

RESEARCH ARTICLE

Long-Range Bidirectional Strand Asymmetries Originate at CpG Islands in the Human Genome

Paz Polak and Peter F. Arndt

Max Planck Institute for Molecular Genetics, Berlin, Germany

In the human genome, CpG islands (CGIs), which are GC- and CpG-rich sequences, are associated with transcription starting sites (TSSs); in addition, there is evidence that CGIs harbor origins of bidirectional replication (OBRs) and are preferred sites for heteroduplex formation during recombination. Transcription, replication, and recombination processes are known to induce specific mutational patterns in various genomes, and therefore, these patterns are expected to be found around CGIs. We use triple alignments of human, chimp, and macaque to compute the rates of nucleotide substitutions in up to 1 Mbps long intergenic regions on both sides of CGIs. Our analysis revealed that around a CGI there is an asymmetry between complementary substitution rates that is similar to the one that found around the OBR in bacteria. We hypothesize that these asymmetries are induced by differences in the replication of the leading and lagging strand and that a significant number of CGIs overlap OBRs. Within CGIs, we observed a mutational signature of GC-biased gene conversion that is associated with recombination. We suggest that recombination has played a major role in the creation of CGIs.

Introduction

Even though CpG islands (CGIs) have been suggested more than decade ago to colocalize with origins of replication (ORIs; Delgado et al. 1998), there has not been a study of the impact of this association on the mutational patterns on a genome-wide level. Probably the reason is that till recently only 30 ORIs were identified in human (Aladjem 2007), and it is not clear if the enrichment of CGIs at ORIs occurs at genome-wide scale. However, a recent large-scale survey could expand the number of characterized ORIs in the human genome about 10-fold, using micro arrays analyzing about 1% of the genome (Cadoret et al. 2008). This study has established the enrichment of CGIs within ORIs; over one-third of the newly identified ORIs overlap CGIs and over 50% of ORIs are in distance of less than 1 kbp from some CGIs in the genome (Cadoret et al. 2008). This link between CGIs and ORIs together with a mutational impact of replication should leave a footprint on substitution patterns along the genome.

The hallmarks of mutations associated with replication are a strand asymmetries in reverse complement mutation rates (Lobry 1996) as a result of the difference in the replication of two DNA strands, which are called leading and lagging strands. It has been suggested that the discontinuous replication of the lagging strand by the Okazaki fragments increases the frequency in which the template of the lagging strand is in single-strand (ss) DNA conformation relative to the template of the leading strand, which is replicated continuously (Frank and Lobry 1999). Because nucleotides on ssDNA are more prone to be damaged (in particular by adenine and cytosine deamination), this can induce differences in the mutational spectrum between the two strands (Frederico et al. 1990). Alternatively, the difference between the mutational spectra of the two DNA strands can be a result of replication of the two

DNA strands by two different polymerases (Kunkel and Burgers 2008) that have different error rates for insertion of nucleotides in the nascent strands (Kunkel and Bebenek 2000; McCulloch and Kunkel 2008). This polarity of replication process has been suggested to induce DNA base composition asymmetry in variety of genomes from prokaryotes (Lobry 1996) to vertebrate both mitochondrial (Faith and Pollock 2003) and nuclear DNA (Touchon et al. 2005).

In bacterial genomes, replication has been associated with violations of Chargaff's second parity rule. This rule states that if the mutational processes do not distinguish between the two DNA strands then there should be an equal frequency of complementary bases in a long piece of ssDNA, that is, $A = T$ and $C = G$ (Lobry 1995). Commonly used quantities to assess strand symmetries are the TA skew ($= (T - A)/(T + A)$) and GC skew ($= (G - C)/(G + C)$); deviations of TA and GC skews from zero imply that mutational processes distinguish between the two DNA strands (Lobry 1996). A positive GC skew and a nonzero TA skew are often used in bacteria to detect the leading strand (Frank and Lobry 1999). A switch in the sign of a skew is indicative for an origin of bidirectional replication (OBR) or the terminus of replication (Mrazek and Karlin 1998; Palleja et al. 2008; Touchon and Rocha 2008); the reason is that DNA polymerases can synthesize DNA only in the direction $5' \rightarrow 3'$ and the replication forks propagate to opposite directions from OBRs.

Although, the vast majority of bacterial genomes share similar skews, comparative genomic analysis revealed that the underlying single nucleotide substitution rates significantly vary between taxa. In each taxa, at least one symmetry of nucleotide substitution rates is broken; often, there is an excess of mutations $C \rightarrow T$, $A \rightarrow G$, and $C \rightarrow G$ on the leading strand when compared with the complementary mutations $G \rightarrow A$, $T \rightarrow C$, and $G \rightarrow C$, respectively. However, there is no one symmetry of nucleotides that is broken in all taxa (Rocha et al. 2006). Therefore, in order to understand the mutational processes coupled to replication in the current days, one has to estimate nucleotide composition skews as well as the mutation rates.

Evidence for strand asymmetries in mutation rates, which are associated with replication, has recently been

Key words: CpG islands, strand asymmetries, origin of bi-directional replication, recombination, biased gene conversion.

E-mail: polak@molgen.mpg.de.

Genome. Biol. Evol. 1(1):189–197. 2009

doi:10.1093/gbe/evp024

Advance Access publication August 3, 2009

© 2009 The Authors

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

found in human. Profiles of TA and GC skews have been used to identify putative ORIs along human chromosomes; Touchon et al. (2005) first analyzed six well-studied ORIs in the human genome and detected on the 5' → 3' strand centered at the ORI, a change from a negative GC and TA skew in the 5' flanking upstream to the ORI to positive skews at the downstream regions exactly as in *Escherichia coli*. By detecting similar changes in compositional skews along human chromosomes, 1,000 putative ORIs (among them 280 are OBRs) were identified in silico, 30-fold more than the number of ORIs that were known at that time (Huvet et al. 2007).

The observed TA and GC skews in the flanking regions of human ORIs imply that there are underlying strand-specific substitution patterns that are coupled to replication in the human genome. Interestingly, the impact of replication on nucleotide composition in the human genome is very similar to the one that induced by transcription process that has been shown to induce positive GC skews in transcripts which are similar to the one on the leading strand (Green et al. 2003; Aerts et al. 2004; Polak and Arndt 2008). As mentioned above, experiments have shown that CGIs are enriched near ORIs (Cadoret et al. 2008). Therefore, the enrichment of ORIs in CGIs is expected to induce strand asymmetries of mutations around CGIs and at the CGIs, there should be a change in the direction of the strand asymmetry. Because CGIs are also origin of transcription, strand asymmetries due to replication might be obstructed by the impact of transcription.

