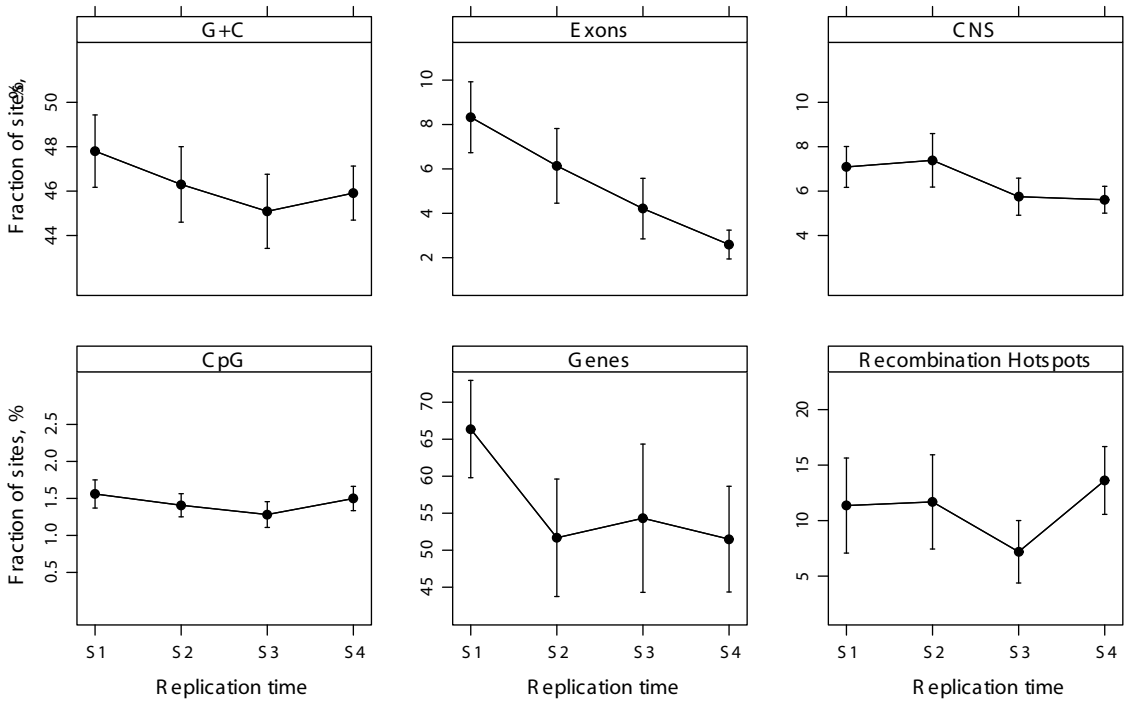
Stamatoyannopoulos JA, Adzhubey I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev S
Time of human DNA replication predicts mutation rate

