GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination

Paul A Ginno, Yoong Wearn Lim, Paul L Lott, et al.

Genome Res. published online July 18, 2013
Access the most recent version at doi:10.1101/gr.158436.113

GC skew at the 5' and 3' Ends of Human Genes Links R-loop Formation to Epigenetic Regulation and Transcription Termination

Paul A. Ginno[1 * ‡], Yoong Wearn Lim [1 *], Paul L. Lott [2], Ian Korf [1,2],

and Frédéric Chédin [1,2 †]

[1] Department of Molecular and Cellular Biology, and [2] Genome Center, One Shields Avenue, University of California, Davis CA 95616.

* These authors contributed equally

‡ Present address: Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland

Running Title: GC skew and R-loops at the ends of human genes

Keywords: GC skew, R-loops, CpG islands, DNA methylation, Promoters, Epigenomics, Transcription Termination, *cis*-acting elements.

† To whom correspondence should be addressed. Email: flchedin@ucdavis.edu

Tel: 1-530-752-1800

Fax: 1-530-752-3085

Abstract

Strand asymmetry in the distribution of guanines and cytosines, measured by GC skew, predisposes DNA sequences towards R-loop formation upon transcription. Previous work revealed that GC skew and R-loop formation associate with a core set of unmethylated CpG island (CGI) promoters in the human genome. Here, we show that GC skew can distinguish four classes of promoters, including three types of CGI promoters, each associated with unique epigenetic and gene ontology signatures. In particular, we identify a strong and a weak class of CGI promoters and show that these loci are enriched in distinct chromosomal territories reflecting the intrinsic strength of their protection against DNA methylation. Interestingly, we show that strong CGI promoters are depleted from the X chromosome while weak CGIs are enriched, a property consistent with the acquisition of DNA methylation during dosage compensation. Furthermore, we identify a third class of CGI promoters based on its unique GC skew profile and show that this gene set is enriched for Polycomb group targets. Lastly, we show that nearly 2,000 genes harbor GC skew at their 3' ends and that these genes are preferentially located in gene-dense regions and tend to be closely arranged. Genomic profiling of R-loops accordingly showed that a large proportion of genes with terminal GC skew form R-loops at their 3'-ends, consistent with a role for these structures in permitting efficient transcription termination. Altogether, we show that GC skew and R-loop formation offer significant insights into the epigenetic regulation, genomic organization, and function of human genes.

Introduction

Epigenetic modifications in eukaryotes, including DNA methylation and histone marks, represent a critical layer of biological information employed to maintain genomic organization and stability and to regulate an array of nuclear processes such as transcription, replication, and recombination (C. David Allis 2007). Extensive mapping efforts have identified combinations of marks that define broad and conserved epigenetic domains in eukaryotes (Filion et al. 2010; Ernst et al. 2011; Roudier et al. 2011). Such patterns can predict the presence and activity of non-coding *cis*-acting regulatory elements such as promoters and enhancers. Despite these intense efforts, many fundamental questions regarding the establishment and maintenance of these epigenetic marks remain unanswered.

Elucidating the role that DNA sequence plays in determining epigenetic patterns is a promising area of investigation. While conventional thinking dictates that epigenetic patterns exist “above and beyond” the DNA sequence, epigenetic patterning often relies on sequence-driven recognition modules. In the case of DNA methylation, a critical mark associated with transcriptional silencing, plants make use of a specific RNA-directed DNA methylation system to target transposons (Wassenegger et al. 1994). In animals, a similar transposon-targeted silencing mechanism is carried out through the piRNA pathway (Aravin et al. 2008). In both systems, epigenetic modifications are deposited to specific sequences by virtue of small, trans-acting RNAs that likely guide epigenetic regulators through complementary RNA:DNA or RNA:RNA interactions. In vertebrates, numerous protein-DNA interactions contribute to guiding the establishment of DNA methylation patterns; *de novo* DNA methyltransferases possess significant preferences for certain sequences (Wienholz et al. 2010; Xie et al. 2012) and can be targeted to specific sites through interactions with DNA-binding proteins (Brenner and Fuks

2006; Hervouet et al. 2009). DNA sequence characteristics also play key roles in protecting loci from DNA methylation. Hence, certain DNA sequences can escape this prevalent modification because they are recognized and bound by specific proteins. This was elegantly demonstrated using the lacO-LacI system (Lin et al. 2000) and more recently by the identification of a large set of constitutive and lineage-specific hypomethylated regions highly enriched for DNA binding factors (Hodges et al. 2011; Lienert et al. 2011).

