Human mutation rate associated with DNA replication timing

John A Stamatoyannopoulos1,4, Ivan Adzhubei2,4, Robert E Thurman1, Gregory V Kryukov2, Sergei M Mirkin3 & Shamil R Sunyaev2

Eukaryotic DNA replication is highly stratified, with different genomic regions shown to replicate at characteristic times during S phase. Here we observe that mutation rate, as reflected in recent evolutionary divergence and human nucleotide diversity, is markedly increased in later-replicating regions of the human genome. All classes of substitutions are affected, suggesting a generalized mechanism involving replication time-dependent DNA damage. This correlation between mutation rate and regionally stratified replication timing may have substantial evolutionary implications.

Evolutionary divergence and inferred mutation rates are known to vary across the human genome1–3, and it has long been speculated that this is a consequence of covariance with an epigenetic feature1,2. In human cells, the temporal order of DNA replication during an S phase lasting approximately 10 h shows marked regional variability across the genome4,5. To parallel the conventional division of S phase into four sequential temporal states (S1–S4), we used a hidden Markov model6 to perform unbiased four-state partitioning of continuous, high-resolution replication timing measurements across 1% of the human genome7. We then determined human-chimpanzee sequence divergence rates and the density of well-ascertained SNPs8 at putatively neutral evolving sites within each temporal state, excluding any bases within annotated exons, repetitive elements, CpG islands, transcriptional start sites (± 2 kb), intronic splice sites and conserved noncoding sequence9 densities in sliding (250 kb) windows, with equivalent results (Supplementary Fig. 1). We binned each distribution into three classes (low, medium and high content), with an equal number of windows at each level and carried out separate tests for three-way interactions using each factor as a controlling variable (total 12 tests). All were highly significant, with P values not exceeding 3.0 × 10−12 (Table 1), as were repeated tests with the additional permutation resampling of temporal states (P < 5.0 × 10−6 for divergence; P < 2.2 × 10−4 for SNPs; Table 1).

To address potential interplay between more than one variable, we developed multiple regression models of both divergence and diversity, confirming the independent effect of replication timing (Supplementary Tables 1 and 2 and Supplementary Fig. 2 online). These models suggest that replication time alone may explain 40–70% of the variability explained by the full model, and ~8% of overall variability in diversity and divergence. The observed correlation between rates of nucleotide change and replication timing is therefore highly unlikely to be caused by variation in G+C content or by a mutagenic effect of recombination. To rule out any hidden dependence on window size, we repeated all analyses conditioned on smaller (30 kb) and larger (100 kb) windows, with equivalent results (Supplementary Fig. 3 online). The direction of effect of replication timing on evolutionary divergence and SNP density is highly similar when other genomic features are controlled. These findings are compatible with a process that influences mutation rate, which should affect both diversity and divergence in a stable fashion over evolutionary time. Furthermore, the findings persist across the spectrum of selected sites, from ancestral repeats and fourfold degenerate sites to conserved noncoding sequences and nondegenerate coding sites (Supplementary Fig. 4 online), and across the human and chimpanzee lineages following the split from macaque (Supplementary Fig. 5 online). We note, however, that the effect on SNP density appears larger than that on evolutionary divergence.

We next considered whether the relationship with mutation rate and replication time could be explained by variation in another genomic feature for which replication timing might be acting as a surrogate. Regional variation in G+C content2,3 and, independently, recombination rate2,3 have been invoked as potential causes of human mutation rate variation. We therefore obtained the distribution of G+C content, CpGs, recombination hot spots9, and gene, exon and conserved noncoding sequence8 densities in sliding nonoverlapping 50-kb windows (approximating the size of chromosomal domains linked to replicons) across each temporal replication state (Supplementary Fig. 1 online). We binned each distribution into three classes (low, medium and high content), with an equal number of windows at each level and carried out separate tests for three-way interactions using each factor as a controlling variable (total 12 tests). Additionally, we repeated all analyses with an independent set of randomly ascertained SNPs (Celera individual A versus NCBI build 35), with nearly identical effect (P < 9.69 × 10−22).

