Transcription induces strand specific mutations at the 5’end of human genes

Paz Polak and Peter F Arndt

Supplemental Research Data

Supplemental Results and Discussion

Regional patterns of nucleotide composition

As was mentioned in the Introduction, the nucleotide composition in human genes is known to be dependent on the distance from the TSS. We repeated this analysis in the same DNA sequences that were used for substitution frequency estimation analysis in Figure 2. These sequences do not contain DNA insertions or deletions that happened after the split of human, chimpanzee and rhesus from their last common ancestor. The profiles of nucleotide composition showed four regional behaviors, as can be seen in Figure S3. First, the densities of C+G and CpGs near the TSS are much higher than genomic average and even within distances of 5 kbp (Saxonov et al. 2006). Second, TA skew =(T-A)/(T+A) and GC skew =(G-C)/(G+C) are positive in transcribed regions but not in the 5’ upstream regions (Touchon et al. 2003). Third, TA skew was found downstream to the 3’end (Louie et al. 2003). Last, the TA and GC skews within introns peak near the TSS and monotonically decrease as the distance from the TSS increases (Aerts et al. 2004; Touchon et al. 2003).

The current single nucleotide substitution rates can lead to TA and GC skews

The GC and TA skews in introns are suggested to be a by-product of a bias in complementary nucleotide substitution rates that is preceding the divergence of apes and new world monkeys. We tested if the recent substitution rates in humans can lead to TA and GC skews in intronic regions. In order to address this question, in each window (see Experimental Procedures), stationary nucleotide densities were derived by a genome-evolution model, which assumes evolution of DNA by neighbor-independent single-nucleotide substitution and the CpG process, with at the rates that were estimated in the vicinities of the 5' and 3'ends of genes (Fig. 2). Using this evolutionary model, we found that in upstream regions to the TSS, GC and TA skews are absent for stationary nucleotide-densities as in the case of the current nucleotide composition (Fig. S3). In similar fashion, in intronic regions, TA and GC skews were observed, but they were much stronger than the current skews (Fig. S3).

Strand asymmetries in UTRs, FFD sites and repeats

The substitution asymmetries that we measured in the main text can be shaped either by mutational processes or by selection pressures on introns. These two processes should leave two different mutational signatures. While mutational processes should not distinguish between the different parts of the transcript, selection effects are limited to functional elements. If these functional elements are intron specific, the asymmetries should be restricted to introns. We tested this hypothesis by analyzing the substitution rates in three different parts of transcript, 5' untranslated regions (UTR), 3'UTR and four fold degenerate (FFD) sites. We started by estimating the substitution rates in 5'UTR regions; however, since the average length of 5'UTR is about 200 bp long, we had enough data for estimating the rates in 5'UTR sequences, which overlap the first two 200 bp long windows downstream to the TSS. We found out that in 5'UTR sequences that are located within the first 200 bp of the transcript the rate of C->T (0.0034 per nucleotide) exceeds G->A (0.0025), whereas, the rates of G->A and T->C are similar (0.00311 and 0.003, respectively). In the next 200 bp long window, the C->T occur in higher frequencies than G->A, but the degree of asymmetry between these two rates is lower than in the first window (0.00314 and 0.0029 respectively), and the asymmetry of G->A (0.00284) versus T->C (0.25) becomes more pronounced. Comparison with intronic regions at similar distances from the TSS showed that the rates in 5'UTR are lower in about 20% than the rates in introns, but the ratios of C->T / G->A and of A->G / T->C were the same in introns and 5'UTR. Similar observations were made in the analysis of 3'UTR sequences.

The excess of A->G over T->C was detected in 3'UTR sequence that are located in the one kbp long region upstream to the 3'end of the genes (Fig. S4). As in 5'UTR sequences, the substitution rates in 3'UTR sequences are lower by about 20% compared to intronic regions (see in Fig. 2 and in Fig. S4), but the ratio of A->G / T->C was the same for both (Fig. S5). The lower substitution rates in 3'UTR and 5'UTR sequences indicates that they these regions are under negative selection.

For FFD sites we found the A->G rates being higher than T->C rates along the whole transcript (Fig. S6). The localized excess of C->T over G->A substitutions was also found in FFD sites located proximal to the TSS. However, in contrast to introns, this bias in the substitution rates is not statistically significant due to the low amount of sequence data on FFD sites.

The UTRs and FFD sites are assumed to evolve under stronger constraints than introns. On the other hand, repetitive elements are assumed to have less functional constraints. Indeed, the substitution rates in repetitive elements (Fig. S8) are higher than in non-repetitive regions (Fig. S9). Yet, the ratios between complementary transition rates were the same in repetitive and non-repetitive regions (Fig. 3).

The asymmetries are not due to functional elements in intron edges

The existence of the local and global asymmetries in different parts of the transcripts, which are under different level of evolutionary constraints, implies that the asymmetries are most likely invoked by a bias in the molecular mutational processes, as a result of selection acting on functional elements, which common to all different parts of the transcript including repeats. Candidates for such functional elements that are common to introns, UTR, and FFD sites would be splicing elements. These elements are usually found in 30 bp at the edges of introns and exons (Chamary and Hurst 2004; Hoffman and Birney 2007). Since previous works suggest that splicing element might be part of the 200 bp ends of introns (Louie et al. 2003; Touchon et al. 2004), we re-analyzed our data, using sequence data after excluding the 200 bp at each end of the introns. The removal of these elements did not affect the profile of the ratios of complementary substitution frequencies (Fig. 3); hence, it seems that the bias is not due to elements that are located in the edges of introns. Nevertheless, there are several works that suggest that the first introns are enriched with splicing motifs (Keightley and Gaffney 2003; Touchon et al. 2004). Even though, the first introns are usually long, and removing them leaves us with little sequence data for analysis, we could observe similar substitution patterns in introns which are not first introns (Fig. S7).

The local asymmetry is correlated to GC content and distance form the TSS

Both the distance from the TSS and the GC content are correlated with the local asymmetry. The excess of C->T over G->A is limited to the first 1 kbp region in the start of the transcript, a region that is also extremely enriched in G+C nucleotides (Fig. 3). It is not clear if the local asymmetry is correlated with the distance from the TSS or if these factors are correlated indirectly through a third factor of G+C density. In order to isolate the impact of the following two parameters GC content and distance length on the mutation spectrum, for each window in a certain length from the TSS we built collections of two types of sequences, GC-rich (above 50%) or GC-poor (below 40%). In GC-rich sequences, the ratios of C->T / G->A and of A->G / T->C were similar to the ratios in all genes (Fig. 3). The similarity between the ratios at the 2 kbp downstream is expected, because the enrichment of GC in these regions. The lack of the bias in C->T over G->A in GC-rich windows that are located further downstream from the TSS implies that this bias in the first 1 kbp of the transcript is not just due to the GC content (Fig. S10, Fig. 3). However, in GC-rich windows, the ratio between A->G over T->C is lower than the overall ratio and fluctuates between 1 and 1.2 (Fig. S10, Fig. 3); this can indicate that the GC-rich sequences that are located several kbp from the 5'end of genes are subject to different molecular mechanisms than GC-poor sequences. On the other hand, in GC-poor windows, the bias between C->T and G->A is absent at the immediate 1 kbp downstream to the TSS, but in the same regions A->G exceeds T->C (Fig. S11, Fig. 3). Therefore, it seems that the both two genomic parameters, the distance from the TSS together with the G+C density, are correlated with the localized asymmetry. Another interesting observation is the dependency of CpG transition rates in the local GC content (Duret and Arndt, PLoS Genetics (2008) in press). At distances of more than 2000 bp from the TSS, the CpG loss rates in GC-poor windows (Fig. S11) were more than 25% higher than in GC-rich windows (Fig. S10). The CpG loss rates in AT-rich sequences were 10 times the rates in GC-rich sequences near the TSS, whereas the C->T transitions rates in non-CpG sites in AT-rich sequences in the first kbp of the transcript was lower than in GC-rich sequences (Fig. S10, Fig. S11).