In vertebrate genomes, CpG islands (CGI) represent a key set of *cis*-acting loci that remain protected from DNA methylation. Such protection is essential given that CGIs function as promoters for ~60% of human genes, particularly broadly expressed “housekeeping” genes (Illingworth and Bird 2009). The mechanistic underpinnings of this protection are the subject of intense curiosity and accumulating evidence suggests that DNA sequence plays a key role in the process. The CFP1 protein specifically recognizes unmethylated CpG-rich regions by virtue of its CXXC domain and recruits the DNA methylation-repulsive H3K4me3 mark (Thomson et al. 2010), thereby ensuring the formation of a protected chromatin layer (Ooi et al. 2007). Motif analyses have also associated particular DNA sequences with methylated or unmethylated CGIs (Feltus et al. 2003; Feltus et al. 2006) and analysis of genomic methylation profiles revealed a propensity for G-rich sequences to resist DNA methylation (Bock et al. 2006; Straussman et al. 2009). We recently showed that unmethylated CGI promoters are characterized by a strong strand asymmetry in the distribution of G and C residues, a property known as GC skew (Ginno et al. 2012). GC skew undergoes a sharp transition at the transcription start site (TSS) and remains high in the first ~1 kb of the transcription unit, overlapping with the protected region. We further showed that transcription through regions of GC skew causes R-loop structures to form, in which the newly synthesized RNA hybridizes back to the template DNA strand and causes the non-template G-rich DNA strand to remain

looped out in a largely single-stranded conformation (Ginno et al. 2012). Further evidence showed that R-loop formation plays a functional role in protecting against DNA methylation. These findings suggest that the unmethylated CGI state is in part encoded in their DNA sequence and mediated by the co-transcriptional formation of R-loop structures. Here, we refine our analysis of GC skew patterns in the human genome and reveal new features associated with this DNA sequence characteristic.

RESULTS

GC skew patterns identify four broad promoter classes associated with unique epigenetic signatures and gene ontologies.

Class I promoters, corresponding to strong CGI promoters, were characterized by the highest CpG density and GC composition and a sharp rise in positive GC skew at the TSS (Figure 1A; 8,332 unique TSSs in this revised dataset). Genes associated with these promoters were prototypical housekeeping genes based on their gene ontologies and broad expression patterns (Ginno et al. 2012). Class II promoters defined a set of weaker CGIs characterized by shorter lengths (~700 bp as opposed to ~1,250 bp for Class I promoters), reduced CpG density and GC composition, and a sharply reduced shift in GC skew at the TSS (Figure 1B; 5,799 unique TSSs). Associated gene ontologies for Class II promoters were less strongly associated with housekeeping functions, as judged by p-values of enrichment and included genes involved with mitochondrial function, organelle biology, and ZNF type transcription factors (Supplementary Table 1). Their transcriptional output was significantly lower than that of Class I loci (Figure 1E). Class III promoters corresponded to CpG-poor loci which, for the most part (6,761 out of 7,968 TSSs) did not show strong GC skew as measured by our Hidden Markov algorithm, SkewR (Ginno et al. 2012) (Figure 1C). Gene ontologies associated with Class III promoters were strongly enriched for tissue-specific functions (Supplementary Table 1), in agreement with the fact that their expression levels are low in pluripotent cells (Figure 1E).