We observed a pronounced trend relating the rate of evolutionary divergence and the density of human SNPs to the progress of DNA replication (Fig. 1). Human-chimpanzee substitutions and human SNP density increase 22% and 53%, respectively, during the temporal course of replication, both of which are highly statistically significant (P < 8.43 × 10−26; Cochran-Armitage; Fig. 1a–c,g–i). To rule out potential confounding by the overall low genome-wide rate of human-chimpanzee divergence, we also analyzed human-macaque divergence, with similar results (P < 2.7 × 10−54; Fig. 1d–f). We confirmed the absence of bias due to a sampling or stratification effect across different genomic regions by testing (Cochran-Mantel-Haenszel) for three-way interactions, treating region assignment as controlling variable (P < 7.2 × 10−12, P < 0.00026 for human-chimpanzee divergence and human SNPs, respectively). Additionally, we repeated all analyses with an independent set of randomly ascertained SNPs (Celera individual A versus NCBI build 35), with nearly identical effect (P < 9.69 × 10−22).

The direction of effect of replication timing on evolutionary divergence and SNP density is highly similar when other genomic features are controlled. These findings are compatible with a process that influences mutation rate, which should affect both diversity and divergence in a stable fashion over evolutionary time. Furthermore, the findings persist across the spectrum of selected sites, from ancestral repeats and fourfold degenerate sites to conserved noncoding sequences and nondegenerate coding sites (Supplementary Fig. 4 online), and across the human and chimpanzee lineages following the split from macaque (Supplementary Fig. 5 online). We note, however, that the effect on SNP density appears larger than that on evolutionary divergence.
transcription-coupled repair. To rule this out, we examined introns and intergenic regions separately, and found no significant difference in any parameter (data not shown).

Finally, we examined the possibility that the mutational effect might be restricted to the subset of the genome we analyzed. To test this, we examined a lower-resolution genome-wide dataset comprising early- and late-replicating regions mapped in lymphoblastoid cells. These data also support a mutational effect analogous with that reported above (Supplementary Fig. 6 online), confirming the generality of our observations.

What molecular mechanism might underlie a monotonic increase in mutation rate during S phase? One possibility is that late stages of DNA replication are associated with the slowing or stalling of replication forks owing to exhaustion of the dNTP pool or difficulty in negotiating heterochromatinized templates, with consequent accumulation of single-stranded DNA (ssDNA) regions. ssDNA is more susceptible to endogenous and environmental damage, and can potentiate mutagenesis directly or via triggering of intra-S-phase checkpoints that set in motion low-fidelity polymerases. Another possibility is that the mismatch repair system might erode during S phase or that lesions in late replicating regions simply lack adequate time to undergo effective repair.

To differentiate these scenarios, we examined mutations at CpG dinucleotides, which arise overwhelmingly from spontaneous deamination of methylcytosine into thymine, a process which escapes DNA mismatch repair. We found that both evolutionary divergence and human SNP density.

**Figure 1** Replication time-dependence of evolutionary divergence and human SNP density. (a–i) Shown are replication time-dependence of human–chimpanzee divergence (fraction sites changed) (a–c), human-macaque divergence (d–f) and human polymorphism rate (SNP density) (g–i), computed across 44 regions (ENCODE) comprising 1% of the human genome. Analyses are presented for all putatively neutral sites (a,d,g), for neutral sites with all CpG-prone dinucleotides removed (b,e,h) and for CpG dinucleotides only (c,f,i). CpG-prone dinucleotides were defined as all sites immediately preceded by C or followed by G in either species; and CpG dinucleotides, as sites with C immediately preceding G at least in one of the species. Plots show mutation rates averaged over all 50-kb nonoverlapping windows within each of four temporal replication states (S1–S4), together with 95% confidence intervals for the mean. Significance of each trend is shown within the corresponding panel as a $P$ value (Cochran-Armitage trend test for proportions). For divergence analysis, human-chimpanzee hg17 vs PanTro1 and human-macaque hg17 vs RheMac2 axtNet alignments were processed to exclude all regions with more than a single substitution per sliding 5-nucleotide window, and axtNet fragments which were either shorter than 500 bases or demonstrated average substitution rate above 3% (12% in case of macaque) were also discarded. For polymorphism analysis, we used version 1 bulk SNPs dataset published as part of the Personal Genome Sequence project (Watson).
Table 1 Significance of replication time-dependence of evolutionary divergence and human SNP density