Transcription coupled repair

Substitution rates are the result of three processes: mutations, repair, and a fixation on the population level. The first two processes are at the molecular level, while the last process is a result of the dynamic at the population level. If we assume a model in which genetic drift is the main force i.e. almost every mutation has an equal chance of being fixated, the alteration in substitution rates along the genome is reduced to variations in the level of molecular processes, i.e., the probability that substitution occurs is a combination of the probability that DNA damage occurs and the chance that this damage will not be repaired correctly. In addition, it is reasonable to assume that most of the single nucleotide mutations are the result of single base damage. Strand asymmetries in damage mechanisms or in the repair processes are needed for asymmetry in complement nucleotide substitution rates. Green et al. (2003) suggested that the strand asymmetry is a result of strand specificity of TCR and a bias in misinsertion of A->G over T->C during replication which is not attributed to transcribed / non-transcribed differences of strands. In their model, any base change in the template strand will be correctly repaired by TCR and thus will not contribute to total base-pair transitions. Only base mutations in the non-transcribed strand contribute to transition rates, since a mismatch between the bases in the non-template and template strand are resolved by replacing the “original” base in the template strand with the complement of the mutated base in the non-template strand. Yet by itself, this doesn’t explain the bias between the complementary substitution rates. The substitution rates of A:T->G:C (where A and G are nucleotides in the non-template strand) are the sum of the rates of A:T->G:A->G:C and of the complement processes of A:T->A:C->G:C. Similarly, the complement process T:A->C:G is sum of the rates of T:A->C:A->C:G and of T:A->T:G->C:G. This strand specificity by TCR leads us to the conclusion that the dominant mutation pathways are A:T->G:T->G:C and T:A->C:A->C:G. Hence, one has to assume either that the rates of single-base mutations of the type A:T->G:T are higher than those of A:T->A:C or that there is a difference in the efficiency of repair of different mismatches, i.e. the rates of G:T->G:C are higher than C:A->C:G. These two scenarios are plausible, since it is known that the DNA polymerases are prone to do more misinsertions of G instead of A than of C instead of T (Mendelman et al. 1989). There is also evidence that the G: T mismatch is better recognized than C: A by the mismatched repair protein complex MSH2/MSH6 (Stojic et al. 2004) which lead to higher rates of T:G->C:G over C:A->C:G (since mismatches are resolved by using the non-template, regardless where the damage happened). TCR process doesn’t increase the mutation rates in transcribed regions, but it redistributes the mutations between the two strands. This can lead to an increased level of one substitution type in the transcribed strand compared to its level in the adjacent intergenic region. At the same time the complementary substitution rate decreases in the transcribed strand, as indeed it can be observed for the global asymmetry pattern (Fig. 2). In summary, the current model for the strand bias of

A->G over T->C in transcribed regions is explained by a combination of strand specificity of TCR together with other mutational biases for complementary nucleotides which are independent of transcription.

AID hotspots

There is growing evidence that AID is expressed in germ cells, and it has even been reported to express in the nucleus, which leads to the conjecture that AID can target ssDNA in these cells (Schreck et al. 2006). The probability that cytosine will be targeted by AID is affected by its flanking sequence; AID targets hot spots in the form of RGYW/WRCY, where R is purine, Y is pyrimidine and W is A or T (Beale et al. 2004; Pham et al. 2003; Rogozin and Diaz 2004).

We propose the mechanism of AID, since its mediated mutations (C->T) are confined to the first 1000 bp downstream to TSS regions of its target genes similar to the region we observed for the local asymmetry. In order to test this hypothesis, we divided the cytosine and guanine in each window into two sets, ones that are within hot spots, and the complementary ones not in hot spots and not part of CpG (see Supplemental Methods). Comparing the mutation patterns in these sets revealed that asymmetry does not appear in hot spots but rather in the non-hot-spots sites (Fig. S12). We therefore suggest that AID is not responsible for this localized pattern.

However, there are other DNA deaminases that target ssDNA which have different sequence specificities. For example, APOBEC3F has a preference to CCN or TCN (Beale et al. 2004). Mutations in these motifs are higher in our data set than in the complementary motifs (data not shown), however APOBEC3F has been shown to deaminate ssDNA just in the cytoplasm (Rosenberg et al. 2007) . Other DNA deaminases unfortunately have not yet been analyzed so extensively as AID.

Supplemental Methods

In this research we estimated the substitution pattern in different genomic context.

GC-rich and poor windows: To test the possibility of G+C specific effects, we also repeated the analyses on two subsets of the gene complement: GC-rich (>50% G+C) and GC-poor (<41%). In each window, the sequences that picked up from genes were comprised from intronic triple-alignments sequences of length longer than 50 bps (in human) and that were either GC-rich or GC-poor.

Repetitive elements: We also checked the repetitive elements impact on substitution rates by estimating the substitution in repetitive regions whose annotation was retrieved from Repeatmasker and in the remaining non-repetitive sequences.

UTR: The UTR sequences were retrieved by excluding of the transcripts introns and translatable exons. For the 5 ‘UTR, we had enough sequence data to estimate substitution rates in only two windows of length 200 bp.

Four fold degenerate sites (FFD): Amino acid in which are coded by 4 codons which differ only in their third position are called four fold degenerated (FFD). The third nucleotide in such FFD codons is called FFD site. In order to build a set of such sites, for each Ensemble gene we chose one transcript that is overlap a Refseq transcript. Since the number of FFD sites varies along the transcripts, we used windows of different lengths, which contain about 100,000 FFD sites. This number was determined because of statistical reasons and because it is the amount that we found in the first kbp. Such criteria left us with 3 windows in the 5 kbp regions downstream to the transcription starts; hence, we extended the region of analysis to 20 kbp past the TSS. The estimation of substitution in each gene was done by a Maximum Likelihood Method.

AID hot spots: AID tends to target cytosine which are part of specific sequence motifs, it has been reported that 50% of the mutations mediated by AID happened in target sites which are just 25% of all possible sequences (Pham et al. 2003). In order to test this effect, in each window we divided the cytosine into two parts, according to whether they were part of WRC/GYW (where W=A/T, R=G/A, Y=C/T) or in non-AID hot spots, which also did not originated from CpGs. The substitution rates were calculated by parsimony analysis, which attempts to attach the shortest evolutionary path from the ancestral sequence. In this analysis, we took into account only sequences where the sequences in rhesus and chimp were the same. Hence by parsimonious we assumed that ancestral sequence of human-chimp-rhesus was the one in chimp and rhesus, and the ancestral sequence was mutated in human after the eventual split between human and chimp.

Transcription status in ESC

Guenther et al. (2007) performed a genome wide mapping of histones modifications, that mark transcription initiation and elongation Their results reveal that many “inactive” (non expressed) genes harbour histone marks associated with transcription initiation at the vicinity of their 5’ends. We used Table S3 and Table S5 in the supplementary material of their paper. Table S3 is a list of genes with their chromosomal location of their TSS and the signal of different histone modifications.

The histone marker H3K4me3, is a marker for transcription initiation. We used a threshold of above two in the signal of H3K4me3 in order to determine if a gene experienced initiation. In addition, we used the chromosomal position of the TSS in order to associate an appropriate ensemble gene and genes ideas. The expression status of a gene was retrieved from Table S5 in that paper. This table provides a list of genes and their expression status in ESC, as it has been determined by 3 methods. We subdivided genes into expressed or non-expressed in ESC categories by using the absent/present calls in microarray expression data in ESC (see supplementary methods of Guenther et al. (2007) for further details on the methods).

Nucleotide composition

We calculated the nucleotide composition of human DNA sequences in non-overlapping windows (Fig. 1). The sequences that were used were the same as for the estimation of substitution rates. This estimation was not done for all human genes but only in human sequences that were part of the analysis. This means that we picked human sequences in which triple alignments were found and no alignment gaps were allowed. However, these results are in agreement with genome-wide analysis that was done previously (see above).

References

Aerts, S., G. Thijs, M. Dabrowski, Y. Moreau, and B. De Moor. 2004. Comprehensive analysis of the base composition around the transcription start site in Metazoa. *BMC Genomics* **5:** 34.