In this work, we measured the averaged substitution rates in the flanking regions of human CGIs. The estimation has been done along the human lineage (since the divergence from human chimp last common ancestor). In order to reduce the impact of transcription, we removed transcripts from the analyzed regions for our initial analysis. We have found that, on average, CGIs are origins of long-range bidirectional strand asymmetry. Along ssDNA centered at the CGI, the ratio between the rates of A → G and T → C is smaller than 1 on the 5' side of CGI and greater than 1 on the 3' side. The asymmetry profile implies that there is a significant number of CGIs that serve as OBR at a genome-wide level. To quantify the additional impact of transcription on mutational bias, we also measured the asymmetry in intronic regions of genes, which have a CGI in their promoter. The level of strand asymmetry is much higher in introns than in intergenic regions, implying that transcription has an additional effect. Within CGIs, the level of asymmetry between A → G versus T → C is lower than in its flanking regions. We further calculated the bias between the rates of substitution of weak bases (W = A or T) in strong bases (S = G or C). The ratio W → S/S → W within CGIs is higher than in the flanking regions. Interestingly, the measured W → S/S → W ratio in CGIs that overlap a transcription starting sites (TSSs) is lower than other regions. We also found that W → S/S → W ratio is positively correlated with recombination rates. This suggests that the enrichment of GC nucleotides in CGIs is due to biased gene conversion (BGC; Galtier and Duret 2007). Using the estimated substitution rates, we could calculate the expected GC content in the analyzed regions. We suggest that CGIs are often used as ORI and as a site for heteroduplex formation, in addition of being often associated with TSSs.

Materials and Methods

Substitution Analysis

In this study, we estimated 18 substitution rates, 12 single nucleotide rates (X → Y), and 6 additional context-dependent rates of substitutions of CpG di-nucleotides into TpG, CpA, ApG, CpT, CpC, and GpG. Substitution frequencies have been estimated from triple alignments of genomic sequences from human, chimp, and rhesus (see below). We computed the substitution rates in the branch from the last common ancestor of human and chimp toward the current day human. To do so, we implicitly reconstruct an ancestral sequence of the last common ancestor. This is done within our maximum likelihood (ML) framework (Duret and Arndt 2008) that assumes neither equilibrium nor reversibility of nucleotide substitution process (Squartini and Arndt 2008). Our ML approach correctly handles effects due to back mutations and is able to reliably estimate substitution frequencies from given aligned sequence (Duret and Arndt 2008).

Sequence Annotation and Multiple Alignments

Triple human–chimp–rhesus alignments were retrieved from the Ensembl database, version 50 from July 2008 (Flicek et al. 2008) using the Ensembl API. They are based on the releases *Homo sapiens* (50, 36c), *Pan troglodytes* (50, 2.1), and *Macaca mulatta* (50, 10), and they were generated using the Enredo Pecan Ortheus pipeline (Paten et al. 2008) for four catarrhini species. The annotation for genes, exons, and translatable exons is according to Ensembl version 50.

Classification of CGIs

We retrieved 21,353 CGIs from the Ensembl database, which defines CGIs using the following cutoffs; minimum length is 400 bps long; minimum GC content is 50%; and minimum of observed over expected CpG ratio of 0.6. We also wanted to check if there is a difference between CGIs that are or are not associated with transcription. Therefore, we classified CGIs in the human genome into two main classes according to their distance from the TSS. The first class of CGIs is associated with transcription; CGIs in this class harbor one or multiple TSSs but do not overlap with genes that are transcribed from different strands. This class is denoted by tCGI and includes 8,249 CGIs. The second class of CGIs encompasses 3,378 distal CGIs (denoted by dCGIs) that are intergenic and found at least 10 kbp away from any annotated TSS of a human gene (fig. 1). We also defined two additional subclasses of CGIs; proximal CGIs (pCGIs) that are intergenic CGI and are found at a distance of up to 10 kb from some annotated TSS in Ensembl and genic CGIs (gCGIs), which are CGIs that are completely contained in a gene.

Reference Strand for Substitution Analysis

For estimation of substitution rates, one has to choose one of the two DNA strands as a reference (fig. 1). For dCGIs, we analyzed the forward strand in the National Center