SUPPLEMENTARY FIGURE S1

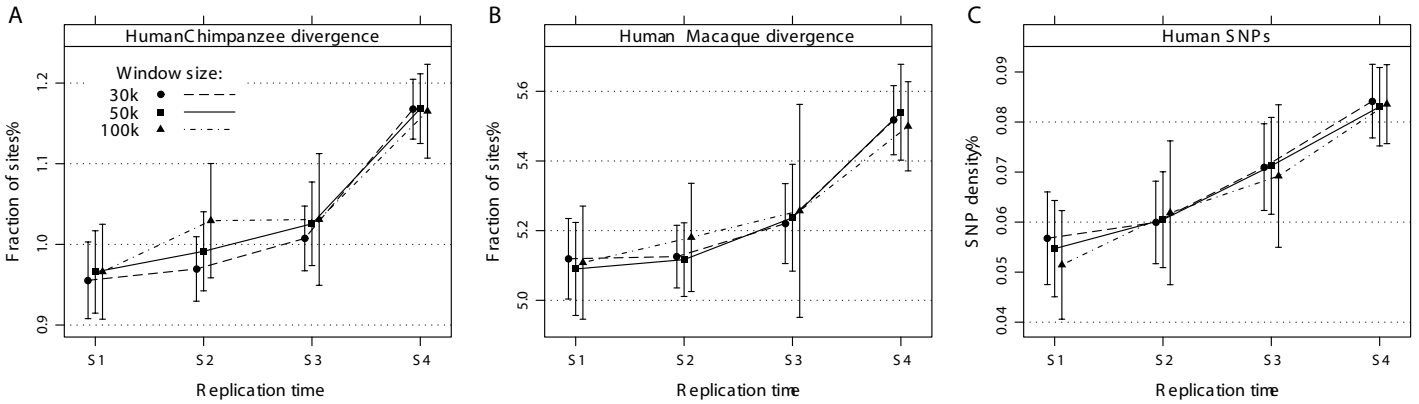


SUPPLEMENTARY FIGURE S1. Distribution of human genomic compositional features vs. time of replication. Shown are distributions vs. replication time of G+C and CpG putatively neutral sites; and densities of annotated exons (RefSeq), genes (RefSeq), conserved non-coding sequences (CNS) (1,2), and recombination hotspots (1,3).

References

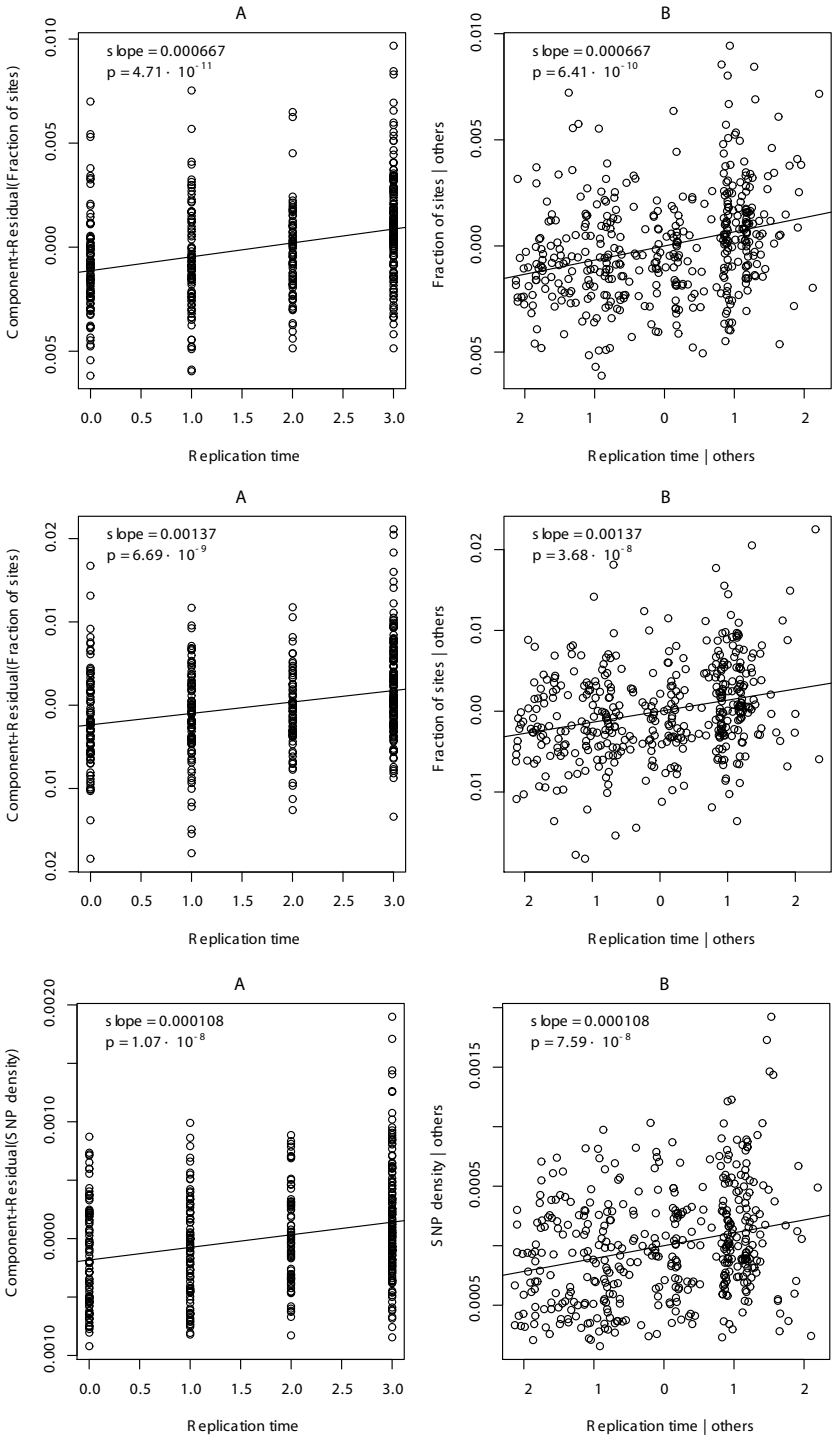
1. Karolchik, D. et al. The UCSC Genome Browser Database: 2008 Update. Nucleic Acids Res 36, D773-9 (2008).
2. Thomas, J.W. et al. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424, 788-93 (2003).
3. McVean, G.A. et al. The fine-scale structure of recombination rate variation in the human genome. Science 304, 581-4 (2004).

