Long range bi-directional strand asymmetries originate at CpG islands in the human genome

Paz Polak and Peter F Arndt

Max Planck Institute for Molecular Genetics,
Ihnestrasse 73, 14195 Berlin, Germany

Supplementary Material

Supplementary Materials and Methods

Substitution analysis in resolution of 100bp around CGIs. We performed a sliding window (100 bp long) analysis within CGIs and their 2.5 kbp long flanking sequences. In contrast to our initial study in the main text of intergenic regions in 2 Mbp regions surrounding the CGIs, in this part of the analysis we included also intronic regions, since for most of tCGIs excluding intergenic regions (and their 5 kbp flanking regions) would leave us without any alignments to analyses. For dCGI the 2.5 kbps long flanking regions are intergenic because of these CGIs are at distance of more than 10kb from a gene. In case of tCGI, the analysis in the proximity of its 5' end is done along intergenic regions (Figure 1), while the analysis in the vicinity of the 3' end of tCGI is done along intronic regions (Figure 1). In Supplementary Figure 6 and Supplementary Figure 7 we estimated the substitution rates in even higher resolution using 100 bps long windows. The analysis was done along the CGI and their 2.5 kbp flanking regions. For this analysis we excluded only exons (coding and UTR regions).

Supplementary Discussion

How are ORIs mechanistically linked to CpG islands? It is not yet known how ORIs are determined in mammalian species. It is even less understood why replication is initiated from CGIs (Aladjem 2007). Since in CGIs are enriched both in set of TSSs and ORIs, it is tempting to suggest the same factors that are associated with transcription initiation are involved in replication initiation (Antequera 2003) such as: low methylation levels in CGI (Antequera and Bird 1999; Rein et al. 1999), particular histone modifications (Lucas et al. 2007), and transcription factor binding sites (Cadoret et al. 2008). It has also been suggested that CGIs harbor the binding site for Origin Recognition Complex (ORC), which is essential for replication initiation (Keller et al. 2002). It is further possible that transcription factors that are usually associated with transcription can also promote replication, e.g. c-Jun or c-Fos (Murakami et al. 1991). Transcription *per se* has been shown not to be necessary for the use of CGIs as ORI but transcription can impact the timing of the replication, transcription is initiated first from transcribed regions (Gomez and Brockdorff 2004).

Currently it is believed that for the vast majority of CGIs that overlap TSS there is a paused RNA polymerases at any given time (Guenther et al. 2007) and therefore an absence of transcription does not exclude that the RNA polymerase and other parts of transcriptional machinery attract the replication machinery. Moreover, recent research suggests that in mammalian CpG islands there is an overproduction of short RNA even in the absence of a full transcript production (Seila et al. 2008). Such short RNA are transcribed both from the sense and antisense strand, where the sense transcripts accumulate downstream to the TSS and the antisense are primarily found upstream to the TSS (Seila et al. 2008). Similar phenomena have been discovered for replication initiation from CpG islands, where overproduction of short DNA sequences that overlap CGI have been detected during S phase of the cell cycle (Gomez and Antequera 2008). It is possible that transcription of several dozens bps of CGIs is part of initiation of replication in particular of the leading strand.

Estimation of unknown bi-directional transcription from CGIs. We have already suggested that the asymmetries within intergenic regions might be caused by replication. An alternative model is that unknown transcription around CGI induces these asymmetries. The fact that the vast majority of the genome is transcribed (Gerstein et al. 2007) may suggest that transcription and TCR are active on a genome wide level (Polak and Arndt 2008) and could

over time have generated the bi-directional asymmetry around CGIs. However, in this case the transcriptional activity has to be biased to transcribe the strand in an outward orientation relative of CGIs.