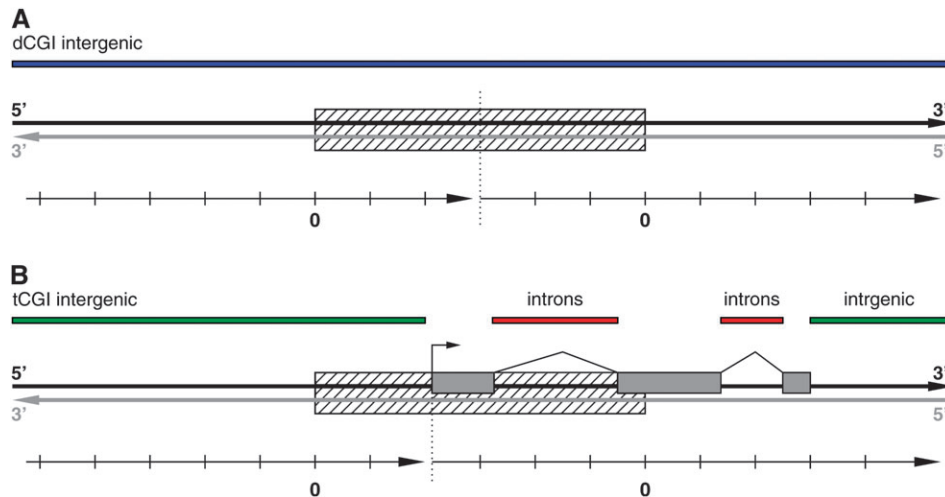


FIG. 1.—Sketch of the analyzed regions around and within two classes of CGIs: dCGIs (A) and tCGIs (B). CGIs are denoted by striped boxes. The bold strand is the reference strand that is used for the substitution analysis and for defining the directionality 5' → 3' relative to the CGI. The substitution rates are estimated relative to the 5' end (3' end) of CGI using a sliding-window analysis. The left (right) coordinate system described the distances of the windows from the 5' end (3' end), which is denoted by the left (right) origin (0 k). The left (right) coordinated system starts (ends) at the middle position to the next CGI upstream (downstream) to the 5' (3') end of CGI and ends (starts) in the dashed line. However, the analyzed regions in both sides of the CGI are restricted to up to 1 Mbp. (A) dCGIs are CGIs that are intergenic and found at distance of at least 10 kbps from a TSS. The reference strand is the NCBI forward strand, and the position of the dashed line is in the middle of the dCGI. (B) A tCGI is a CGI that harbors TSS of a transcript (exons are denoted by shaded areas). The reference strand is chosen to be the nontemplate (or coding) strand of this gene. The dashed line is coincides with the TSS. The colored bars indicate regions that were analyzed in figure 2: dCGI—intergenic regions (blue); tCGI—intergenic regions (green); and tCGI—introns (red) of genes whose TSS is inside of the tCGI.

for Biotechnology Information (NCBI) annotation because we cannot a priori distinguish the two DNA strands. For tCGI, the two DNA strands are distinguishable because on one strand a gene is transcribed (fig. 1). Therefore, for tCGI, we estimated the substitution rates with respect to the nontemplate (i.e., the not transcribed) strand of the associated gene.

CGI Centric View

The aim of our analysis is to derive the profile of nucleotide substitution rates around and within CGIs. In particular, we estimate the dependence of rates relative to the 5' and 3' end of CGIs on the reference strand (see fig. 1). The sequence for the analysis on the 5' side starts up to 1 Mbp upstream to the 5' end of the CGI and ends either at the middle point inside of the CGI for distal CGIs or at the TSS inside of tCGIs (see dashed line in fig. 1). In order to avoid the 2-fold analysis of nucleotides, the upstream region was truncated to the middle position between two CGIs, if the next CGI was closer than 2 Mbp. A similar procedure has been applied to determine sequences for the regional analysis surrounding the 3' end of CGIs (fig. 1).

Sliding-Window Analysis and Pooling

In order to get high resolution of the dependence of 18 nucleotide substitution rates in the distance from CGI, we wished to estimate the rates in sliding windows, of length 10 kbp each, along the flanking region of individual CGI (fig. 1). However, since the divergence in DNA sequence between human and chimp is about 1%, the estimated 18 rates in a single 10-kbp long DNA sequence are noisy. Therefore, in order to perform this kind of analysis, we had to pool data for all CGIs in a similar fashion as we

did before for TSSs (Polak and Arndt 2008). We estimated the 18 substitution rates in genome-wide pooled 10-kbp long nonoverlapping windows, which are located at fixed distances from individual CGIs up to distance of 1 Mbp for a CGI. The coordinate of a window in the 5' and 3' analyzed regions is the distance of center of the window to the 5' and 3' ends of CGIs, respectively (fig. 1). Such analysis is possible with the availability of genome-wide human–chimpanzee–macaque alignments that cover about 85% of the human genome. In the Supplementary Material online, we present an estimation of substitution rate in even higher resolution using 100 bps long windows along CGI and their 2.5-kbp flanking regions.

Examining the Transcriptional Impact

In our earlier study, we have found that transcription can induce strand asymmetries along genes (Polak and Arndt 2008). However, these strand asymmetries are not limited to transcripts, and they are found at the 5-kbp long flanking regions upstream to the TSS and downstream to the 3' end of human genes. In order to avoid possible effects of transcription on mutational asymmetries, we first masked out transcribed regions (UTRs, introns, and exons including their 5 kbps flanking regions) from our analysis, that is, we analyzed just the intergenic regions. However, at a later stage, we also estimated the rates in intronic sequences of genes with TSS inside of tCGI (see transcript in fig. 1).

Recombination Rates and Recombination Hot spots

We retrieved the genome-wide recombination rates and hot spots from the Phase 2 HapMap data set (HapMap 2007), which is estimated from phased haplotypes in

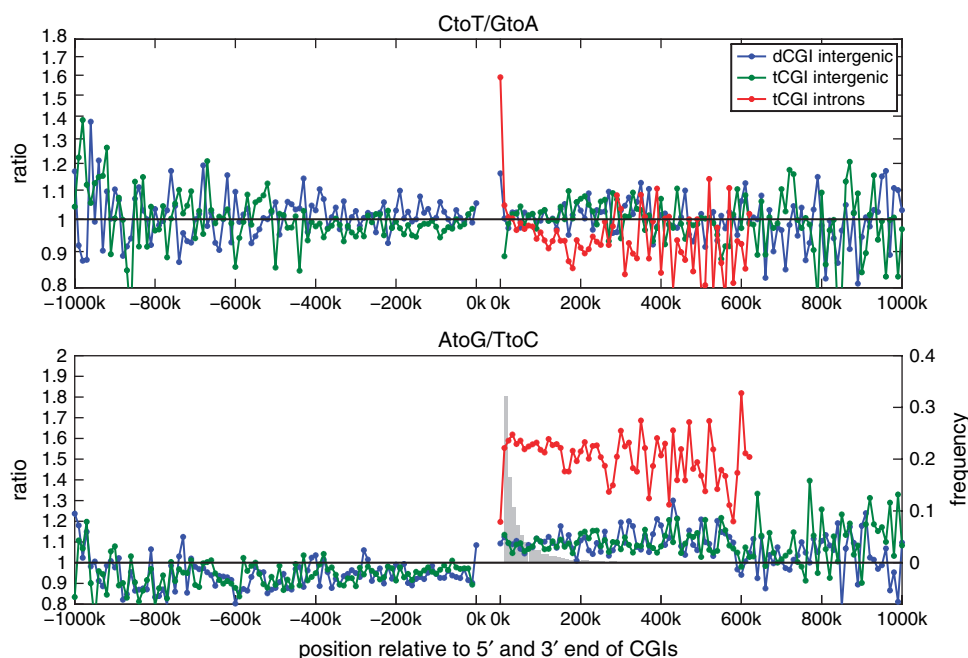


FIG. 2.—Ratios between complementary substitution rates in intergenic (blue, green) and intronic (red) regions. The ratios are plotted against the distance from the 5' end (left 0 k) and 3' end (right 0 k) of CGIs calculated in 10-kbp long windows. For dCGI and tCGIs, the analyzed sequences are intergenic (see corresponding blue and green bars in fig. 1) and are taken from the reference strand as it is described in figure 1. The analyzed intronic sequences are of genes that their TSS is located within tCGI (see red bars in fig. 1). The ratios in these regions are computed using the nontemplate strand of a gene (see fig. 1); intronic sequences are only available for the 3' side (left to the gap) analysis. The ratios at 0 k are calculated within the CGIs, for details, see supplementary figure 6 (Supplementary Material online). A shaded histogram of gene lengths in the human genome is presented at the bottom panel demonstrating that strand asymmetries between $A \rightarrow G$ versus $T \rightarrow C$ extend over distances larger than of a typical length of a gene.