Beale, R.C., S.K. Petersen-Mahrt, I.N. Watt, R.S. Harris, C. Rada, and M.S. Neuberger. 2004. Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. *J. Mol. Biol.* **337:** 585.

Chamary, J.-V. and L.D. Hurst. 2004. Similar Rates but Different Modes of Sequence Evolution in Introns and at Exonic Silent Sites in Rodents: Evidence for Selectively Driven Codon Usage. *Mol Biol Evol* **21:** 1014-1023.

Green, P., B. Ewing, W. Miller, P.J. Thomas, and E.D. Green. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33:** 514 - 517.

Guenther, M.G., S.S. Levine, L.A. Boyer, R. Jaenisch, and R.A. Young. 2007. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell* **130:** 77-88.

Hoffman, M.M. and E. Birney. 2007. Estimating the Neutral Rate of Nucleotide Substitution Using Introns. *Mol Biol Evol* **24:** 522-531.

Keightley, P.D. and D.J. Gaffney. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *PNAS* **100:** 13402-13406.

Louie, E., J. Ott, and J. Majewski. 2003. Nucleotide Frequency Variation Across Human Genes. *Genome Res.* **13:** 2594-2601.

Mendelman, L.V., M.S. Boosalis, J. Petruska, and M.F. Goodman. 1989. Nearest neighbor influences on DNA polymerase insertion fidelity. *J. Biol. Chem.* **264:** 14415-11423.

Pham, P., R. Bransteitter, J. Petruska, and M.F. Goodman. 2003. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424:** 103-107.

Rogozin, I.B. and M. Diaz. 2004. Cutting Edge: DGYW/WRCH Is a Better Predictor of Mutability at G:C Bases in Ig Hypermutation Than the Widely Accepted RGYW/WRCY Motif and Probably Reflects a Two-Step Activation-Induced Cytidine Deaminase-Triggered Process. *J Immunol* **172:** 3382-3384.

Rosenberg, B.R., F.N. Papavasiliou, W.A. Frederick, and H. Tasuku. 2007. Beyond SHM and CSR: AID and Related Cytidine Deaminases in the Host Response to Viral Infection. In *Advances in Immunology*, pp. 215-244. Academic Press.

Saxonov, S., P. Berg, and D.L. Brutlag. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *PNAS* **103:** 1412-1417.

Schreck, S., M. Buettner, E. Kremmer, M. Bogdan, H. Herbst, and G. Niedobitek. 2006. Activation-induced cytidine deaminase (AID) is expressed in normal spermatogenesis but only infrequently in testicular germ cell tumours. *The Journal of Pathology* **210:** 26-31.

Stojic, L., R. Brun, and J. Jiricny. 2004. Mismatch repair and DNA damage signalling. *DNA Repair* **3:** 1091.

Touchon, M., A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes. 2004. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucl. Acids Res.* **32:** 4969-4978.

Touchon, M., S. Nicolay, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS letters* **555:** 579-582.

Supplemental Figures

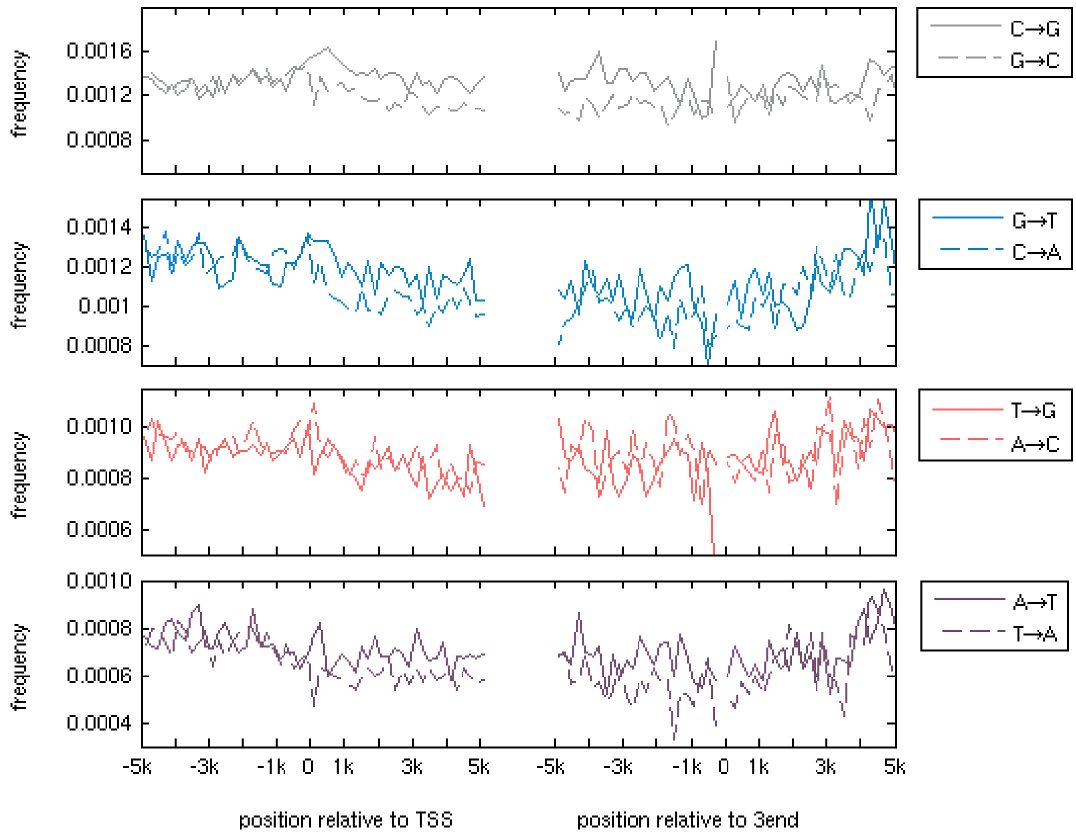

Figure S1. Transversion rates in introns and in intergenic regions in the vicinity of 5' and 3'ends of human genes

The plots show the estimated eight single nucleotide substitution rates deamination rates in non-overlapping 200 bp long windows. The distances of the windows' centers from the 5'end or 3'end are indicated on the x- axes. The estimation of substitution frequencies has been performed using the non-template strand.

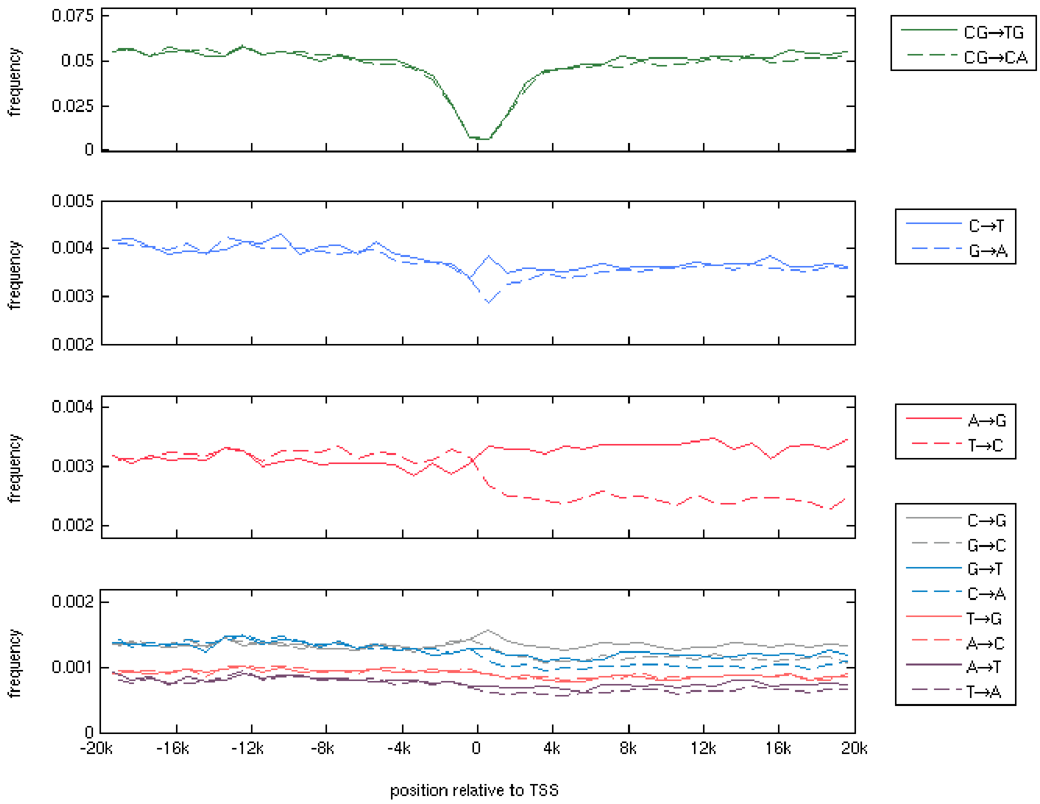

Figure S2. Substitution rates in introns and in intergenic regions in a 20 kbp long regions centered on the 5'end of genes.