In order to quantify the amount of transcription that is needed to generate the observed asymmetry in intergenic regions, we assumed that in intergenic regions downstream to the CGI, a fraction *p* of the sequence is evolved according to an asymmetric model and (1-*p*) fraction is evolved according to a symmetric model, in which complementary substitution rates are equal to each other. Under our model the relation between the frequency of substitution of X in Y in intergenic (*i*) regions and the rates by asymmetric (*as*) and symmetric (*s*) models is described by:

$$r_i\left(X \to Y\right) = p * r_{as}\left(X \to Y\right) + (1-p) * r_s\left(X \to Y\right)$$

The fraction *p* can be different for complementary bases since the frequency of such bases are, in general, not equal to each other in regions that evolved under asymmetric model, for example, there is an excess of Ts over As in intronic regions (Supplementary Figure 8); however, even in regions that have been (so far) known to be subject to the strongest asymmetric mutational forces in the genome, i.e. introns, the bias is not significantly high (less than 10% surplus of Ts over As in introns). Therefore, we approximate the A and T content to be the same in sequences, which are either evolved according to symmetric or asymmetric model, and by that we can reduce the number of parameters in our model.

Therefore, the substitution of A in G and T in C in intergenic regions are described by:

$$r_i\left(A \to G\right) = p * r_{as}\left(A \to G\right) + (1-p) * r_s\left(A \to G\right) \qquad (1)$$

and

$$r_i\left(T \to C\right) = p * r_{as}\left(T \to C\right) + (1-p) * r_s\left(A \to G\right). \qquad (2)$$

Notice that $r_s\left(T \to C\right) = r_s\left(A \to G\right)$ because of the definition of complementary symmetric model. From the above two equations we derive that:

$$p = \frac{r_i\left(A \to G\right) - r_i\left(T \to C\right)}{r_{as}\left(A \to G\right) - r_{as}\left(T \to C\right)}$$

To get the rates of A->G and T->C in intergenic regions we averaged the rates that are given in Supplementary Figure 1, over 10 consecutive windows downstream to the 3' end of dCGI. To get the rates of the asymmetric model we averaged the substitution rates in intronic regions (Supplementary Figure 4) over 10 consecutive windows downstream to the 3' end of tCGI:

$$r_i\left(A \to G\right) = 0.0033 \pm 0.0001, \quad r_{as}\left(A \to G\right) = 0.0034 \pm 0.0001$$
$$r_i\left(T \to C\right) = 0.0031 \pm 0.0001, \quad r_{as}\left(T \to C\right) = 0.0022 \pm 0.0001$$

Using these values we found that at least $p = 0.167$ fraction of the DNA is evolved under asymmetric model, i.e. there is surplus of about 17% in transcription outwards CGIs than inwards. Here, the parameters we used for the asymmetric model are estimated in transcribed regions of known genes. It is likely that the transcription levels of unknown genes are lower than of known genes and therefore the mutational bias should be also higher. Hence, 17% surplus of outward expression is a conservative lower bound of bi-directional transcription from a given CGI and it is reasonable to assume a higher excess.

Supplementary References

Aladjem MI. 2007. Replication in context: dynamic regulation of DNA replication patterns in metazoans. Nat Rev Genet **8**:588-600.

Antequera F. 2003. Structure, function and evolution of CpG island promoters. Cellular and Molecular Life Sciences (CMLS) **60**:1647-1658.

Antequera F, and Bird A. 1999. CpG islands as genomic footprints of promoters that are associated with replication origins Current Biology **9**:R661-R667

Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, and Prioleau M. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. PNAS **105**:15837-15842.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, and Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. Genome Research **17**:669-681.

Gomez M, and Antequera F. 2008. Overreplication of short DNA regions during S phase in human cells. Genes & Development **22**:375-385.

Gomez M, and Brockdorff N. 2004. Heterochromatin on the inactive X chromosome delays replication timing without affecting origin usage. PNAS **101**:6923-6928.

Guenther MG, Levine SS, Boyer LA, Jaenisch R, and Young RA. 2007. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. Cell **130**:77-88.

Keller C, Ladenburger E-M, Kremer M, and Knippers R. 2002. The Origin Recognition Complex Marks a Replication Origin in the Human TOP1 Gene Promoter. J. Biol. Chem. **277**:31430-31440.

Lucas I, Palakodeti A, Jiang Y, Young D, Jiang N, Fernald A, and Le Beau M. 2007. High-throughput mapping of origins of replication in human cells EMBO reports **8**:770-777.