HapMap Release 22 (NCBI 36). We computed the recombination rate for a genomic locus (e.g., individual CGI) by averaging the recombination rate along this region.

Results

Long-Range Bidirectional Asymmetries Originate in CGI

In this research, we derived the profile of 18 nucleotide substitution rates around CGI (see Materials and Methods). We adopted a CGI centered view (see fig. 1), where for each CGI we defined a region centered at the CGI that spreads from this CGI to the middle points to the next CGIs. These regions can be more than 1 Mbp long, but the average distance between two CGI is about 100 kbp. We estimated the substitution rates in intergenic regions in the 5' and 3' sides relative to CGIs that are associated with a transcription start site, tCGIs, and those that are at least 10 kbp away from a TSS, dCGIs (fig. 1, panel A and B). In order to measure the deviation of substitutions from strand symmetric case, we calculated the ratios of between pairs of complementary transition rates, that is, the ratio of the rates of $C \rightarrow T$ over $G \rightarrow A$ and the ratio of rates of $A \rightarrow G$ over $T \rightarrow C$ (fig. 2). In the intergenic upstream regions to tCGI and dCGI, the ratio $A \rightarrow G/T \rightarrow C$ is less than 1 (the mean and standard deviation of the ratio in the first 10 windows upstream to dCGIs is 0.934 ± 0.02), whereas, in intergenic downstream regions, it is greater than 1 (mean and standard deviation in the first 10 windows downstream is 1.09 ± 0.023). In addition, on both sides of the CGI (5' and 3'), the $A \rightarrow G/T \rightarrow C$ ratio is constant along several hundreds

of kilo base pairs. These properties of the ratio profile indicate that there is a bidirectional asymmetry around CGIs that has the CGI as its origin and that the mutational pressure that leads to the asymmetry is constant over long distances. In contrast, the ratio of $C \rightarrow T/G \rightarrow A$ fluctuates around 1 along the flanking intergenic regions of CGI, an indication that the strand symmetry between these rates holds around CGIs. Among the four pairs of complementary transversion rates, three are symmetric and only the symmetry between the rates of $C \rightarrow G$ to $G \rightarrow C$ is broken (supplementary figs. 1 and 2, Supplementary Material online).

We also performed the substitution analysis around the middle point between two consecutive CGIs. On a genome-wide scale, there is a change in the direction of strand asymmetries at the midpoint (supplementary fig. 3, Supplementary Material online). The ratio $A \rightarrow G/T \rightarrow C$ is greater than 1 at the 5' region of the midpoint, whereas, in the 3' region, the ratio is lower than 1. The change in the direction of the asymmetry is smooth (supplementary fig. 3, Supplementary Material online) around the midpoint, in contrast, to the sharp change in the $A \rightarrow G/T \rightarrow C$ ratio at the CGIs (fig. 2). The analysis around the midpoint rule out the scenario that the ratio profiles in figure 2 can be found around any random positions in the genome; and therefore, it establishes the role of CGIs as origins of the bidirectional asymmetry.

The Strand Bias Is Quantitative Similar in tCGIs and dCGIs

In transcribed regions (excluding exons), a similar qualitative asymmetry, that is, $A \rightarrow G/T \rightarrow C > 1$, has been already reported (Green et al. 2003; Polak and Arndt

2008). The excess of $A \rightarrow G$ over $T \rightarrow C$ in introns has been suggested to be a result of transcription-coupled repair (TCR) and a higher misincorporation rate of G's (instead of A's) into the nascent DNA at positions with a template T compared with the reverse complement process. If only transcription induces the asymmetries in intergenic regions around transcripts, then one would expect a higher level of strand bias around tCGI than around dCGIs. However, the ratio profile is quantitative identical around these two classes of CGIs. This observation is a strong indication that the transcription of genes is not the cause of long-range asymmetries (a point which will be further addressed below).

The Ratio $A \rightarrow G/T \rightarrow C$ Is Greater in Introns Than in Intergenic Regions

To check whether transcription has an additional impact on the ratios between complementary substitution rates, we performed a sliding-window analysis also along intronic regions of genes whose TSS is inside of a tCGI and we denote them by tCGI genes (see fig. 1 and Materials and Methods). For these introns, the ratios were calculated in windows of 10 kbp starting from the TSS up to 1 Mbps downstream to the tCGI (supplementary fig. 4, Supplementary Material online). We found out that in introns the ratio of $A \rightarrow G$ to $T \rightarrow C$ rates is 1.6 compared with only 1.1 in intergenic regions at a similar distance from the 3' end of tCGIs (fig. 2). Hence, the greater excess of $A \rightarrow G$ over $T \rightarrow C$ substitutions along the nontemplate strand of tCGI genes compared with intergenic regions suggests that the transcription process has a primary impact on the asymmetries in introns. From the fact that the ratio $A \rightarrow G/T \rightarrow C$ is relative constant along introns (fig. 2), one can conclude that these transcription-associated mutational forces are not dependent on the distance from CGIs. Note that by our definition (see fig. 1 and Materials and Methods), at any given bin, the intergenic regions are pooled out of the 3' intergenic flanking regions of tCGI genes that are shorter than the ones whose sequences are taken for the analysis of the intronic regions.

