = 0.98$, $P < 2.2 \times 10^{-16}$), implying that these estimates are robust. The correlation was weaker for mLMRs ($R = 0.75$, $P < 2.2 \times 10^{-16}$; supplementary fig. S14, Supplementary Material online).

Since mutation rates differ between nucleotides (Lynch 2010), our LMR estimates as described above may be confounded by the differences in the nucleotide composition of genomic windows. To address this, we additionally used an alternative procedure for LMR estimation that accounts for the nucleotide composition. We calculated, for each species b and genomic window h, the expected number of mutations M based on its nucleotide composition:

$$M_{h,b} = \mu_b(\mathrm{G} \leftrightarrow \mathrm{C})n_{h,b}(\mathrm{G}+\mathrm{C}) + \mu_b(\mathrm{A} \leftrightarrow \mathrm{T})n_{h,b}(\mathrm{A}+\mathrm{T}),$$

where $\mu_b\mu_b$ is the genomic rate of the corresponding mutation in species b; and $n_{h,b}$ is the number of corresponding nucleotides in this window in species b. We then defined LMR as the ratio of the observed and expected numbers of mutations. This procedure yielded very similar results to those in the main text (supplementary figs. S15 and S16, Supplementary Material online).

Explained Variance of the LMR

Using linear regression as implemented in the lm function in *R*, we calculated the fraction of the variance in LMR between genomic windows in a species that can be explained by genomic features in the same and/or different species, and/or by the human LMR. Adjusted R^2 values were used to minimize the effect of the number of explanatory variables. To calculate the contribution of each feature independently of the contributions of other features, we performed ANOVA type III tests using drop1 function in *R*. 95% CIs for R^2 values were obtained by bootstrapping genomic windows in 200 bootstrap trials. Local polynomial regression fitting in figure 1 was performed using the loess function of *R*.

Correlating Differences between LMR and Genomic Features

For each window, we calculated the difference in normalized LMR values between human and mouse or chimpanzee, and the differences in normalized feature values between the same two species. We then calculated Pearson's correlation coefficients between these values. We also performed ANOVA type III tests to estimate the independent

contributions of changes in different genomic features to changes in the mutation rate, explaining changes in LMR between species with the drop1 function corresponding to changes in each feature separately. 95% CIs for R^2 values were obtained by bootstrapping genomic windows in 200 bootstrap trials.

Correlations with Phylogenetic Distance

To estimate the significance of the correlation between the phylogenetic distance and the R^2 values, we obtained the distribution of Spearman correlation coefficients by bootstrapping genomic windows in 10,000 bootstrapping trials.

Identification of HARs, HDRs, CARs, and CDRs

We aimed to select the genomic windows such that the human (chimpanzee) LMR was substantially increased or decreased, compared with the chimpanzee (respectively, human) LMR. In this section, raw LMR values were used, that is, normalization to mean = 0 and variance = 1 was not applied. First, we predicted the expected LMR $u_{b,h}(\exp)$ in phylogenetic branch b for each genomic window h, accounting for the mean LMR of this window across all branches $u_h = \frac{\sum_i u_{i,h}}{i}$ and the length of the branch leading to this species $v_b = \frac{\sum_j u_{b,j}}{j}$, and normalizing by the mean LMR across all windows in all species:

$$u_{b,h}(\exp) = \frac{\sum_i u_{i,h} \sum_j u_{b,j}}{\sum_i \sum_j u_{i,j}}.$$

We compared this value for human or chimpanzee with the corresponding observed value of LMR, $u_{b,h}(\mathrm{obs})$. To single out the genomic windows with LMR changes in the species of interest, we then selected all windows where the magnitude of change in this species sp_1 (human or chimpanzee) was greater than the magnitude of change in the sister species sp_2 (respectively, chimpanzee or human):

$$\left| lg \frac{u_{sp_1,h}(\mathrm{obs})}{u_{sp_1,h}(\exp)} \right| > \left| lg \frac{u_{sp_2,h}(\mathrm{obs})}{u_{sp_2,h}(\exp)} \right|.$$

We ranked these windows by the magnitude of change in LMR in sp_1 compared with sp_2,

$$\Delta_h = \frac{u_{sp_1,h}(\mathrm{obs})}{u_{sp_1,h}(\exp)} \Big/ \frac{u_{sp_2,h}(\mathrm{obs})}{u_{sp_2,h}(\exp)},$$

and identified accelerated regions (ARs) and decelerated regions (DRs) as the 100 windows with the highest and lowest values of Δ_h. Summary statistics describing the properties of ARs and DRs are presented in supplementary table S2, Supplementary Material online.

For all density plots, we used the kernel density estimation implemented in ggplots R-package with the default optimized bandwidth of smoothing (bw.nrd0). The scores for recombination and gBGC for all genomic windows were normalized to mean = 0 and variance = 1.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Author Contributions

N.V.T., V.B.S., R.A.S., and G.A.B. planned and designed the experiments, V.B.S. and G.A.B. directed the research, N.V.T. analyzed the data, and R.A.S. assisted in statistics and performed the power test analysis. All authors contributed to the preparation of the article.