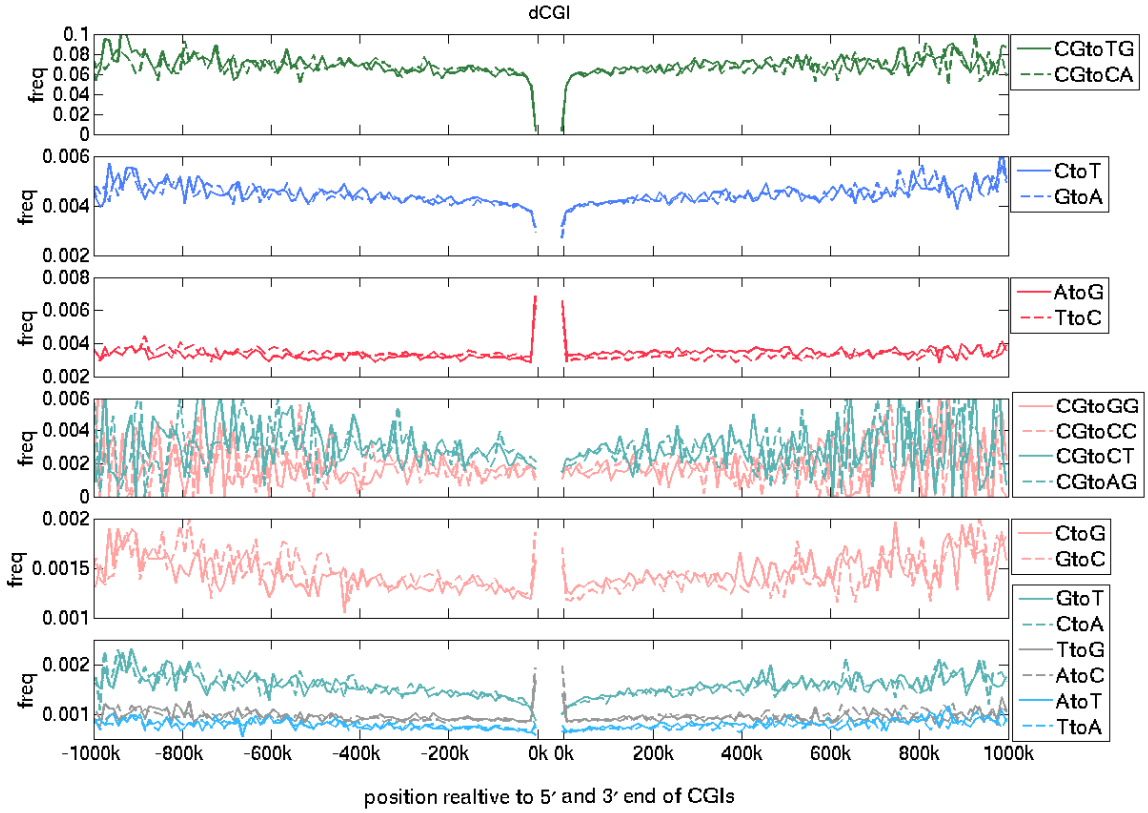
Murakami Y, Satake M, Yamaguchi-Iwai Y, Sakai M, Muramatsu M, and Ito Y. 1991. The nuclear protooncogenes c-jun and c-fos as regulators of DNA replication. PNAS **88**:3947-3951.

Polak P, and Arndt PF. 2008. Transcription induces strand-specific mutations at the 5' end of human genes. Genome Research **18**:1216-1223.

Rein T, Kobayashi T, Malott M, Leffak M, and DePamphilis ML. 1999. DNA Methylation at Mammalian Replication Origins. J. Biol. Chem. **274**:25792-25800.