The plots show the estimated twelve single nucleotide substitution rates, as well as, the CpG deamination rates in non-overlapping 1000 bp long windows (see Figure 1).

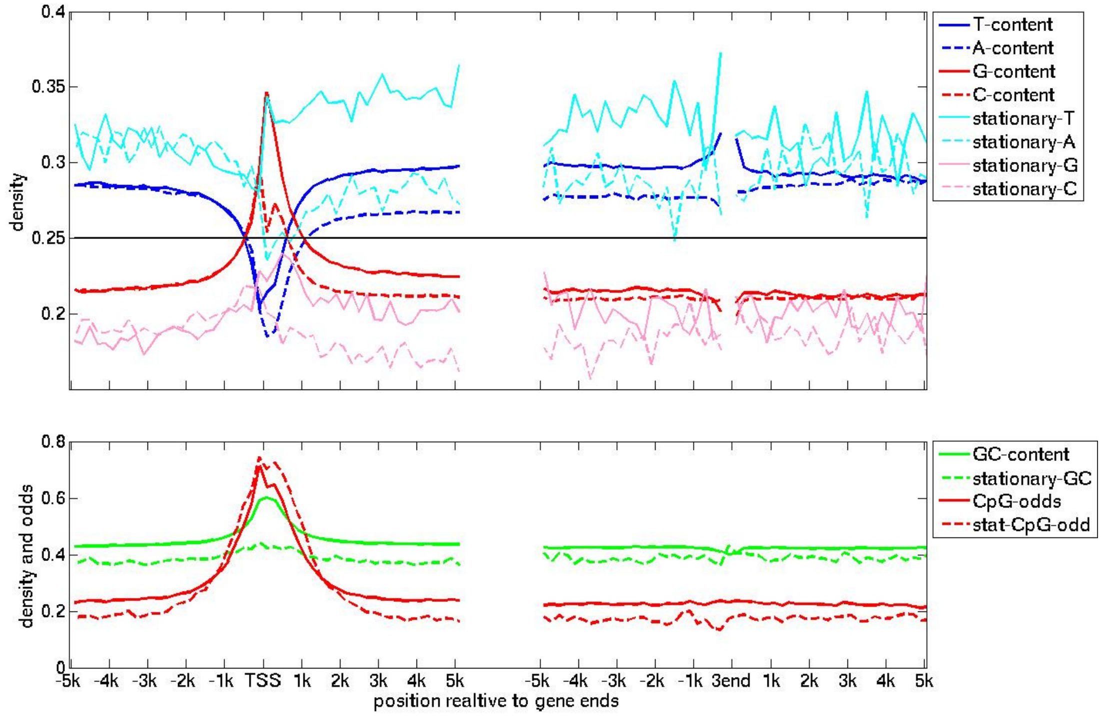

Figure S3. The current and stationary base-composition of introns and intergenic flanking regions of genes as a function of distance from the 5'end and from the 3'end of genes

The base composition is calculated in each window with the sequences that were used in Figure 2. The stationary base composition is the predicted genomic content in stationary state by genomic evolution model that used the estimated substitution rates as predicted in Figures 2 (See also Supplemental Results and Discussion and Supplemental Methods for further details). The CpG odd plot is frequency of observed/expected CpGs in each window.

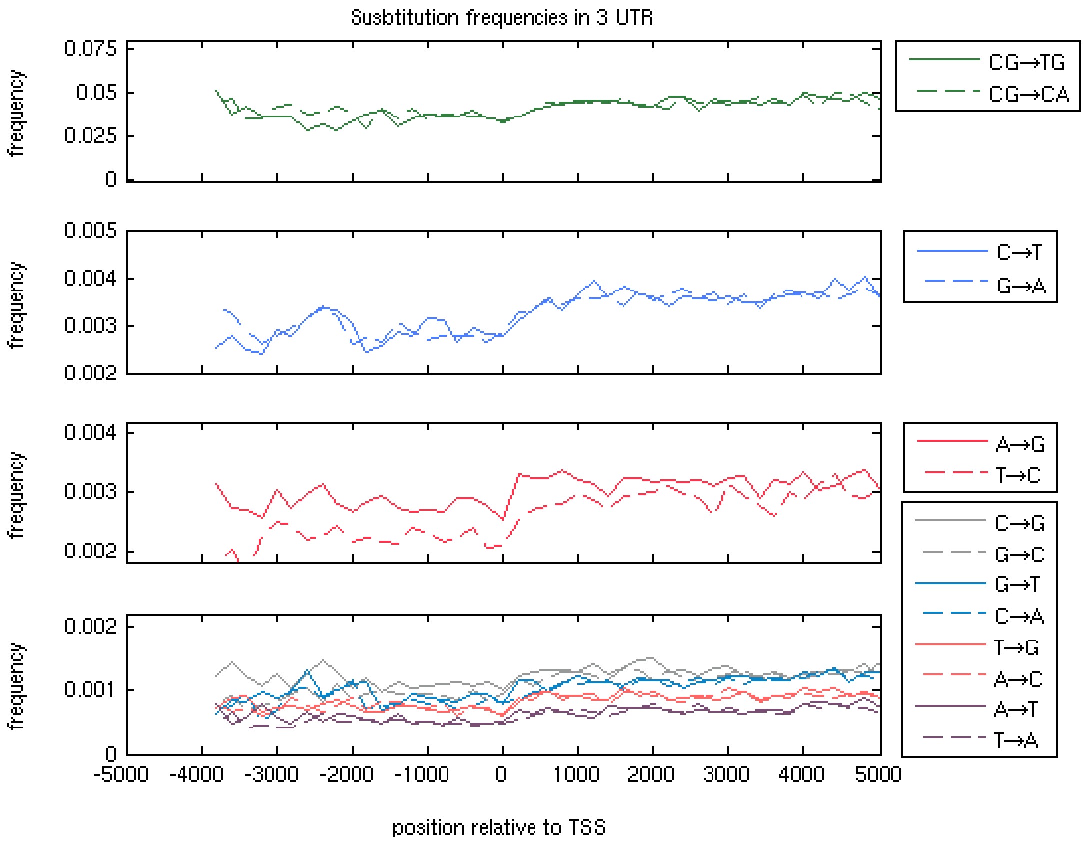

Figure S4. Substitution rates in 3'UTR and 3' downstream intergenic region of human genes.

Each point in a plot is the estimated substitution rate in a window of length 200 bp that is located at certain distance from the 3'end.