In order to further test whether the transcription process is the primarily force that shapes the asymmetries in introns, we measured the ratios in intronic regions pooled from genes, whose TSS is not inside a CGI (denoted by non-tCGI genes). We divided the set of non-tCGI genes into two classes according to the direction of their transcription relative to the closest CGI, which can be either dCGI or tCGI. The first class is composed of genes that are transcribed outward from the closest CGI. The second class called inward genes and it contains genes that are transcribed toward the closest CGI. We found that the ratio $A \rightarrow G/T \rightarrow C$ is about 1.4 on the nontemplate strand of both classes of genes (supplementary fig. 5, Supplementary Material online). Therefore, for some genes, the direction of the asymmetries is opposite to the one in the intergenic regions. In introns of genes that are transcribed outward of CGIs, the direction of the asymmetry is the same as in intergenic regions (supplementary fig. 5, Supplementary Material online), whereas introns of genes that are transcribed toward CGIs have opposite asymmetries than in the average intergenic regions at similar distances from the CGI

(see supplementary fig. 5 [Supplementary Material online] and fig. 2). Hence, the direction of the strand asymmetries along genes is determined by the direction of transcription.

Low $A \rightarrow G/T \rightarrow C$ Ratio in Introns That Overlap tCGIs

Interestingly, in intronic regions that overlap the tCGI, the ratio $A \rightarrow G/T \rightarrow C$ is about 1.1 (this is the first bin in the analysis of 3' intronic regions in fig. 2, and more detailed analysis can be found in supplementary fig. 6, Supplementary Material online), whereas in intronic regions downstream to the tCGI, the ratio fluctuates between 1.5 and 1.7 (fig. 2). Moreover, this ratio in introns that overlap tCGIs is similar to the one in intergenic regions (fig. 2). Therefore, the asymmetry level within tCGI is not dependent on transcription; this implies that TCR is not active in the transcribed parts of CGIs or that its impact on substitution pattern is obscured by other mutational processes.

Additional Asymmetry between $C \rightarrow T$ and $G \rightarrow A$ Is Found in Intronic Regions

In contrast to intergenic regions, we observed in intronic regions that there is an additional broken symmetry that is between $C \rightarrow T$ and $G \rightarrow A$ rates. This symmetry is broken in two different ways and on different scales. The first asymmetry is an excess of $C \rightarrow T$ over $G \rightarrow A$ that is restricted to intronic regions that overlap CGI (fig. 2 and supplementary fig. 6 [Supplementary Material online]). This bias is the same localizes asymmetry, which we have reported before, at the first 2 kbp downstream to the TSS (Polak and Arndt 2008); the broken of the symmetry between $C \rightarrow T$ and $G \rightarrow A$ in these regions has been suggested to be a consequence of transcription-associated mutagenesis. An opposite and weaker bias is found in regions that are located 50–500 kbp downstream to the CGIs (fig. 2) as it has been previously observed by Green et al. (2003) and has been suggested to be a result of TCR.

CpG Methylation–Deamination Rates Are Lower in CGIs

Because CGIs are richer in their GC and CpG content from other genomic regions, we also focused on the CpG methylation–deamination rates within CGIs and $W \rightarrow S/S \rightarrow W$ ratio (supplementary fig. 7, Supplementary Material online). The CpG methylation–deamination (transition) rates drop in CGIs in almost two orders of magnitudes. In the flanking regions of tCGI and dCGI, the average surplus of the $CpG \rightarrow TpG$ and $CpG \rightarrow CpA$ substitution rate is 0.08 per dinucleotide, whereas in both tCGI and dCGI, this rate is about 0.001 and 0.002 per CpG, respectively (supplementary fig. 7, Supplementary Material online). Interestingly, the transversion rates, $CpG \rightarrow ApG/CpT$, are similar within and in the flanks of CGIs. The sharp decrease in $CpG \rightarrow TpG/CpA$ in CGIs has been suggested to be a result of lower methylation levels in CGI that in turn leads to lower number of $C \rightarrow T$ transitions (Saxonov et al. 2006). However, the similar rates of methylation-mediated transversions of

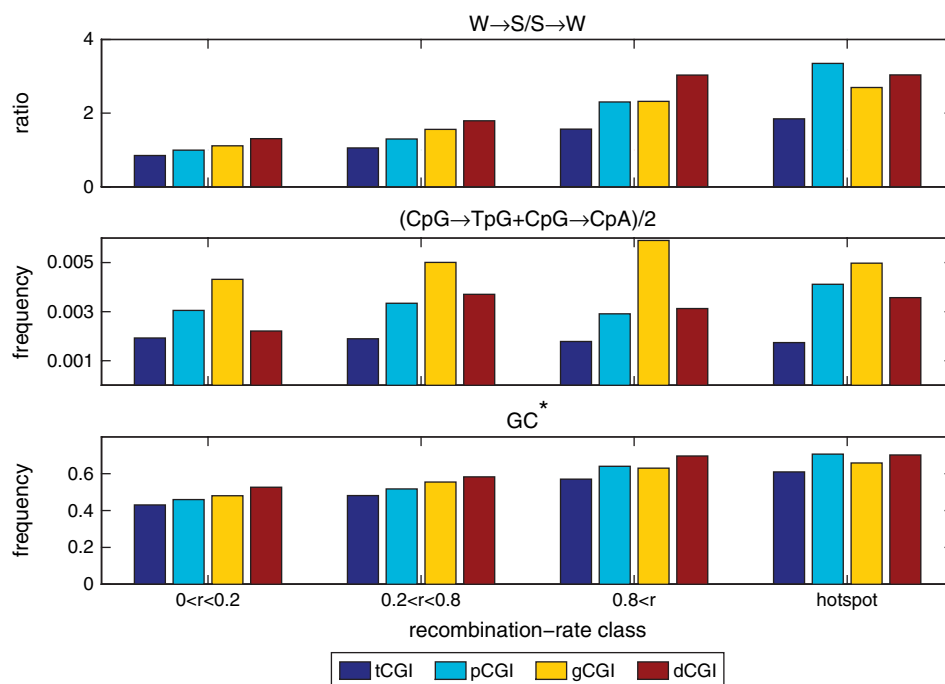


FIG. 3.—Dependence of $W \rightarrow S/S \rightarrow W$ ratio, CpG deamination frequencies, and the stationary GC content from the recombination rate for four classes of CGIs (see Materials and Methods). The CGIs in the four (t-,p-,g-,d-) CGI classes are subdivided according to four recombination rate ranges ($0 < r < 0.2$, $0.2 < r < 0.8$, $0.8 < r$, hot spots), which are denoted on the horizontal axis.

CpGs in CGIs and in their flanks imply that there are additional factors that shape the profile of CpG mutations in CGIs beside the methylation levels themselves.