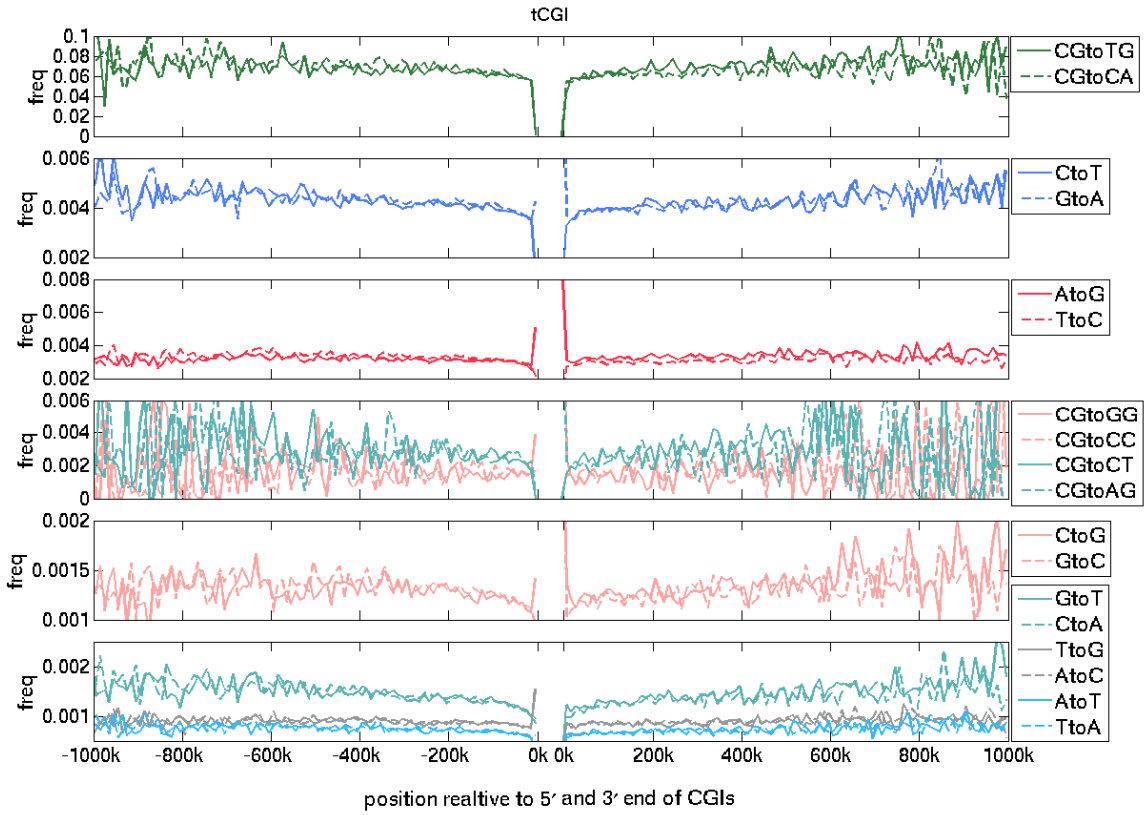
Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, and Sharp PA. 2008. Divergent Transcription from Active Promoters. Science **322**:1849-1851.