Intriguingly, we identified a fourth class of promoters (Class IV; 2,099 TSSs) overlapping with a strong peak of GC skew in the “reverse” orientation compared to the canonical Class I promoters (i.e., negative GC skew for genes transcribed on the + strand and vice-versa). Class IV promoters displayed CpG densities, GC percentages, and lengths comparable to Class I strong CGIs but they showed clear negative GC skew upstream of the TSS (Figure 1D). We

ruled out that this “reverse” pattern could be due to the presence of a large proportion of bidirectional promoters transcribing a Class I gene on the minus strand; annotated bidirectional genes have been filtered out from the data shown in Figure 1D. Significantly, Class IV genes were highly enriched for gene ontologies related to transcriptional regulation, morphogenesis, and cell fate commitment and included numerous genes encoding for transcription factors spread out among various gene families (Figure 1F and Supplementary Table 1). Class IV genes showed lower average expression than Class I genes in pluripotent cells (Figure 1E).

To delve deeper into the function of these promoters, we determined whether each promoter class could be associated with specific epigenetic signatures. For this, we analyzed the promoter proximal patterns of a number of epigenetic marks, including DNA methylation and a variety of histone modifications, as measured in human embryonic stem cells. In the case of DNA methylation, CGI promoters (Class I, II and IV) all showed protection around the TSS (Figure 2A), while Class III promoters were susceptible to this mark, as previously observed (Ginno et al. 2012). Amongst CGI promoters, Classes I and Class IV showed strong protection around the TSS, while Class II promoters were characterized by a weaker and narrower protected region, consistent with the respective CpG density and GC skew of these three classes. In terms of histone modifications, Class I and Class II CGIs could be most readily distinguished by their profiles of H4K20me1 and H3K79me2 deposition (Figure 2B, C). These two marks have been linked to transcription initiation and/or elongation (Nguyen and Zhang 2011; Beck et al. 2012) and their levels are tightly correlated with gene expression (Karlic et al. 2010; Vavouri and Lehner 2012). In both cases, Class I CGIs showed a higher density of modification over Class II, while Class IV CGIs were intermediate. Modest differences could be observed for the distribution of H3K4me3, H3K36me3, and the H2A.z histone variant among the 4 classes (Supplemental Figure 1). Interestingly, Class IV CGI promoters were enriched for the

Polycomb-mediated H3K27me3 mark (Figure 2D). In addition, these promoters were preferentially associated with the H3K27 methyltransferase EZH2, and the RING1B protein, two key members of the PRC2 and PRC1 complexes, respectively (Figure 2E, F). This shows that Class IV CGIs represent a subset of CGIs that are more likely to recruit and be regulated by the Polycomb regulatory complexes (PRCs). This is consistent with the observed gene ontologies associated with Class IV promoters, namely enrichment for developmental regulators, transcription factors, and other well-known developmentally regulated genes that are Polycomb-group targets. Altogether, these data reveal that GC skew is a useful DNA sequence metric for identifying unique promoter classes associated with distinct gene ontologies and epigenetic signatures.

Weak and strong CGI promoters show distinct genomic distributions reflecting fluctuations in chromosomal gene density.

To further analyze possible differences between Class I and Class II CGI promoters, we determined the genomic distribution of these loci on a chromosome by chromosome basis. Interestingly, Class I promoters were more likely to be located on gene-poor chromosomes as evidenced by a clear negative correlation between the percentage of Class I genes on a given chromosome and gene density (Figure 3A). In sharp contrast, Class II promoters showed a clear positive correlation with gene density (Figure 3B). In fact, nearly half of the Class II genes mapped to the 8 most gene-rich chromosomes, particularly chromosomes 19 and 17, for which they showed highly significant enrichment (data not shown), while Class I promoters were highly enriched on chromosomes 13 and 18, among the most gene-poor chromosomes. In a similar analysis, the distribution of Class III or IV promoters on individual chromosomes was only minimally affected by overall gene densities (Supplemental Figure 2). This suggests that the type of promoter CGI (Class I vs. Class II) and the accompanying strength of protection against

DNA methylation (strong vs. weak, respectively) reflect the genomic architecture and gene density of chromosomal territories.

The X chromosome shows a unique distribution of Class I and II promoters that may reflect dosage compensation.