SUPPLEMENTARY FIGURE S2



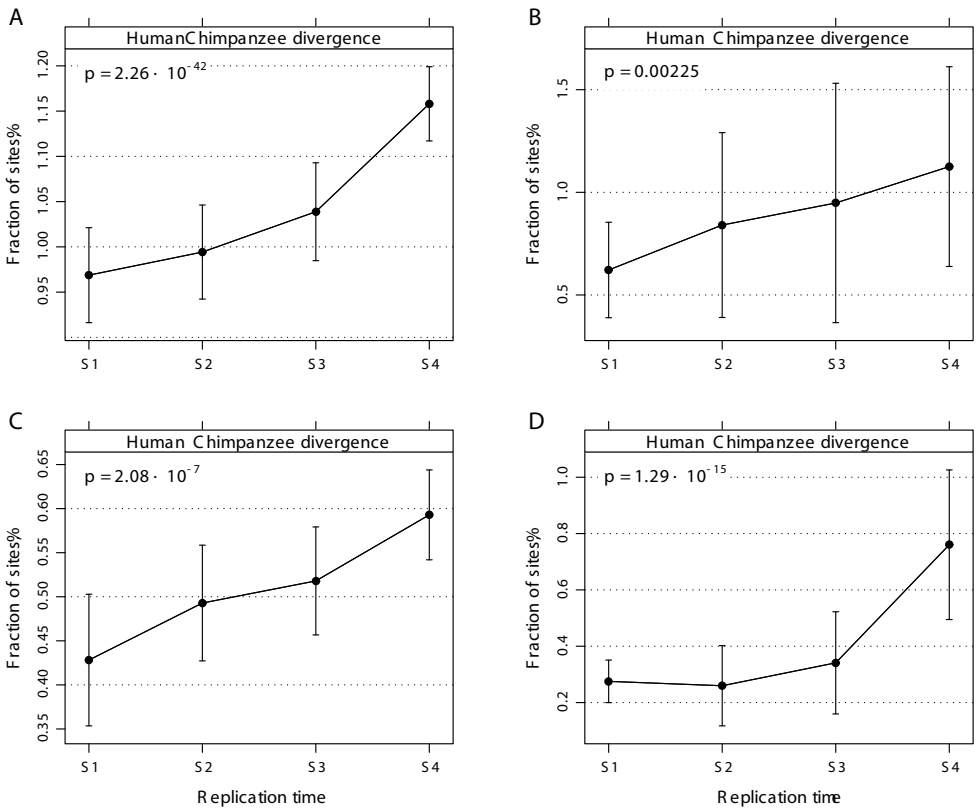
SUPPLEMENTARY FIGURE S2. Linear regression analysis of the partial model controlling for all predictors except replication time. Shown are partial residual (A) and partial regression (B) plots of partial regression model with the replication time predictor removed, fitted to human-chimpanzee divergence (top row), human-macaque divergence (middle row), and human SNP density (bottom row), non-CpG neutral sites. See Supplemental Table S2 for full regression model parameters. Lines on plots represent linear fit of residuals vs. replication state component. In all cases, the fitted lines demonstrate similar linear increasing trend indicating significant and consistent contribution of replication time to response (mutation rate) when simultaneously controlling for all other predictors in the mode