Supplementary Figures



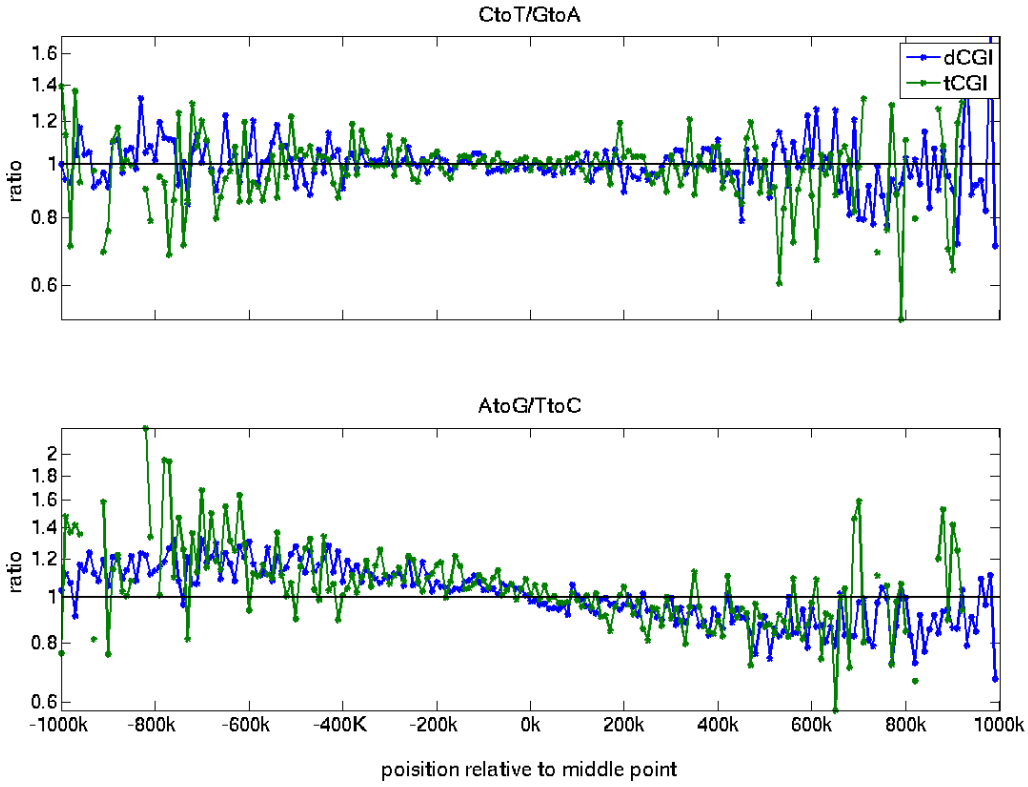
Supplementary Figure 1

Substitution frequencies within dCGIs and along their 1 Mbps long upstream and downstream intergenic regions. The plots show the estimated twelve single nucleotide substitution frequencies and 6 CpG deamination frequencies in non-overlapping 10 kbp long windows. The 5' and 3' ends of dCGIs are represented by left and right 0k, respectively. The distances of the windows' centres from the 5'end and 3' end of dCGIs are indicated by negative and positive values, respectively. The estimated of rates within dCGIs are indicated by the data point at 0ks (see Supplementary Figure 8 for the amount for sequence in each window).