We observed a striking exception to the correlations between gene density and the chromosomal distribution of Class I and II promoters for the X chromosome. Promoters of X-linked genes showed depletion for Class I promoters (Figure 3A) and enrichment for Class II promoters (Figure 3B). These deviations from the autosomal trends take on particular biological significance given that the X chromosome undergoes inactivation in females, a process characterized by the acquisition of DNA methylation at hundreds of promoters along this chromosome. Because of this unique process, X-linked promoters are likely under specific evolutionary constraints to enable epigenetic silencing and dosage compensation on the inactive X chromosome while at the same time retaining sufficient protection to ensure expression on the active X. The depletion of strongly protected Class I promoters and the concomitant enrichment of weakly protected Class II promoters observed on the X seem to satisfy this requirement. This hypothesis predicts that the relative depletion of Class I promoters and enrichment of Class II promoters should correlate with X-inactivation efficiency along the chromosome. Indeed, we find evidence that the most evolutionarily ancient X chromosome strata (XCR), which undergoes X-inactivation to nearly 100% efficiency, showed the lowest representation of Class I promoters (Figure 3D left). The S2a strata, which undergoes inactivation with ~90% efficiency, showed slightly more Class I promoters than XCR. Likewise, the evolutionarily more recent XAR strata, which was acquired from autosomes and undergoes X-inactivation to a lower degree (~71%, (Carrel and Willard 2005)), showed a frequency of Class I promoters below what is expected from the autosomal distribution. Finally, the frequency of Class I promoters on the short pseudo-autosomal PAR1 region was consistent with the

autosomal average. In contrast, Class II promoters showed the opposite distribution patterns across the X-chromosome (Fig 3D right), supporting the contention that the relative distribution of Class I and II CGI promoters on the X may have evolved in response to dosage compensation. It should be noted that, in addition to these findings, Class III (CpG-poor) promoters actually represented the largest class of promoters on the X (42%, versus 29% and 26% for Class I and II, respectively). This is consistent with the X chromosome being enriched for genes involved in germline- and brain-specific functions.

Genes in gene-rich neighborhoods are enriched for GC skew at their 3'-ends.

To profile GC skew patterns more thoroughly, we decreased the stringency of our SkewR Hidden Markov Model (HMM) and re-analyzed the human genome (see Methods for details). This led to a total of ~66,000 peaks as compared to ~19,000 under the previous model. As expected, promoter annotations still represented the largest signal in the dataset and Class I and Class II genes were now part of one large GC-skewed promoter class. Interestingly, the 3'-ends of human genes now represented a significant category of all GC skew peaks. Using a window of -500 +1500 bp around the transcription termination signal (TTS) of all genes in our gene list, we observed that a total of 2,044 TTSs associate with peaks of GC skew conducive to R-loop formation (i.e. for which positive GC skew is co-oriented with transcription) at their 3' ends. Metaplot analyses further revealed that, as was observed at the 5'-end of genes, GC skew underwent a sharp transition at the TTS and rapidly decayed in the 1-2 kb downstream of TTS (Figure 4A). The average amplitude of the shift in GC skew at the TTS was lower than that observed for Class I CGI promoters but nonetheless equal to or stronger than that of Class II CGI promoters. While a clear shift is observed precisely at the TTS, it is interesting to note that the start sites of the 3' GC skew peaks are distributed in a ~ 3 kb window around the TTS such that a significant fraction of genes may encounter GC skew even before reaching the polyA

signal. The frequent occurrence of terminal GC skew suggests that thousands of human genes may experience R-loop formation at their 3'-ends.