SUPPLEMENTARY FIGURE S3



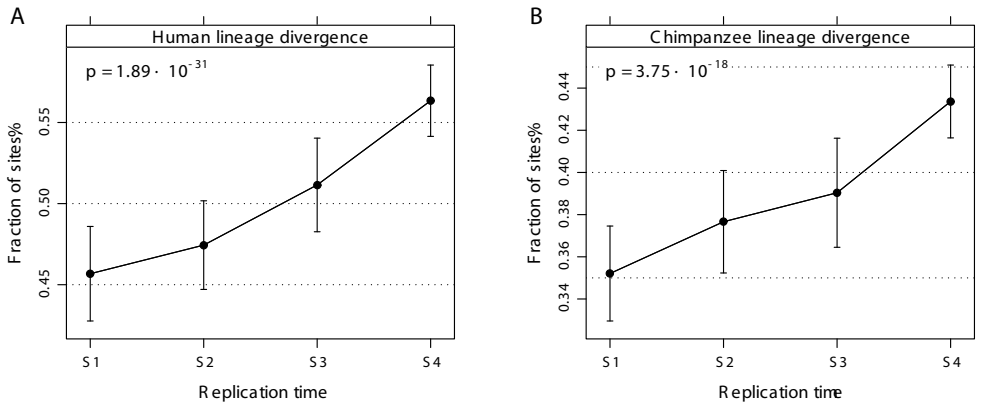
SUPPLEMENTARY FIGURE S3. Comparison of performance of different window sizes for sampling divergence and SNP density. Shown is dependence of human-chimpanzee divergence (A), human-macaque divergence (B), and human SNP density (C) on time of replication in putatively neutral non-CpG sites. Non-overlapping windows of 30kb, 50kb, and 100kb in size were tested in each case and revealed identical trends. Larger windows were not utilized due to the high resolution of the replication timing partitioning data (average segment size <75kb). Both 30kb and 50kb windows demonstrate similar performance but the former showed high levels of sampling errors when utilized for more sparse datasets (data not shown), while the use of 100kb window size result in inflated variance and decreased resolution.