An Excess of $W \rightarrow S$ over $S \rightarrow W$ Is Found within dCGI

The balance of $W \rightarrow S$ versus $S \rightarrow W$ rates impacts the GC content; higher rates of $W \rightarrow S$ than $S \rightarrow W$ will lead to enrichment of GC nucleotides in the analyzed regions. Because, the GC content of CGIs is much higher than in the rest of the genome, we wanted to see if this is reflected in the ratio of $W \rightarrow S/S \rightarrow W$ within CGIs compared with the ratio in their flanking regions. In particular, we wanted to test if this ratio is invoked by other factors, then methylation. Therefore, we calculated the $W \rightarrow S$ rates without including of methylation-deamination rates. We found out that the ratios are different between tCGI and dCGI. In the last, the ratio of $W \rightarrow S/S \rightarrow W$ is about 2, whereas, in tCGI, the ratio of $W \rightarrow S/S \rightarrow W$ is about 1 (supplementary fig. 7, Supplementary Material online). In the immediate flanking regions of both types of CGIs, this ratio is lower than 1 (supplementary fig. 7, Supplementary Material online). Hence, the $W \rightarrow S/S \rightarrow W$ profile implies that GC-enriching substitution pattern is a feature of the CGIs, but the level of this enrichment is higher in tCGIs.

In the next step, we used the estimated substitution rates to calculate the stationary GC content, denoted by GC^* , that is, the GC content that is acquired after a long time assuming that the substitution rates do not change (Duret and Arndt 2008). GC^* in dCGI (about 0.6; supple-

mentary fig. 7, Supplementary Material online) is higher than GC^* in tCGI (about 0.42; supplementary fig. 7, Supplementary Material online). The GC content in both classes of CGI is lower in equilibrium than the current GC frequency of 0.65 in both types of CGIs (supplementary fig. 7, Supplementary Material online).

The Recombination Rate Is Correlated with the $W \rightarrow S/S \rightarrow W$ Ratio within CGIs

The primary process that is known to increase the $W \rightarrow S/S \rightarrow W$ ratio is recombination through GC-BGC (Galtier and Duret 2007). In regions where a heteroduplex is formed mismatches can occur. It has been suggested that repair of mismatches between S bases (GC) and W bases (AT) will be corrected preferentially into S bases (Meunier and Duret 2004). Therefore, we further subdivided each class of CGIs according to the estimated recombination rates in every CGI. At first stage, CGIs that overlap with a recombination hotspot were grouped into a subclass called hotspot. At the second stage, the remaining CGIs were subdivided into further three subclasses according to their recombination rates: below 0.2 and 0.2–0.8 cM/Mb and above 0.8 cM/Mb. Indeed, there is a positive correlation between the ratio $W \rightarrow S/S \rightarrow W$ and the recombination rate in tCGI and in dCGI (fig. 3), this correlation exists also for gCGIs and pCGIs (fig. 3). In addition, the $W \rightarrow S/S \rightarrow W$ ratios in dCGIs are greater than the one in tCGI at any recombination level; this indicates that recombination is not the only factor that impacts the $W \rightarrow S/S \rightarrow W$ ratio but also the absence or presence of TSS on CGI contributes to the bias of $W \rightarrow S$ over $S \rightarrow W$.

Discussion

Using triple alignments of human, chimpanzee, and macaque, we could estimate the substitution rates in human CGIs and their flanks. Our analysis reveals two novel findings: first, the existence of long-range strand asymmetries in the intergenic flanking to the CGI regions; second, the ratio $W \rightarrow S/S \rightarrow W$ is higher in dCGI than in tCGI and it is positively correlated with the estimated recombination rates in CGIs. What processes can generate these two patterns? How these mutational processes are associated with CGI?

Replication Forms the Bidirectional Strand Asymmetries in Intergenic Regions

CGIs Are OBRs

The association of CGIs with OBR (Antequera and Bird 1999; Gomez and Antequera 2008) has been suggested before and recently further established in a genome-wide study (Cadoret et al. 2008). Similar patterns of $A \rightarrow G/T \rightarrow C$, that is, bidirectional asymmetries, have been associated with ORI in bacterial and mitochondrial genomes. Therefore, we suggest that the asymmetries around CGIs and the change in the direction of the asymmetry in the CGIs (fig. 2) are due to initiation of bidirectional replication in these regions.

The range of the asymmetries is several hundreds kilo base pairs that is similar to average length of a replicon and one order of magnitude longer than the average size of transcripts (fig. 2). In case only transcription would invoke these asymmetries, CGIs would be origins of unknown transcripts of length of hundreds of kilo base pairs in both outward directions from CGIs. In particular, such transcripts should be originated not only for tCGIs but also for dCGIs. Recent articles suggest that over 90% of the genome is transcribed; therefore, most of what we annotated as intergenic regions are transcribed at least once (The ENCODE Project Consortium 2007). If such unknown transcripts induced this asymmetry, their orientation would be biased relative to the CGIs, that is, the transcription would tend to occur in outward direction from the CGIs. Using the estimated rates for introns (supplementary fig. 4, Supplementary Material online) and in intergenic region (supplementary fig. 1, Supplementary Material online), we suggest that a surplus of at least 17% of the nucleotides in the so intergenic regions around CGI should be transcribed in outward orientation relative to CGIs (see Supplementary Material); however, there is no sufficient data at this time to check this prediction.

Are all CGIs associated with OBRs? Cadoret et al. (2008) have found that 99 of 506 CGIs in the ENCODE regions serve as OBRs in a specific cell line. It is possible that different CGIs are ORIs in different cell types because the factors that link between CGIs and OBRs might be cell specific (for further discussion on these factors, see the Supplementary Material online). The substitution rates that we measured in the present study are the result of mutations that occurred in or during the production of human germ cells. Up to now, data on OBRs in these cell types are not available.

Strand Asymmetries Are Due to Difference in DNA Polymerases Error Spectra

The association of OBR with strand asymmetries in the human genome implies that on the leading strand, which serves as the template strand of a new lagging strand, the ratios $A \rightarrow G/T \rightarrow C$ and $C \rightarrow G/G \rightarrow C$ are greater than 1 (fig. 2; supplementary figs. 1 and 2, Supplementary Material online) as in many bacterial genomes (Rocha et al. 2006). Also in human mitochondrial DNA (mtDNA), one observes that along the so-called heavy strand, the rate of $A \rightarrow G$ substitutions is higher than of $T \rightarrow C$ substitution (Faith and Pollock 2003).