<table>
<thead>
<tr>
<th></th>
<th>P value</th>
<th>G+C</th>
<th>Cpg</th>
<th>Exons</th>
<th>Genes</th>
<th>CNS</th>
<th>Recombination hot spots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Chimpanzee divergence</td>
<td>Stratification</td>
<td>$2.8 \times 10^{-47}$</td>
<td>$1.1 \times 10^{-81}$</td>
<td>$3.3 \times 10^{-43}$</td>
<td>$1.5 \times 10^{-49}$</td>
<td>$7.1 \times 10^{-44}$</td>
<td>$1.2 \times 10^{-43}$</td>
</tr>
<tr>
<td>Human SNP density</td>
<td>Stratification</td>
<td>$2.0 \times 10^{-13}$</td>
<td>$1.2 \times 10^{-22}$</td>
<td>$2.9 \times 10^{-13}$</td>
<td>$8.1 \times 10^{-14}$</td>
<td>$3.0 \times 10^{-12}$</td>
<td>$1.8 \times 10^{-13}$</td>
</tr>
<tr>
<td>Human-Macaque divergence</td>
<td>Stratification</td>
<td>$2.9 \times 10^{-30}$</td>
<td>$3.7 \times 10^{-43}$</td>
<td>$8.9 \times 10^{-29}$</td>
<td>$1.0 \times 10^{-30}$</td>
<td>$1.4 \times 10^{-28}$</td>
<td>$1.5 \times 10^{-28}$</td>
</tr>
</tbody>
</table>

Human-chimpanzee and human-macaque divergence and human SNP density were stratified by one of six controlling variables including G+C content, the number of CpGs; exons (RefSeq); gene units (RefSeq); conserved noncoding sequences (CNS); and recombination hot spots (see Supplementary Fig. 1). P values were computed by both stratification (generalized Cochran-Mantel-Haenszel) and permutation resampling (Monte-Carlo) approaches. The distribution of each compositional feature was produced by sliding a nonoverlapping 50-kb window across each of the four temporal replication states (S1–S4) and binning composition values obtained into three classes (low, medium and high content), with an equal number of pooled windows under each level. Cochran-Mantel-Haenszel tests for three-way interactions were then done using each compositional factor as a controlling variable. Monte-Carlo tests were carried out by running 200,000 random permutations of replication timing assignments for a set of generated 50-kb windows for each feature, followed by calculation of Cochran-Mantel-Haenszel statistic as described above.

and human nucleotide diversity at CpG sites (Fig. 1c,f.i) correlate with replication timing, closely paralleling other types of sites (Fig. 1a,b,d,e,g,h). The parallelism between CpG and non-CpG sites cannot be explained by alterations in the dNTP pool, nor by reduced polymerase fidelity, nor by defective mismatch repair. In addition, we found all classes of evolutionary transitions and transversions to show strong replication timing-dependence with a characteristically similar trend (Supplementary Fig. 7 online). This indicates that the effect is not due to biases in the genesis of specific mutational events nor to their handling by the repair machinery.

Our results therefore suggest that a simple consequence of the process of DNA replication—accumulation of single-stranded DNA within later replicating regions—may provide the most parsimonious explanation. Because ssDNA is highly susceptible to endogenous DNA damage, including alkylation, oxidation and deamination, it is reasonable to expect that there will be discrepancies between our calculations and those that might be made from germ cells were data available. The correlation reported herein should therefore be regarded as a lower limit estimate of actual dependence of mutation rate on replication timing.

Our results further suggest that replication timing is the dominant factor responsible for the reduced nucleotide diversity around exons, which may have a mechanistic or a selection-based explanation.

Although exons preferentially reside in early replicating regions with lower mutation rate, it is notable that a number of human genes controlling developmental fate, differentiation and cell proliferation reside in repressed, late-replicating chromatin in most adult cell types. This suggests that increased mutation rate affecting late replicating regions of the human genome may reflect a significant evolutionary cost for sequestering specific gene subsets within a repressed nuclear compartment.13

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank A. Kondrashov, M. Przeworski and J. Cameron for helpful discussions. This work was supported by US National Institutes of Health grants U54HG003042 and R01GM071852 to J.A.S.; R01GM078598, R01MH084676 and U54LM008748 to S.R.S. and R01GM06987 to S.M.M.

AUTHOR CONTRIBUTIONS


Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/