Acknowledgments

We thank the members of Alexey Kondrashov's, Georgii Bazykin's, and Shamil Sunyaev's labs for discussion and helpful comments on the article. This work was supported by the Russian Foundation for Basic Research grant no. 15-34-21135-mol_a_ved to G.A.B. and by the Molecular and Cellular Biology Program of the Russian Academy of Sciences.

References

Ananda G, Chiaromonte F, Makova K. 2011. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biol* 12:1–18.

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc Natl Acad Sci U S A* 112:2109–2114.

Asthana S, Schmidt S, Sunyaev S. 2005. A limited role for balancing selection. *Trends Genet* 21:30–32.

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, Leffler EM, Bowden R, Aneas I, Broxholme J, et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. *Science* 336:193–198.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14:708–715.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 10:e1003537.

Brunschwig H, Levi L, Ben-David E, Williams RW, Yakir B, Shifman S. 2012. Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics* 191:757–764.

Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011. Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics* 187:1225–1234.

Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. 2013. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet* 9:e1003684.

Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* 513:195–201.

Charlesworth D. 2010. Don't forget the ancestral polymorphisms. *Heredity* 105:509–510.

Consortium TEP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Coop G, Myers SR. 2007. Live hot, die young: transmission distortion in recombination hotspots. *PLoS Genet* 3:e35.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485:376–380.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc R Soc Lond B Biol Sci* 279:3491–3500.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* 4:e1000071.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* 10:285–311.
Eyre-Walker YC, Eyre-Walker A. 2014. The role of mutation rate variation and genetic diversity in the architecture of human disease. *PLOS One* 9:e90166.
Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, Genome of the Netherlands C, van Duijn CM, Swertz M, Wijmenga C, et al. 2015. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* 47:822–826.
Gaffney DJ, Keightley PD. 2005. The scale of mutational variation in the murid genome. *Genome Res* 15:1086–1094.
Glemin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. *Genome Res* 25:1215–1228.
Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective population size within a genome. *Genetics* 189:1389–1402.
Harris K, Pritchard J. 2016. Rapid evolution of the human mutation spectrum. *bioRxiv*. doi: https://doi.org/10.1101/084343.
Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res* 21:349–356.
Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* 12:756–766.
Imamura H, Karro J, Chuang J. 2009. Weak preservation of local neutral substitution rates across mammalian genomes. *BMC Evol Biol* 9:89.
Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 42:D764–D770.
Kuruppumullage Don P, Ananda G, Chiaromonte F, Makova KD. 2013. Segmenting the human genome based on states of neutral genetic divergence. *Proc Natl Acad Sci USA* 110:14699–14704.
Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499:214–218.
Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339:1578–1582.
Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* 18:337–340.
Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478:476–482.
Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* 107:961–968.
Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* 17:704–714.
Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538:201–206.
Martincorena I, Seshasayee ASN, Luscombe NM. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485:95–98.
Mathieson I, Reich DE. 2016. Variation in mutation rates among human populations. *bioRxiv* doi: 10.1101/063578
McVean G. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.
McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* 5:e1000471.
Michaelson Jacob J, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al. 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* 151:1431–1442.
Nusbaum C, Mikkelsen TS, Zody MC, Asakawa S, Taudien S, Garber M, Kodira CD, Schueler MG, Shimizu A, Whittaker CA, et al. 2006. DNA sequence and analysis of human chromosome 8. *Nature* 439:331–335.
Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol* 4:675–682.
Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. Recombination initiation maps of individual human genomes. *Science* 346:1256442.
Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* 10:e1004525.
Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483:169–175.
Schuster-Bockler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 488:504–507.
Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* 489:75–82.
Tyekucheva S, Makova K, Karro J, Hardison R, Miller W, Chiaromonte F. 2008. Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol* 9:R76.
Veltman JA, Brunner HG. 2012. De novo mutations in human genetic disease. *Nat Rev Genet* 13:565–575.
Webster MT, Hurst LD. 2012. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet* 28:101–109.
Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G, Patterson N, Myers SR, Curran JE, Duggirala R, et al. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* e04637.
Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, Baker R, Thach DC, Iyer RK, Vockley JG, Niederhuber JE. 2016. New observations on maternal age effect on germline de novo mutations. *Nat Commun* 7:10486.
Woo YH, Li W-H. 2012. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat Commun* 3:1004.
Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen J-Q, Hurst LD, Tian D. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* 523:463–467.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
Yazdi PG, Pedersen BA, Taylor JF, Khattab OS, Chen Y-H, Chen Y, Jacobsen SE, Wang PH. 2015. Increasing nucleosome occupancy is correlated with an increasing mutation rate so long as DNA repair machinery is intact. *PLoS One* 10:e0136574.
Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515:355–364.
Zhou X, Cain CE, Myrthil M, Lewellen N, Michelini K, Davenport ER, Stephens M, Pritchard JK, Gilad Y. 2014. Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol* 15:547.