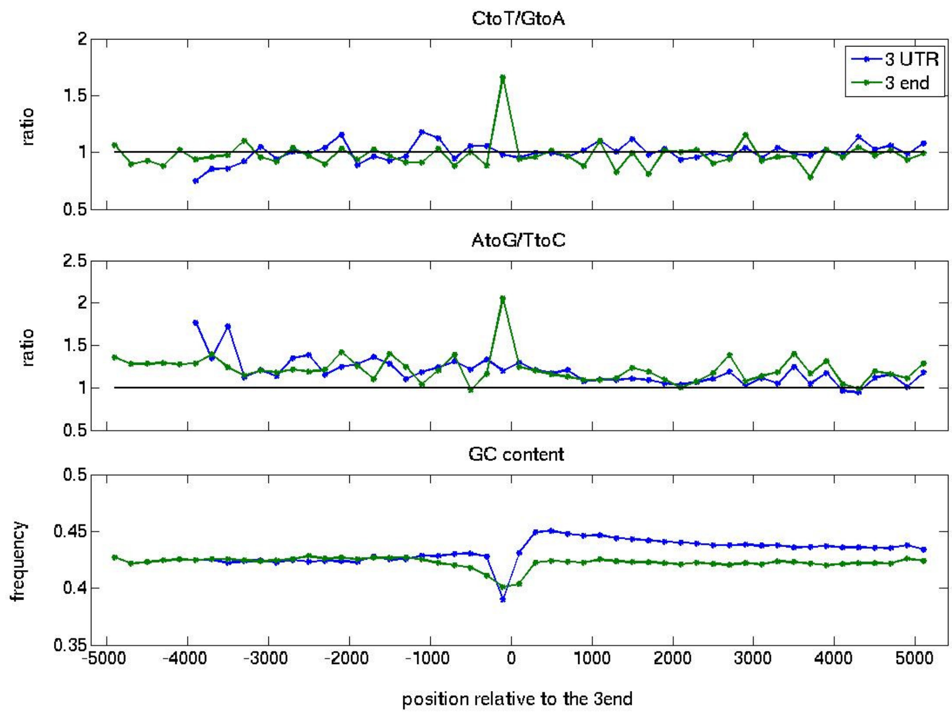

Figure S5. The ratios between complementary transition rates plotted against the distance from the 3'end of genes

The ratios in 3'UTR and introns were calculated using the estimated substitution rates in Figure S4 and Figure 2, respectively.

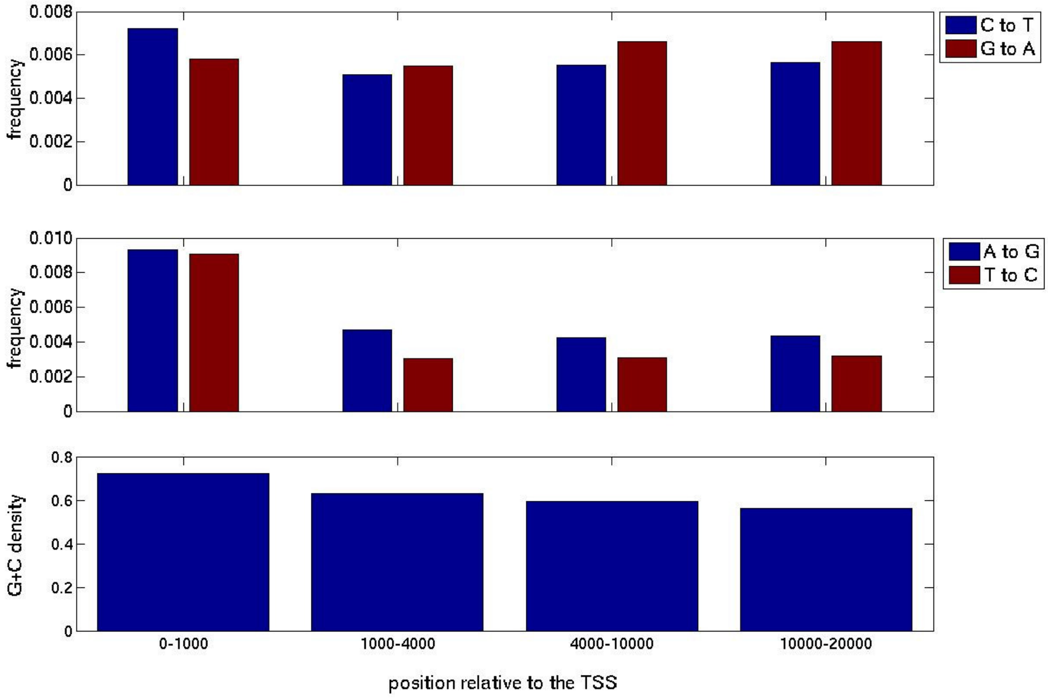

Figure S6. Transition rates and GC content in FFDs in non-CpG sites and in windows with varying lengths

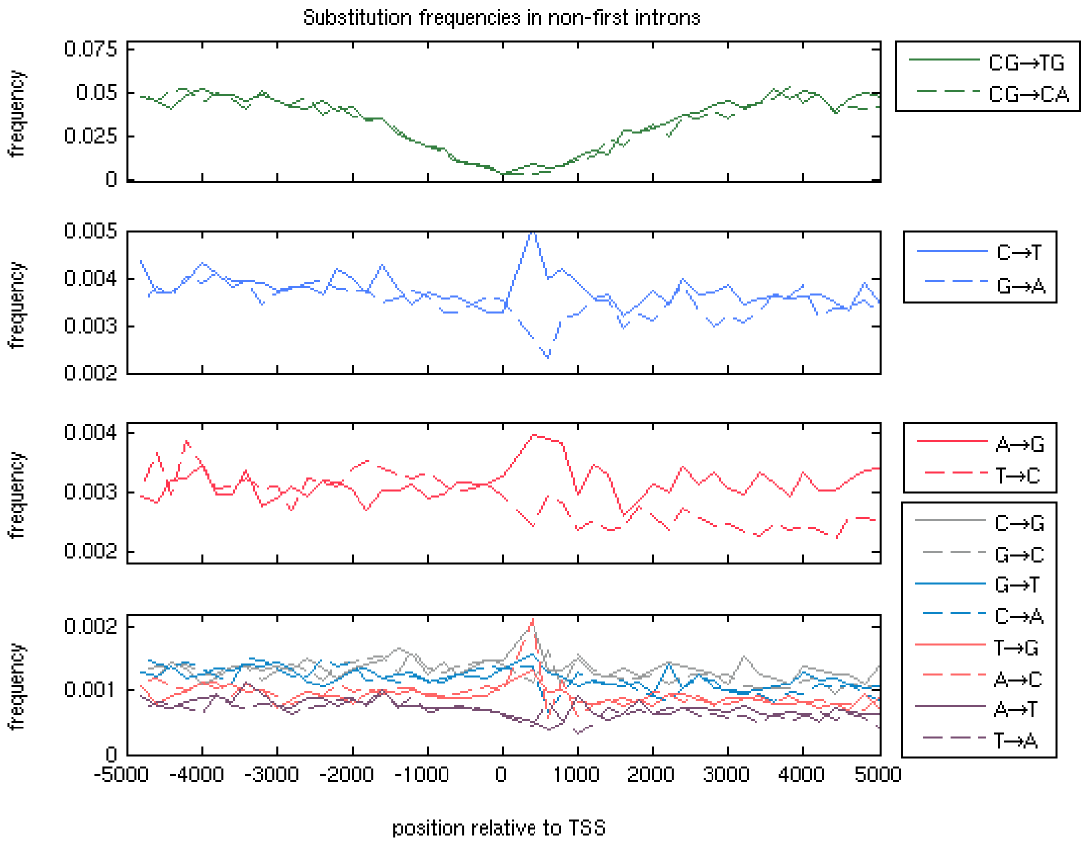

Figure S7. Substitution rates in DNA sequences in a 10 kbp long region centered on the 5'end after excluding first introns in addition to exons