SUPPLEMENTARY FIGURE S4



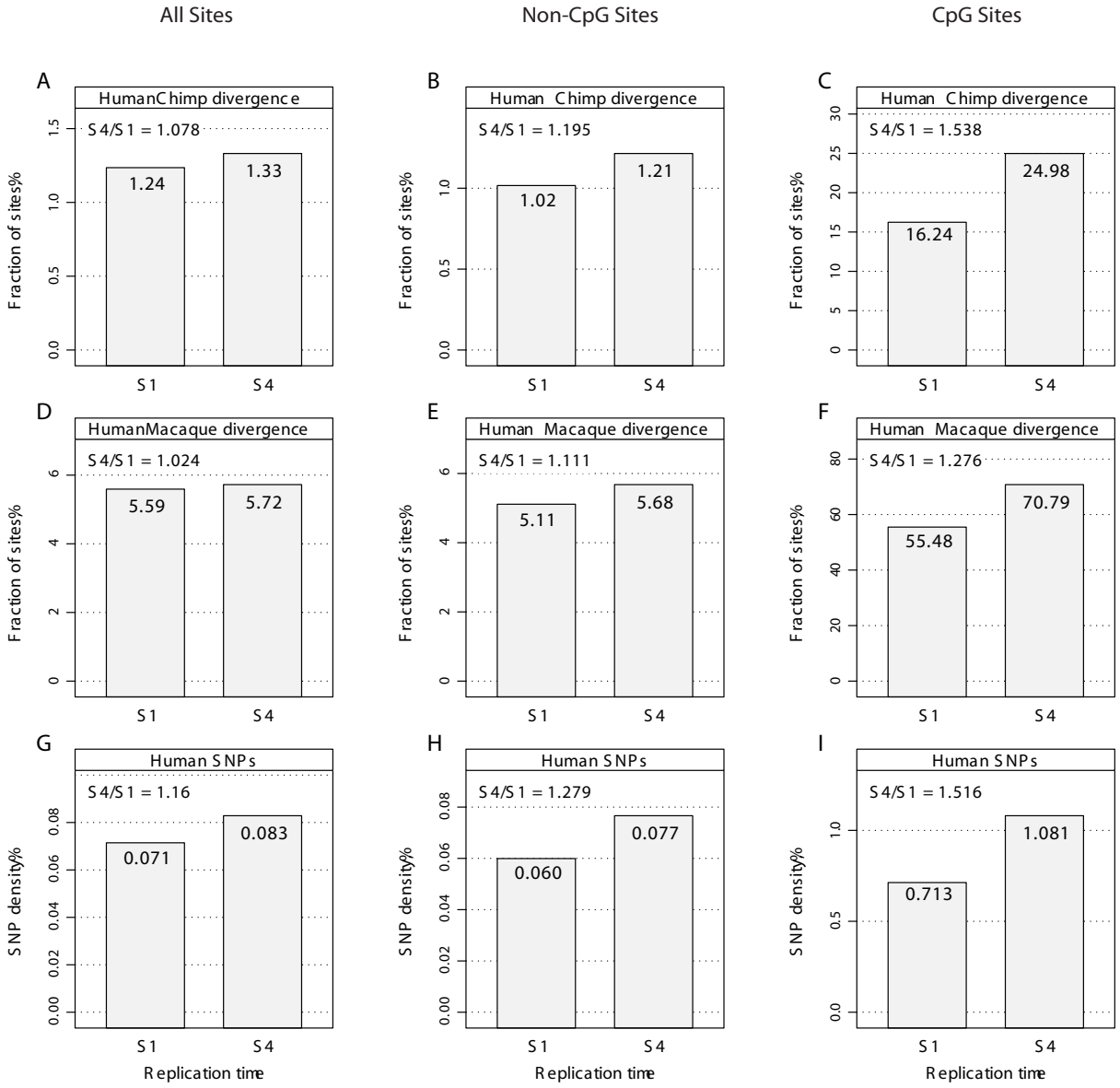
SUPPLEMENTARY FIGURE S4. Dependence of human-chimpanzee divergence on time of DNA replication at different classes of sites. Shown is dependence of human-chimpanzee divergence on time of replication in non-CpG sites annotated as ancestral repeats (A), coding 4-fold degenerate (B), conserved non-coding (C), and coding non-degenerate (D). The first two examples represent sites under relaxed selection while the latter two are assumed to be under strong selective pressure. In all cases, the same increasing trend is observed, with S4/S1 gain in substitution rates of 20%, 81%, 38%, and 176% for the four types of sites, respectively. Note: Results for coding 4-fold degenerate sites are unreliable due to the low frequency of mutations detected (total number of such substitutions = 196).