There are two main mechanisms that can invoke a ratio $A \rightarrow G/T \rightarrow C$ greater than 1 on the leading strand (and on the heavy strand of mtDNA). In bacteria (and for mtDNA), it has been suggested that adenine on the leading (heavy) strand is subject to higher deamination rates because the leading (heavy) strand is found in ssDNA conformation and adenine is less stable in this conformation than when it is base paired to T in dsDNA conformation. However, the exposure of ssDNA should also induce even stronger asymmetry between $C \rightarrow T$ and $G \rightarrow A$ because the rates of cytosine deamination in ssDNA are 140 faster than in dsDNA (Frederico et al. 1990). And indeed along the heavy strand of the mitochondrial genome, one observed a stronger excess of the rates of $C \rightarrow T$ over $G \rightarrow A$ than of $A \rightarrow G$ over $T \rightarrow C$ (Faith and Pollock 2003). We find that the ratio $C \rightarrow T/G \rightarrow A$ is close to 1 (fig. 2), and therefore, it is unlikely that the exposure of the leading strand is the main cause for the excess of $A \rightarrow G$ over $T \rightarrow C$ on this strand (even though, it is possible that human cells have repair mechanism that can handle better the cytosine deamination damage than bacteria). This suggests that the mechanism that has introduced asymmetry along the human mitochondrial DNA is different than the one that has generated the asymmetry in the nuclear DNA.

Another potential source of the asymmetry between the leading and lagging strands has been found in yeast. A recent study (Kunkel and Burgers 2008) demonstrated that the lagging and leading strands are synthesized by two different DNA polymerases, pol δ and pol ϵ , respectively. Assuming that also two different polymerases are used in human, we suggest that the observed asymmetries are due to a difference between the error spectra of these two DNA polymerases. During synthesis of the nascent DNA, there are 12 possible single base–base mismatches that can occur. However, there are two DNA polymerases participating in DNA replication and therefore potentially 24 misincorporation rates. Each of the 12 single-base mutation rates is a combination of two errors that are caused by pol δ and pol ϵ . The couple of replication errors that contribute to substitution of A in G on the leading strand are misincorporation of G opposite to template T by pol ϵ and of C opposite to template A by pol δ that result in G–T and A–C (leading lagging) mismatches. These mismatches become substitution mutations $A \rightarrow G$ when they are repaired into G–C base pair. In a similar manner, substitution T in C on the leading strand is a result of misincorporation by pol δ of C opposite template A and of G opposite template T by pol ϵ , which results in C–A and T–G mismatches.

Unfortunately, the error rate during replication of these two polymerases, pol δ and pol ϵ , is currently only known for yeast (Thomas et al. 1991) but not for the human polymerases. Therefore, we cannot examine this model but rather we predict that the rates that lead to $A \rightarrow G$ on the leading strand are higher than the error rates that lead to $T \rightarrow C$.

Implication on Genomic Architecture: Coorientation of Transcription and Replication

The association of CGI with replication and transcription suggests that these processes are coupled to each other through CGIs. Such association has several evolutionary consequences. Coinciding of ORI and TSSs implies that genes with CGI promoter are transcribed from the leading strand in human. It has been suggested before that in human, transcription is coordinated with the replication fork progress (Huvet et al. 2007) similar to some bacterial species (Rocha and Danchin 2003; Rocha 2008). The coorientation of transcription and replication is suggested to reduce head-on collisions between DNA and RNA polymerases in bacteria (Rocha and Danchin 2003). On the other hand, there is an ongoing discussion whether most human genes are coded on the leading strand or not (Huvet et al. 2007; Necsulea et al. 2009).

Possible Higher Rates of BGC in CGI *GC-BGC*

A by-product of recombination is considered to be the primary mutational process that increases the ratio $W \rightarrow S/S \rightarrow W$ in the genome (Hwang and Green 2004; Meunier and Duret 2004; Galtier and Duret 2007; Duret and Arndt 2008). Indeed, the $W \rightarrow S/S \rightarrow W$ ratio is correlated with the estimated recombination rates in the different CGI categories (fig. 3). However, although, the ratio $W \rightarrow S/S \rightarrow W$ is greater in CGIs than in their flanks; the average recombination rate is the same in CGIs and in their immediate flanking regions (supplementary fig. 7, Supplementary Material online). This implies that recombination cannot solely explain the profile of $W \rightarrow S/S \rightarrow W$. We suggest two explanations that can be accounted for this difference. First, the recombination rates that we use in this study are actually an estimation of the rate of crossing-over events, and they miss the non-crossing-over events. Therefore, it is possible that the rates of recombination are locally higher within CGI regions, but crossing-over is suppressed due to purifying selection. Second, the $W \rightarrow S$ versus $S \rightarrow W$ bias can be shaped by positive selection for high GC content in CGIs due to their role in gene regulation (Strathdee et al. 2004). The GC-rich sequence of CGIs is associated with an open chromatin structure that allows the binding of RNA polymerase and transcription factors to the DNA (Antequera and Bird 1999). The above explains the preference for high stationary GC content in tCGI. The stronger bias of $W \rightarrow S$ versus $S \rightarrow W$ in dCGI might point to distinct mutational mechanisms that shape the GC content in tCGI and dCGI. We suggest that the gaining of GC bases by BGC is a transient state. All CGIs have first emerged in the genome by BGC activity (which also increases the total substitution rates); but once a CGI becomes functional in the context of near by genes (tCGI), it also becomes constrained

and has lower mutation rates that preserve its GC content. Indeed, there is a difference in the total substitution rates between the two classes; within dCGI, the rate is greater in 40% than in tCGI (supplementary fig. 7, Supplementary Material online). Interestingly, the total substitution rate is higher in both tCGI and dCGI than in their flanking regions, suggesting that these regions have undergone a period of positive selection for high GC content or BGC (supplementary fig. 7, Supplementary Material online).

Are CGIs Vanishing?