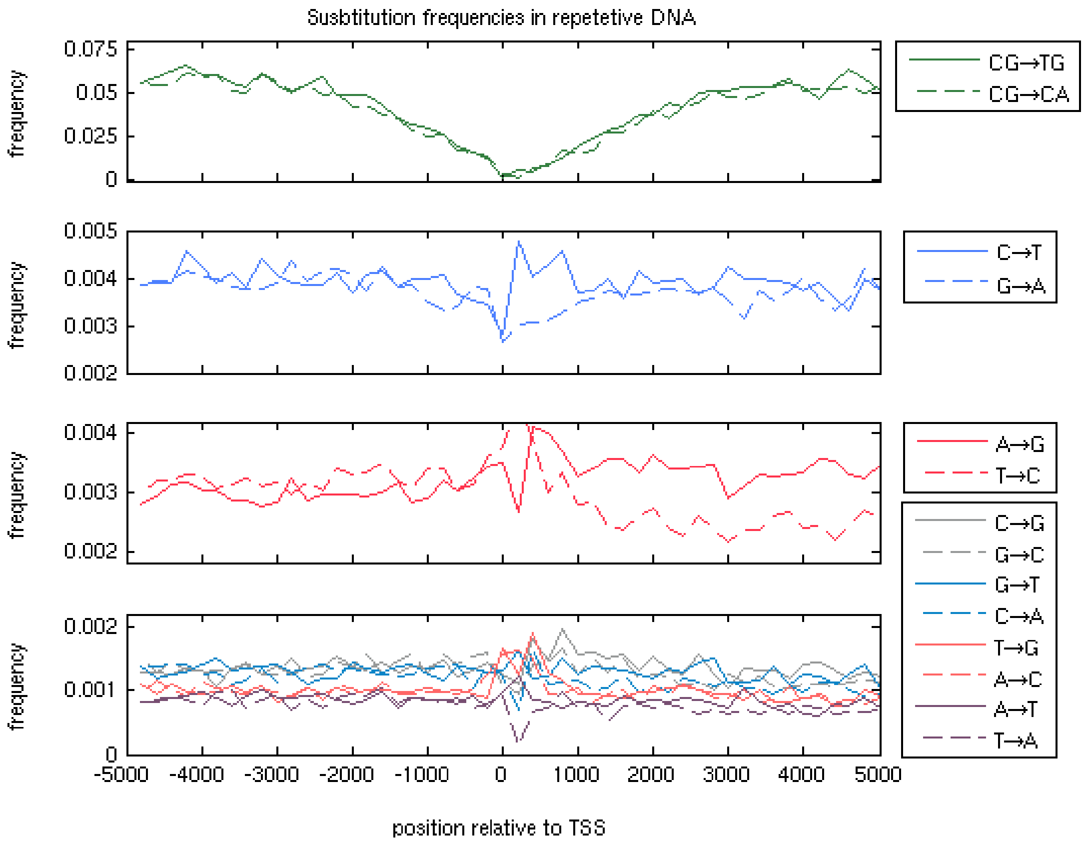

Figure S8. Substitution rates in repetitive DNA sequences in a 10 kbp region centered on the 5'end of genes

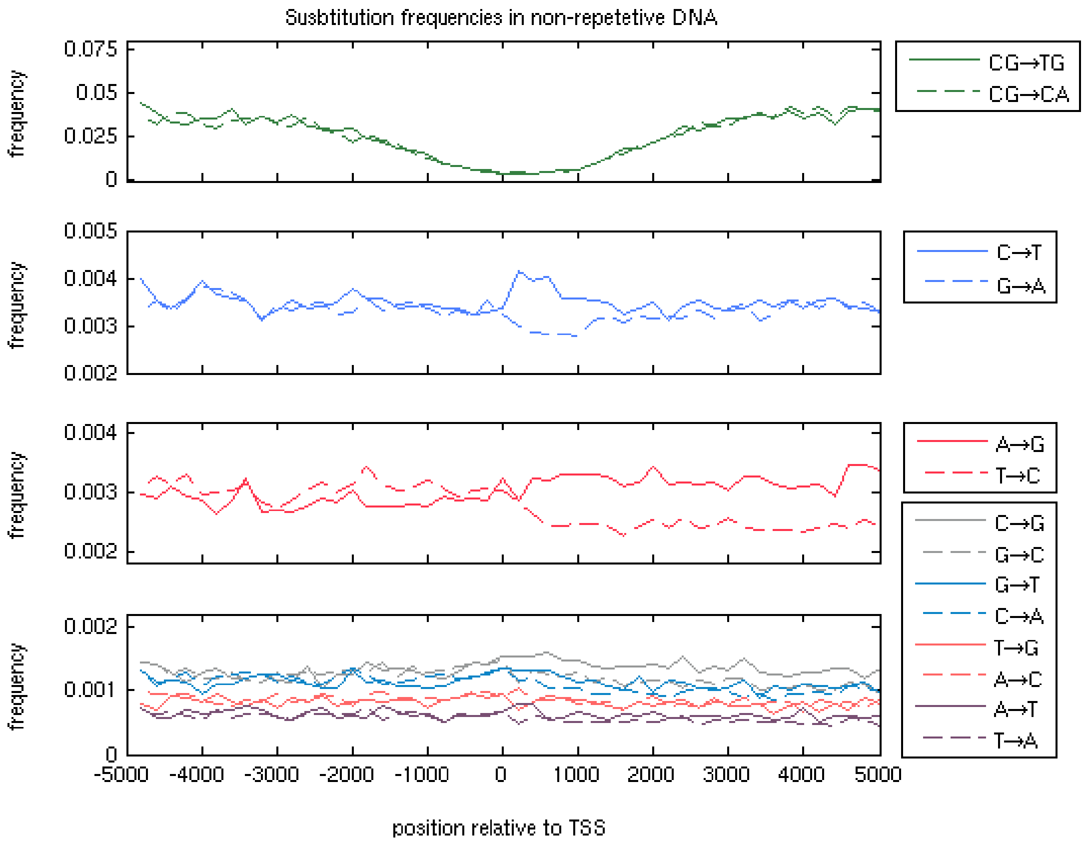

Figure S9. Substitution rates in non-repetitive sequences in a 10 kbp long region centered on the 5'end.

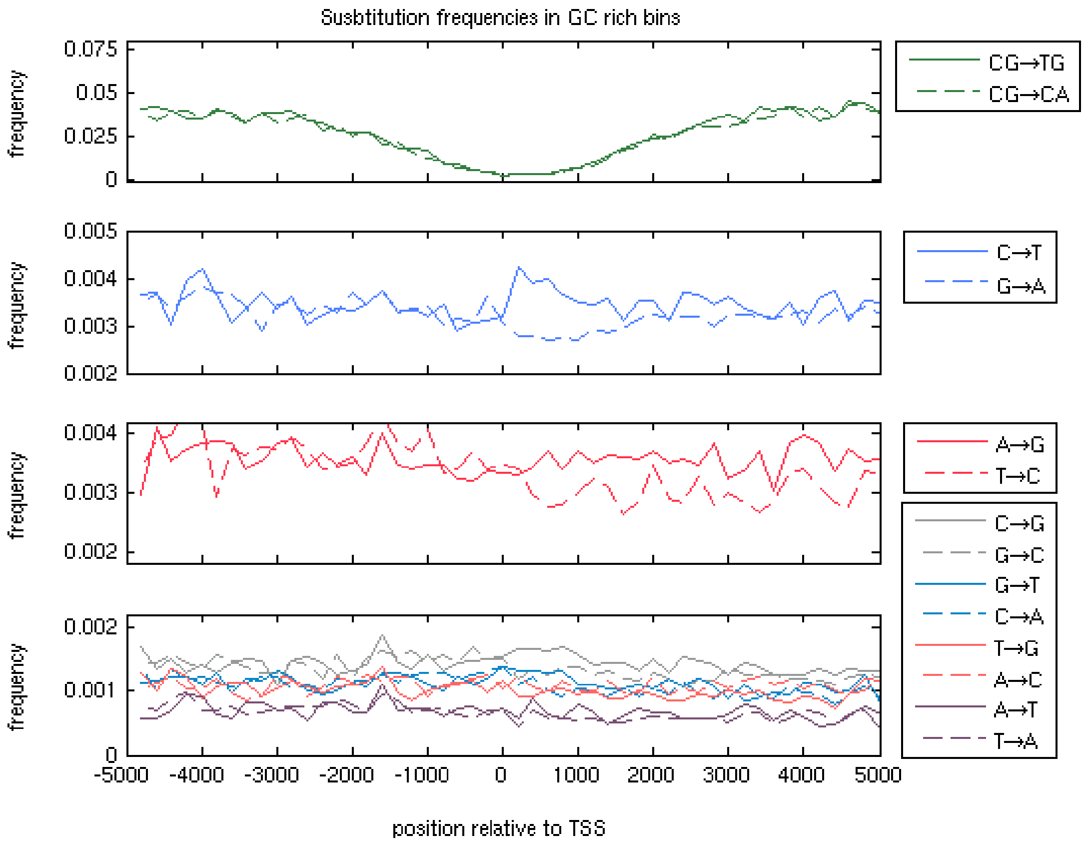

Figure S10. Substitution rates in GC-rich pooled sequences in a 10 kbp long region centered on the 5'end

In each window, just sequences with above 50% GC content were pooled out from all genes.