SUPPLEMENTARY FIGURE S5



SUPPLEMENTARY FIGURE S5. Dependence of lineage-specific divergence on time of DNA replication. Shown is a comparison between dependence of human-chimpanzee divergence on time of replication in the human lineage (A) and in the chimpanzee lineage (B). A three-way human-chimpanzee-macaque alignment was used to polarize and analyze separately substitutions in the human and chimpanzee lineages. Non-parsimonious substitutions were discarded and only non-CpG putatively neutral sites were utilized (see Fig. 1). The increase in substitution rate between S4/S1 temporal replication states equals 23% in the case of both human and chimpanzee lineages ($p < 3.75 \times 10\text{-}18$).

SUPPLEMENTARY FIGURE S6.

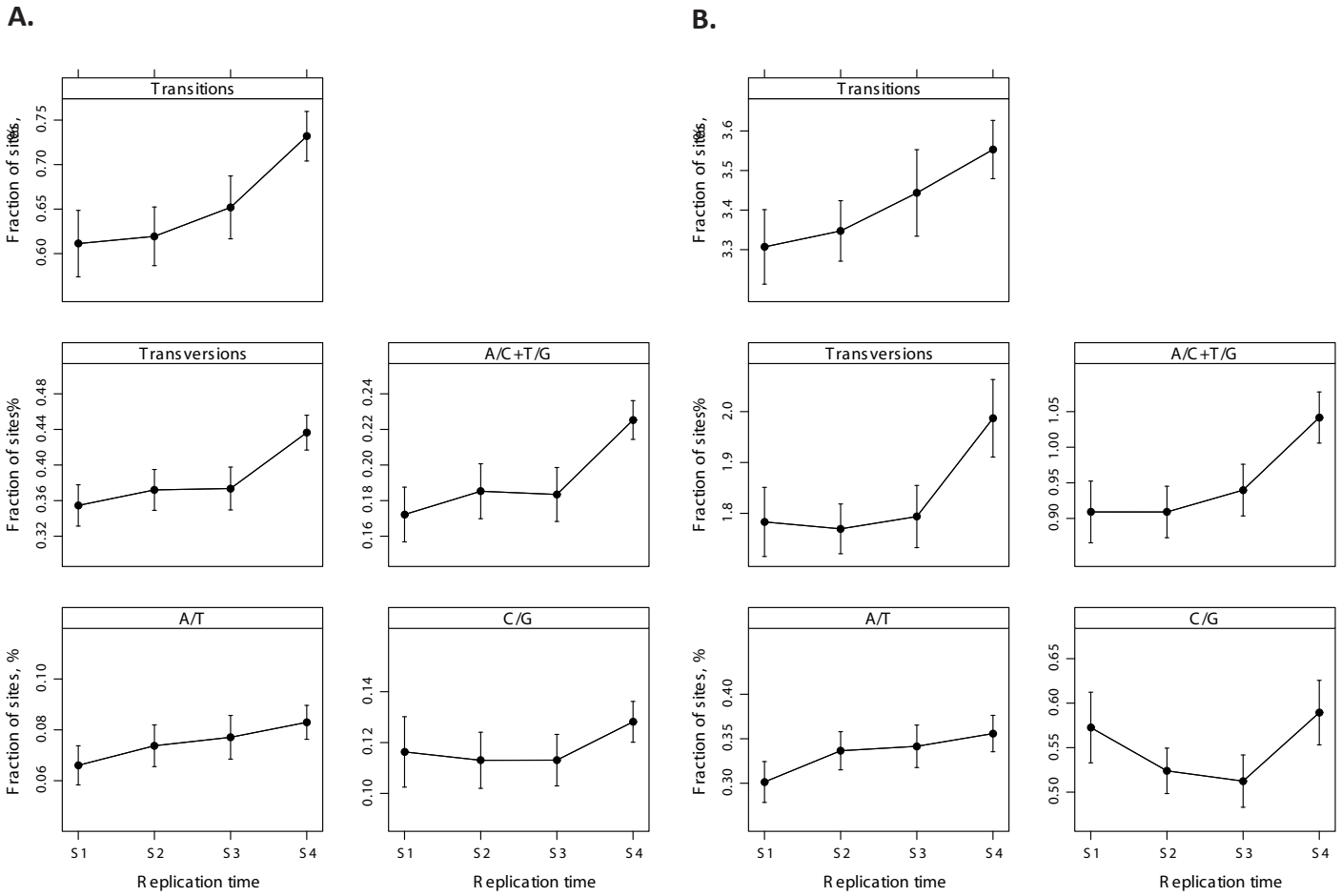


SUPPLEMENTARY FIGURE S6. Replication time-dependence of divergence and SNP density genome-wide. Shown are human-chimpanzee divergence (A-C), human-macaque divergence (D-F), and human SNP density (G-I) computed for all putatively neutral sites (left column), non-CpG neutral sites (middle column), and neutral CpG sites (right column) using genome-wide replication timing (early vs. late) data (1). For the purpose of the analysis, 20% of the lowest scoring (ratio < 1.24) and 20% of the highest scoring (ratio > 1.66) segments from 1Mb human genome dataset4 were selected and designated as S4 (late replicating) and S1 (early replicating), correspondingly. The fraction of sites in each was calculated as described in Fig. 1 but with the nucleotide counts pooled without window sampling. The estimated increase in mutation rates between the S1 and S4 temporal replication states (see panel legends) was highly significant at $p < 2.2 \times 10^{-16}$ (Chi-square) in all cases. The relative increase in rates of polymorphism and divergence between replication states is lower for all sites combined than for CpG-sites and non-CpG prone sites separately. This is not because