In this study, a CGI is defined as DNA sequence longer than 400 bp that fulfills two conditions: 1) the GC content is above 50% and 2) the CpG observed/expected ratio is greater than 0.6 (see Materials and Methods). However, the value of the stationary GC content in tCGIs ($GC^* = 0.41$) suggests that these regions are predicted to lose their CGI properties in the long run, whereas dCGI with ($GC^* = 0.58$) will keep them. This result is counter intuitive because tCGIs are assumed to play a major role in gene regulation (Antequera and Bird 1999), and one would expect that substitution rates would maintain the GC content in the vicinity of TSS. It is possible that the mutational forces that increase the GC content and build up the CGIs are more active in these regions only prior to time of the association of CGI with transcription. When CGI harbors TSS and becomes a dominate regulator of gene transcription, the mutational forces that induce mutations might have deleterious impact. An alternative explanation might be that other mutational forces such as insertion and deletion can lead to an increase of GC content that counterbalance the impact of nucleotide substitution rates.

Conclusion

The availability of primate genomes and their corresponding alignments together with high-quality genome annotation enables us to gain insights on difference in mutational processes in different contexts along human chromosomes. In particular, one can address the question about the mutational signature that is associated with different cellular processes. In an earlier work, we used the annotation of genes to study the impact of transcription on mutation patterns in the vicinity of 5' end and 3' end of genes (Polak and Arndt 2008). In this paper, we presented an analysis of substitution pattern within and around CGIs, which are mammalian sequence features. Because many of the CGIs overlap with TSSs, these two studies suggest that CGIs are associated with five substitution patterns that are the mutational signatures of the main cellular processes in human germ cells. First, a long-range bidirectional substitution asymmetry between $A \rightarrow G$ and $T \rightarrow C$, which is induced by replication. Second, an excess of $W \rightarrow S$ over $S \rightarrow W$ in dCGIs that is invoked (via BGC) by higher recombination rates. Third, a strong bias of $A \rightarrow G$ over $T \rightarrow C$ in intronic regions in addition to higher rates of $G \rightarrow A$ over $C \rightarrow T$ (at distance of more than 50 kb downstream to the tCGI); this asymmetries are results of TCR. Forth, an excess of $C \rightarrow T$ over $G \rightarrow A$ in transcribed regions of tCGIs and limited to the first 2 kb that are induced by

transcription-associated mutagenesis (Polak and Arndt 2008). Fifth and last, a 100-fold drop in the CpG deamination rates in CGI relative to the genome-wide rate levels due to lower level of methylation in CGIs (Polak and Arndt 2008). The increasing amount of functional genomic data such as transcription levels, methylation levels, histone positioning, histone modifications, positions of transcription factor-binding site, and other genetic features, all on a genome-wide level (Bock and Lengauer 2008) will enable us to further explore the interplay between mutational and functional processes, especially, the role of epigenetics in shaping substitution patterns.

Supplementary Material

Supplementary figures 1–7 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Literature Cited

- Aerts S, Thijs G, Dabrowski M, Moreau Y, De Moor B. 2004. Comprehensive analysis of the base composition around the transcription start site in Metazoa. *BMC Genomics*. 5:34.
- Aladjem MI. 2007. Replication in context: dynamic regulation of DNA replication patterns in metazoans. *Nat Rev Genet*. 8: 588–600.
- Antequera F, Bird A. 1999. CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr Biol*. 9:R661–R667.
- Bock C, Lengauer T. 2008. Computational epigenetics. *Bioinformatics*. 24:1–10.
- Cadoret JC, et al. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci USA*. 105:15837–15842.
- Delgado S, Gómez M, Bird A, Antequera F. 1998. Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J*. 17:2426–2435.
- Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 4:e1000071.
- Faith JJ, Pollock DD. 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics*. 165:735–745.
- Flicek P, et al. 2008. Ensembl 2008. *Nucleic Acids Res*. 36: D707–D714.
- Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*. 238:65–77.
- Frederico L, Kunkel T, Shaw B. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*. 29:2532–2537.
- Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet*. 23:273–277.
- Gomez M, Antequera F. 2008. Overreplication of short DNA regions during S phase in human cells. *Genes Dev*. 22:375–385.
- Green P, Ewing B, Miller W, Thomas PJ, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet*. 33:514–517.
- HapMap. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 449:851–861.
- Huvet M, et al. 2007. Human gene organization driven by the coordination of replication and transcription. *Genome Res*. 17:1278–1285.
- Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA*. 101: 13994–14001.
- Kunkel TA, Bebenek K. 2000. DNA replication fidelity. *Annu Rev Biochem*. 69:497–529.
- Kunkel TA, Burgers PM. 2008. Dividing the workload at a eukaryotic replication fork. *Trends Cell Biol*. 18:521–527.
- Lobry J. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol*. 40:326–330.
- Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*. 13:660–665.
- McCulloch SD, Kunkel TA. 2008. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res*. 18:148–161.
- Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol*. 21:984–990.
- Mrazek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci USA*. 95:3720–3725.
- Necsulea A, Guillet C, Cadoret J-C, Prioleau M-N, Duret L. 2009. The relationship between DNA replication and human genome organization. *Mol Biol Evol*. 26:729–741.
- Palleja A, Guzman E, Garcia-Vallve S, Romeu A. 2008. In silico prediction of the origin of replication among bacteria: a case study of *Bacteroides thetaiotaomicron*. *OMICS: J Integr Biol*. 12:201–210.
- Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. 2008. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res*. 18:1814–1828.
- Polak P, Arndt PF. 2008. Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res*. 18:1216–1223.
- Rocha EPC. 2008. The organization of the bacterial genome. *Annu Rev Genet*. 42:211–233.
- Rocha EPC, Danchin A. 2003. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet*. 34:377–378.
- Rocha EPC, Touchon M, Feil EJ. 2006. Similar compositional biases are caused by very different mutational effects. *Genome Res*. 16:1537–1547.
- Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci USA*. 103:1412–1417.
- Squartini F, Arndt PF. 2008. Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Mol Biol Evol*. 25:2525–2535.
- Strathdee G, Sim A, Brown R. 2004. Control of gene expression by CpG island methylation in normal cells. *Biochem Soc Trans*. 32:913–915.
- The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 447:799–816.
- Thomas D, et al. 1991. Fidelity of mammalian DNA replication and replicative DNA polymerases. *Biochemistry*. 30:11751–11759.
- Touchon M, et al. 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc Natl Acad Sci USA*. 102:9836–9841.
- Touchon M, Rocha EPC. 2008. From GC skews to wavelets: a gentle guide to the analysis of compositional asymmetries in genomic data. *Biochimie*. 90:648–659.

Eugene Koonin, Associate Editor

Accepted July 22, 2009