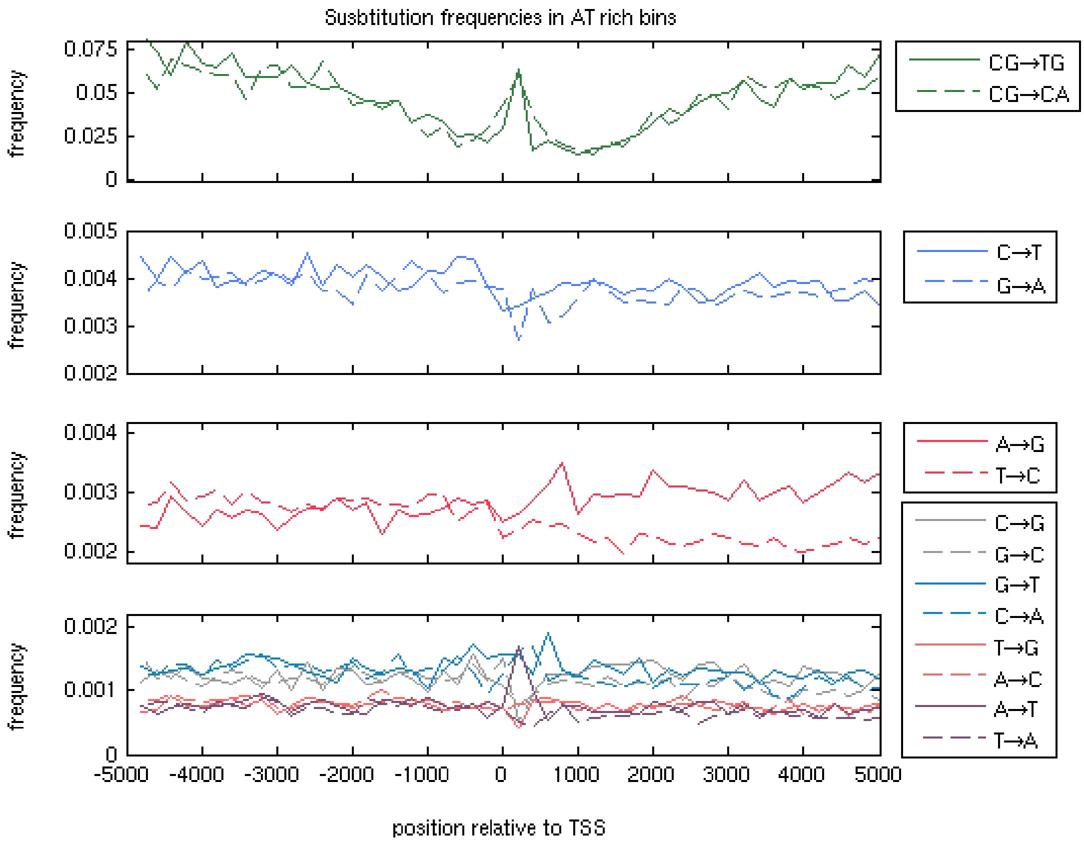

Figure S11. Substitution rates in AT-rich (GC-poor) pooled sequences in a 10 kbp long region centered on the 5'end.

In each window, just sequences with less then 41% GC content were pooled out from all genes.

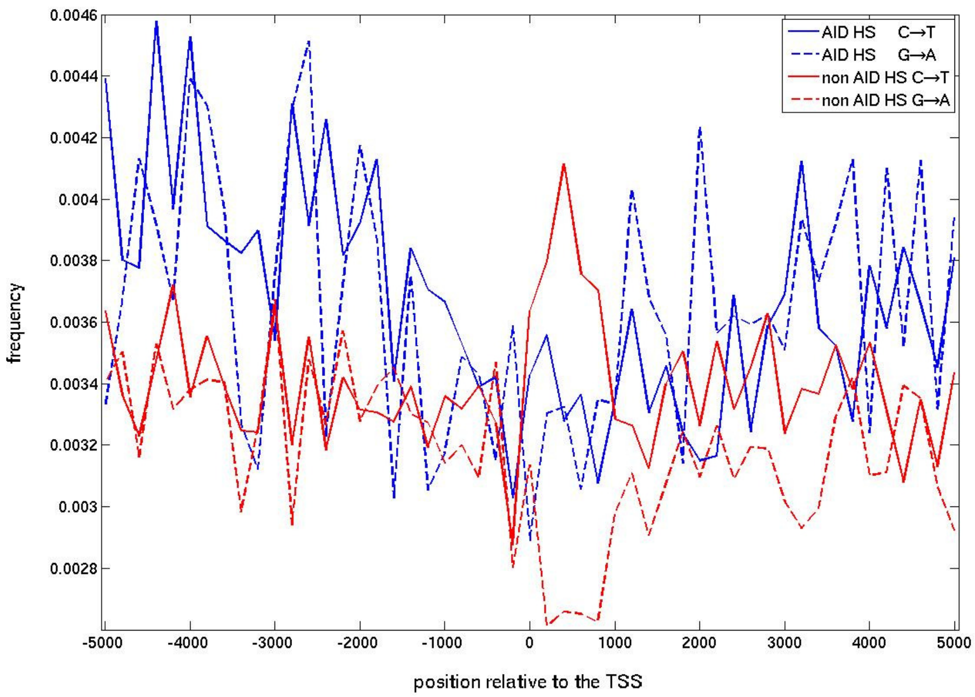

Figure S12. Transition frequencies of cytosine and guanine in/off AID hot spots

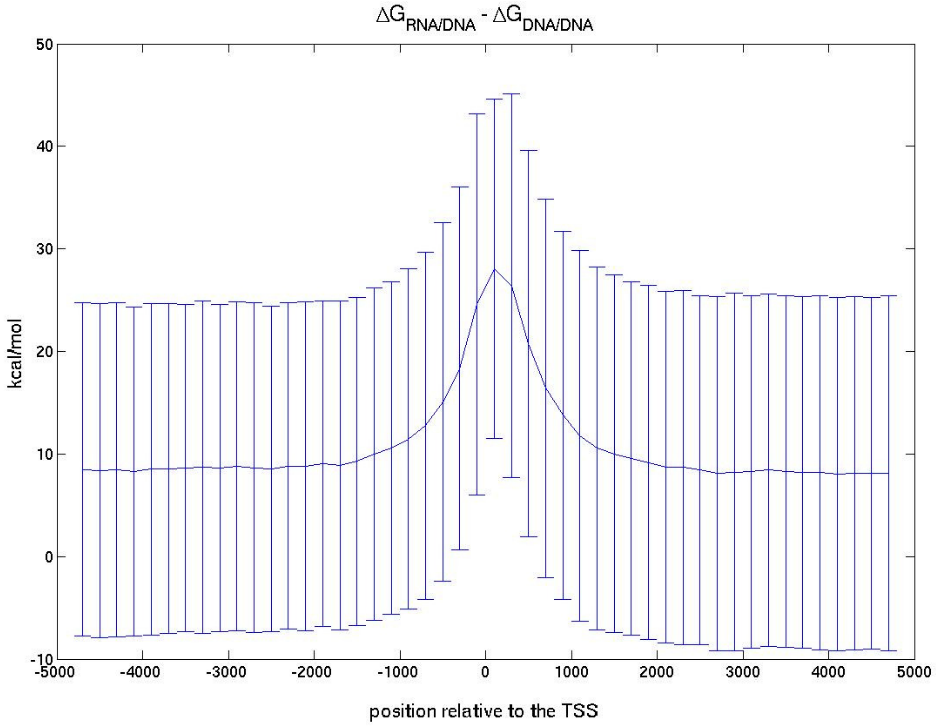

Figure S13. The mean (and the variation) of the energetic gap between the free energy of RNA/DNA and of DNA/DNA hybrids.