of the weaker effect in CpG-prone sites that do not have CpG in either genome. Rather, it is a simple consequence of the reduced fraction of CpG sites (and CpG-prone sites) in late replicating regions compared to early replicating regions. Due to much higher rate of mutations in CpG sites, this reduction has a larger effect on the numerator than on the denominator of the ratio.

References

1. Woodfine, K. et al. Replication timing of the human genome. Hum Mol Genet 13, 191-202 (2004).

SUPPLEMENTARY FIGURE S7



SUPPLEMENTARY FIGURE S7. Replication time distribution of pooled transitions, pooled transversions, and individual mutation types in putatively neutral sites. For human-chimpanzee (A) and human-macaque (B) divergence, shown are distributions vs. replication time of pooled transitions, pooled transversions, and transversions that can be enumerated individually.SUPPLEMENTARY FIGURE S7

SUPPLEMENTARY TABLE S1

Replication time	ENCODE regions	Aligned	Neutral all			Neutral non-CpG			Neutral CpG		
			Sites	Subst.	SNPs	Sites	Subst.	SNPs	Sites	Subst.	SNPs
		Human-Chimpanzee alignment:									
S1	5,767,800	5,031,238	983,224	10,904	687	624,040	5,977	353	16,192	2,958	132
S2	6,726,550	5,994,908	1,673,947	18,598	1,140	1,082,257	10,674	665	24,208	4,769	217
S3	5,770,050	5,120,731	1,744,713	2,0076	1,432	1,131,742	11,732	839	24,395	5,080	284
S4	8,790,650	7,562,018	2,963,786	38,673	2,835	1,915,383	22,117	1,597	46,746	10,062	563
		Human-Macaque alignment:									
S1	5,767,800	3,498,478	717,155	38,066	-	457,124	23,203	-	12,404	7,426	-
S2	6,726,550	4,310,481	1,231,462	64,938	-	799,409	40,881	-	18,885	11,734	-
S3	5,770,050	3,606,482	1,289,537	69,395	-	839,306	43,930	-	19,371	12,392	-
S4	8,790,650	5,274,039	2,156,487	122,054	-	1,400,788	76,549	-	35,929	22,511	-

SUPPLEMENTARY TABLE S1. Distribution of raw frequencies (counts) of nucleotides found in each replication timing state within the set of regions analyzed (ENCODE).