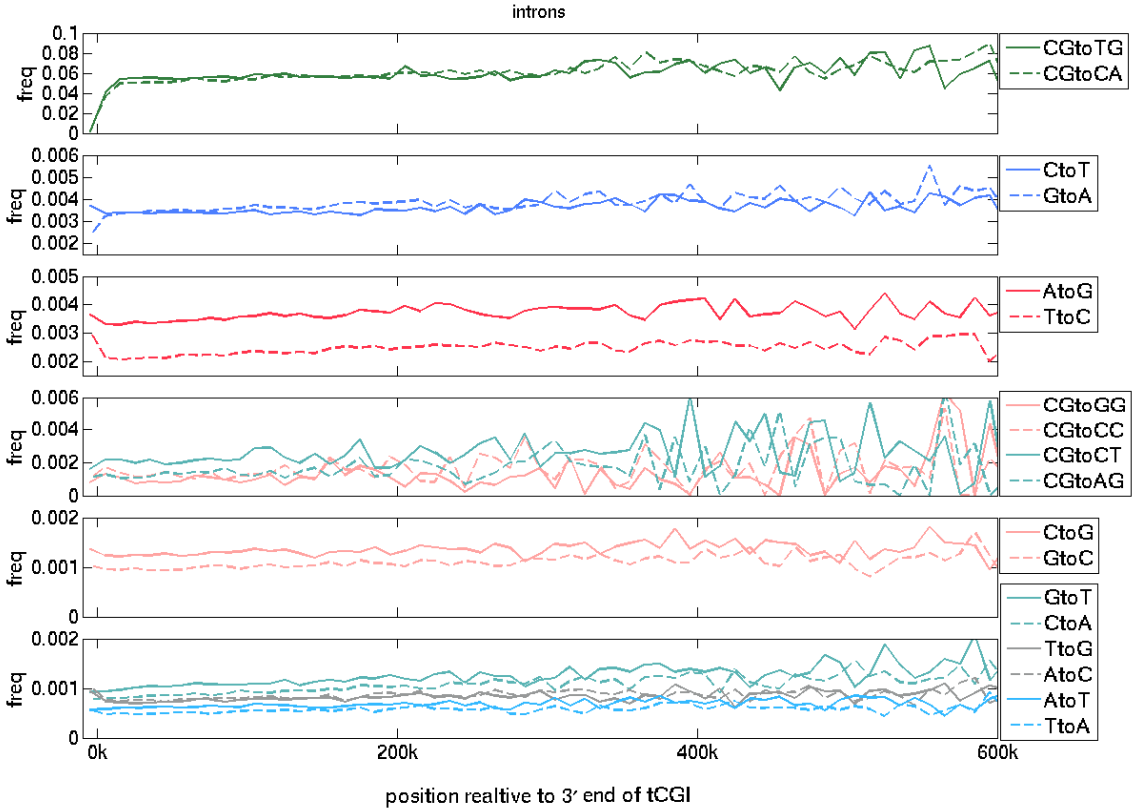
Supplementary Figure 2

Substitution rates within and around tCGIs (see Supplementary Figure 1 for further details).



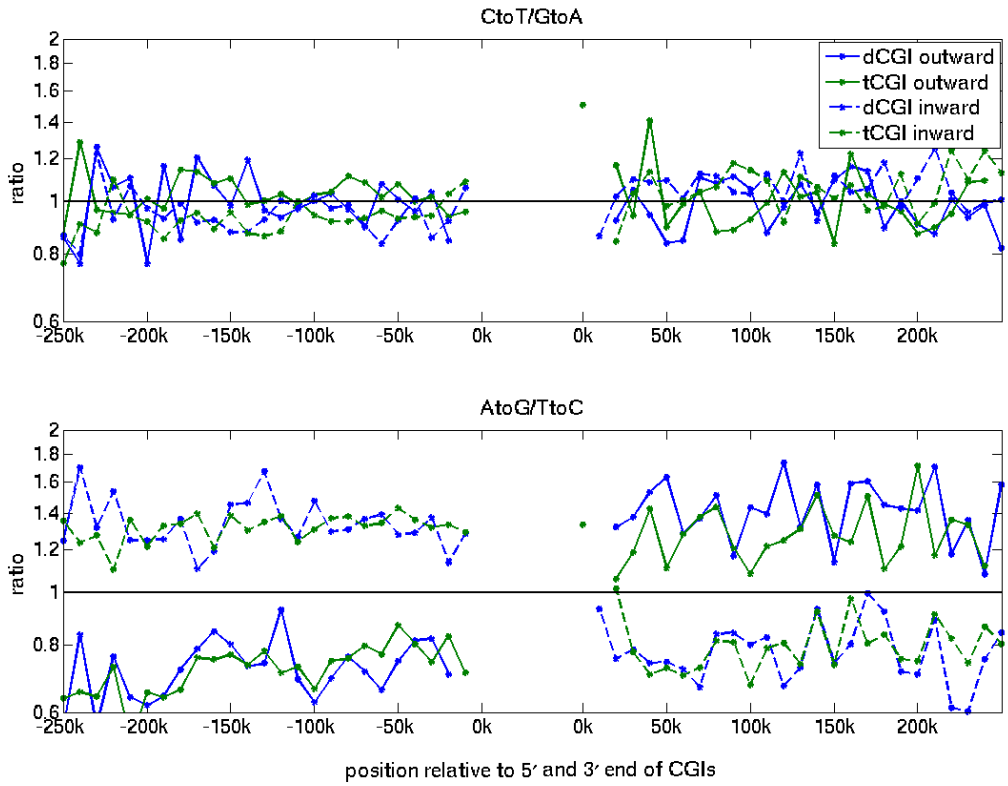
Supplementary Figure 3

Ratios of complementary transitions rates in intergenic plotted against distance from the middle point (0k) between consecutive CGIs calculated in 10 kbp long windows. All analyzed regions that were analyzed are interegenic. The middle point is defined between dCGI and tCGI (that are located left to the 0k) to their 5' nearest CGIs (right to the 0k).