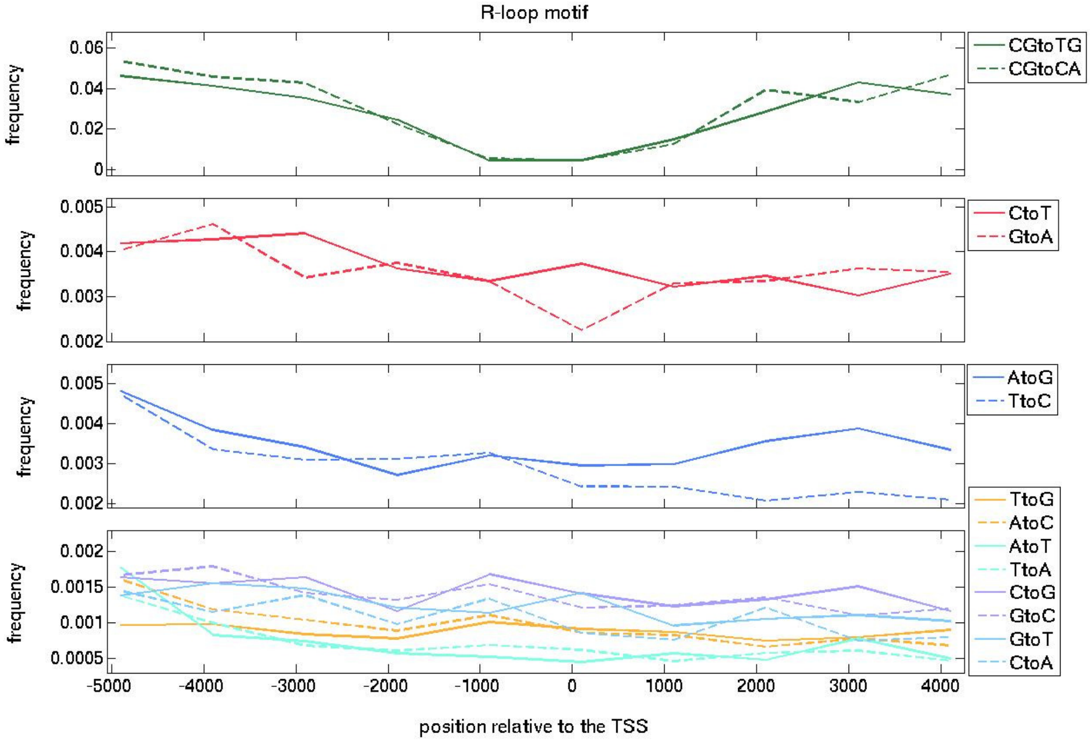

Figure S14. Substitution rates along genes that contain the R-loop initiation region (see main Text).

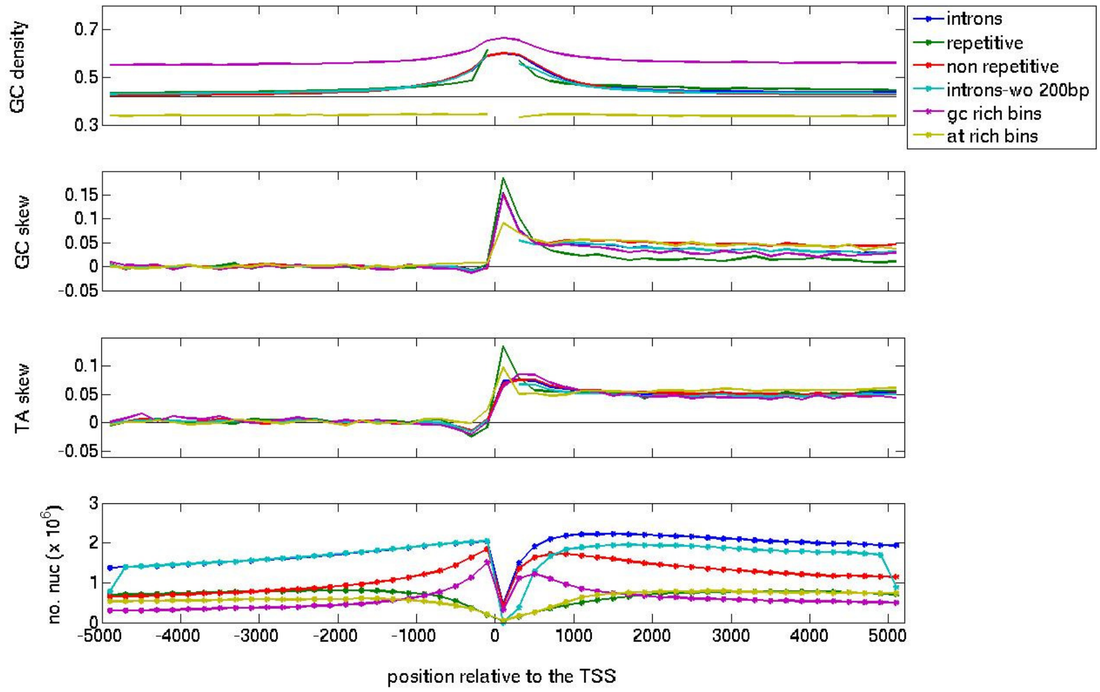

Figure S15. Statistical features of sequences that were used for the estimation substitution rates in Figure 3

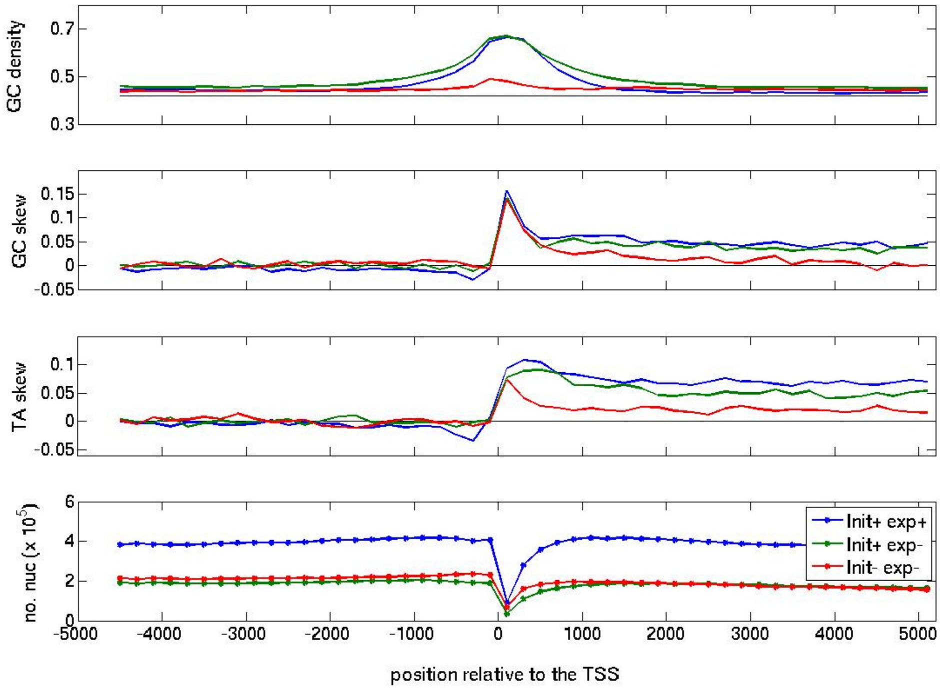

Figure S16. Statistics of the nucleotide composition that were used for the estimation substitution rates in Figure 4