Aligned sites in two-way alignments are shown under '**Aligned'**, and the putatively neutral subset of aligned sites under '**Sites**' (see text). Counts are shown for all putatively neutral sites under '**Neutral all';** neutral sites with CpG-prone dinucleotides removed under '**Neutral non-CpG';** and neutral CpG dinucleotides under '**Neutral CpG'** (see **Fig. 1**). Also included are counts of substitutions between two species ('**Subst.'**) and enumerated human SNPs ('**SNPs'**) within the corresponding sets of putatively neutral sites.

Nature Genetics: doi:10.1038/ng.363

SUPPLEMENTARY TABLE S2

Predictor	*t*-Value[a]	*p*-Value[b]	VIF[c]	Variability Explained[d]	Partial R^2
	Human-Chimpanzee divergence				
Replication time	6.54	$1.80 \cdot 10^{-10}$	1.12	0.40	0.083
CNS	-5.23	$2.69 \cdot 10^{-7}$	1.24	0.30	0.053
Genes	-2.86	0.0045	1.13	0.08	0.016
Exons	2.11	0.036	1.52	0.04	0.0086
Recombination hotspots	1.92	0.056	1.09	0.04	0.0071
G+C content	-0.16	0.87	1.19	0.0002	$4.88 \cdot 10^{-5}$
Multiple R^2					0.20
Adjusted R^2					0.19
	Human-Macaque divergence				
CNS	-6.55	$1.73 \cdot 10^{-10}$	1.20	0.45	0.085
Replication time	5.83	$1.13 \cdot 10^{-8}$	1.10	0.35	0.067
Exons	3.62	$3.3 \cdot 10^{-4}$	1.47	0.14	0.026
Recombination hotspots	2.45	0.015	1.09	0.06	0.012
G+C content	-2.28	0.023	1.19	0.05	0.010
Genes	-1.23	0.22	1.14	0.02	0.003
Multiple R^2					0.19
Adjusted R^2					0.18
	Human SNP density				
Replication time	5.60	$3.87 \cdot 10^{-8}$	1.12	0.70	0.068
Exons	2.58	0.010	1.51	0.10	0.015
CNS	-2.50	0.013	1.24	0.10	0.014
Genes	-2.08	0.038	1.12	0.10	0.009
G+C content	-0.59	0.55	1.19	0.008	$7.6 \cdot 10^{-4}$
Recombination hotspots	-0.18	0.86	1.09	$7.0 \cdot 10^{-4}$	$7.2 \cdot 10^{-5}$
Multiple R^2					0.097
Adjusted R^2					0.084

SUPPLEMENTARY TABLE S2. Linear regression analysis of divergence and SNP density. Shown are linear regression analyses of human-chimpanzee divergence, human-macaque divergence, and human SNP density distribution in a set of putatively neutral non-CpG sites utilizing a model with replication time together with five other genomic features as predictors, excluding CpG content (see **TABLE 1**). Notes: [a]Value of *t*-statistic; [b]*p*-Value of testing null hypothesis that a predictor's coefficient is equal to zero; [c]Variance inflation factor[*]; [d]Received contribution of a predictor to variability explained by the full model, RCVE[**]. For methods details see (1) and (2) below.

(1) Davis, C.E., Hyde, J.E., Bangdiwala, S.I. & Nelson, J.J. An example of dependencies among variables in a conditional logistic regression. In: Modern Statistical Methods in Chronic Disease Epidemiology, Eds. S.H. Moolgavkar & R.L. Prentice, pp. 140-147. New York: Wiley, 1986;

(2) Kvikstad, E.M., Tyekucheva, S, Chiaromonte F & Makova, K.D. A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Comput Biol* **3**, 1772-82 (2007).