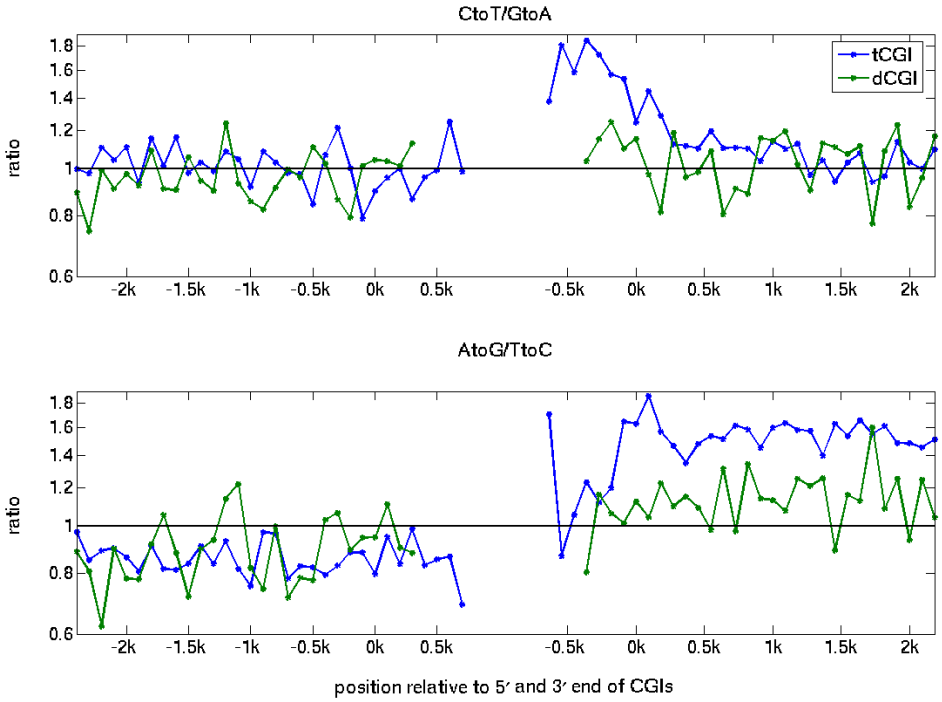
Supplementary Figure 4

Substitution rates in intronic regions of genes that have TSS inside of tCGIs. The positions are relative to the 3' end of tCGI (see Supplementary Figure 1 for further details).



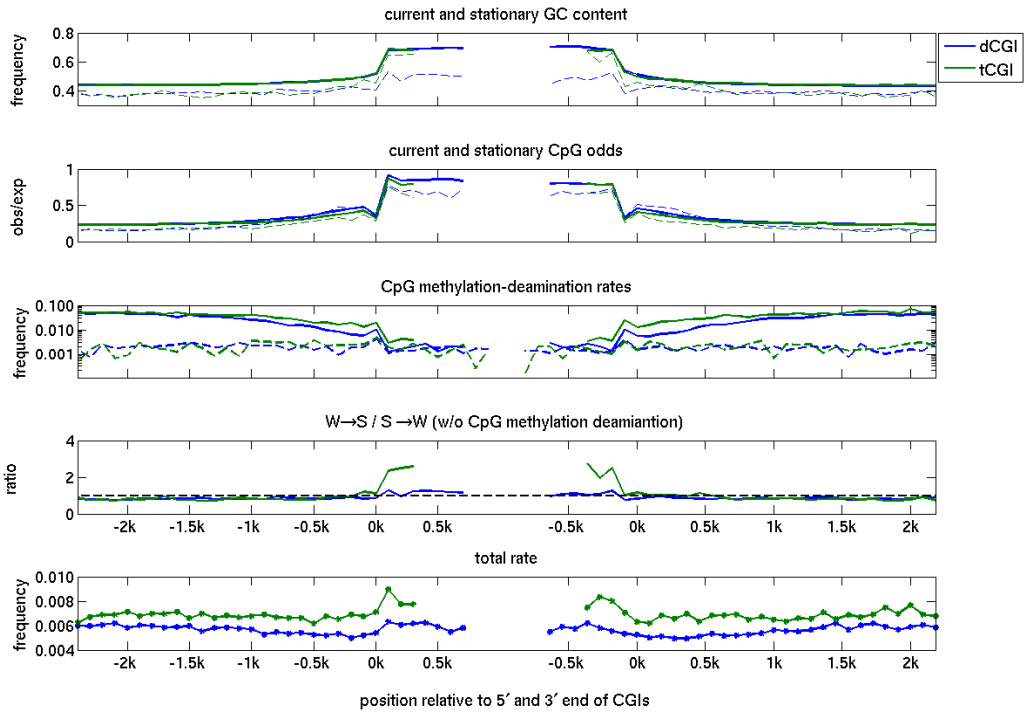
Supplementary Figure 5

Ratios of complementary transition rates in intronic regions of inward and outward genes relative to tCGIs and dCGIs. All rates are calculated using the reference strand as defined in the method section.



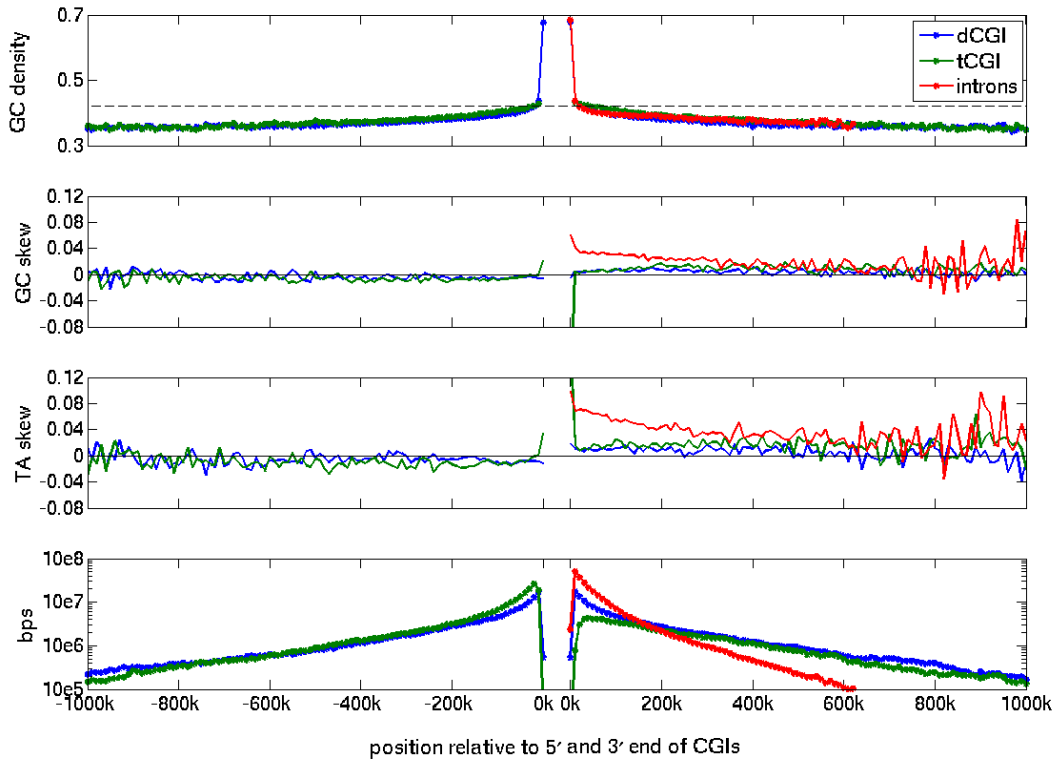
Supplementary Figure 6

Ratios of complementary transition rates are plotted against distance from the 5'end (left 0k) and 3'end (right 0k) of CGIs calculated in 100 bp long windows. The analyzed regions are non-exonic. The ratios that are presented between the left and the right 0k are calculated within the CGIs. The data points between the left and the right 0k are calculated within the CGIs. The data points between the left and the right 0k are calculated within the CGIs.



Supplementary Figure 7

GC content, CpG odds, CpG methylation-deamination rates, W->S/S->W ratio and total substitution rates in the proximity of 5' and 3' ends (left and right 0k, respectively) of dCGI and tCGI. In the top two panels, the current values of GC content and CpG odds (continuous lines) are compared with the corresponding stationary quantities (dashed lines). In the middle panel, [CpG->CpA+CpG->TpG]/2 (continuous) are compared with [CpG->CpT+CpG->ApG]/2 (dashed). The data points between the left and the right 0k are calculated within the CGIs.



Supplementary Figure 8

Statistical features of sequences that were used for the estimation substitution rates in Figure 2 and Supplementary Figures 1, 2 and 4. TA- and GC-skews are defined in the main text.