To further address the biological significance of 3' GC skew, we performed gene ontology analysis. Unlike promoter classes, genes with terminal GC skew were not enriched in any particular category (data not shown), arguing that functional classification is not the main determinant of 3' GC skew. We next asked whether gene density was a factor in determining 3' GC skew enrichment. For this, we analyzed the distribution of genes with 3' GC skew on a chromosome basis as a function of gene density. This revealed a clear positive correlation between both parameters (Figure 4B). Thus, genes located on gene-rich regions are much more likely to show 3' GC skew compared to genes located on gene-poor chromosomes. Interestingly, the X chromosome fit well within this trend, unlike previous promoter-centric observations (Figure 3). This suggests that 3' GC skew may have little, if anything, to do with X-inactivation or epigenetic regulation but that instead, it is determined by other constraints related to local gene density. In support of this, we determined that the distance separating the TTS of a gene with 3' GC skew to the nearest neighboring TSS or TTS (whichever was closest) was 30 kb on average. In comparison, the TTS of genes without terminal GC skew were separated by an average of 120 kb from their nearest neighbor, a statistically significant difference (p-value < 0.0001; Mann-Whitney test). Furthermore, 40.7% of genes with terminal GC skew had their TTS located 2 kb or closer from the nearest downstream gene while only 8.8% of non-skewed TTS could be found within the same distance from their nearest neighbor (Figure 4C). Of these closely arranged genes, genes with terminal GC skew were ~12 times more likely to lie in a tandem arrangement with their neighbor compared to genes devoid of terminal GC skew (Figure 4C). Within the class of closely arranged genes with terminal GC skew, tandem oriented genes were also more frequent than converging genes (Figure 4C). Interestingly, 68% of genes with 3'

GC skew also showed 5', or promoter, GC skew, arguing that GC skew delineates the beginning and end of a significant number of human genes.

GC skew leads to frequent R-loop formation at the 5' and 3' ends of human genes

GC skew strongly predisposes a sequence to form R-loops upon transcription. In order to profile these structures genome-wide, we used a previously developed technique termed DRIP-seq (DNA:RNA immunoprecipitation coupled to sequencing; (Ginno et al. 2012)). Two independent DRIP-seq experiments were conducted using genomic DNA extracted from human pluripotent Ntera2 cells. To improve the resolution of DRIP-seq, the DNA was fragmented with two different cocktails of restriction enzymes chosen to cleave the target DNA in a distinct and complementary way (see Methods for details). After immunoprecipitation, high-throughput sequencing and computational mapping of the sequencing reads back to the human reference genome, DRIP signal was assigned back to restriction fragments and consensus DRIP-seq peaks were called if overlapping DRIP peaks could be identified in both datasets. This resulted in a total of 4,181 consensus DRIP-seq peaks. This likely represents a sharp underestimate of the total number of R-loop peaks given the imperfections of the distributions of cleavage sites in the genome (i.e., peaks identified in one dataset often landed in regions of dense cleavage in the other dataset, thus precluding their identification in both datasets).

Location analysis revealed that nearly two-thirds of these stringent DRIP-seq peaks mapped to the 5' end (1,587 peaks) or the 3' end (1,052 peaks) of human genes. Representative examples of DRIP profiles at the TSS (Figure 5A) and TTS (Figure 5B) are shown. Figure 5C shows an example where an R-loop formed at the TTS of two convergent genes, while Figure 5D illustrates a gene with both TSS and TTS R-loop peaks. Shifts in GC

skew and R-loop formation are therefore a feature of thousands of genes in the human genome. As expected from the dependence of R-loop formation on the superior thermodynamic stability of RNA:DNA hybrids formed with a G-rich RNA, DRIP-seq peaks were highly enriched for highly skewed CGI promoters (Class I; 956 peaks), less enriched for weakly skewed CGI promoters (Class II; 350 peaks), and even less so for non-CGI promoters (Class III; 127 peaks) and "reverse" CGIs promoters (Class IV; 154 peaks) (Figure 5E). DRIP-seq peaks mapping to the 3'-end of genes were also much more frequent at regions harboring GC skew (805 peaks) than at regions without GC skew (247 peaks). R-loop formation at the 5' and 3'-ends of human genes therefore strongly correlates with GC skew.

GC skew is highly correlated with the unmethylated epigenetic state over the length of the first exon and at the 3' end of genes.

To further investigate the relationship between GC skew and epigenetic states, we focused on Class I promoters and clustered them into three subclasses based on the length of GC skew downstream of the TSS: promoters with short skew (less than 300 bp; 1,288 promoters), medium skew (between 300 and 700 bp; 4,674 promoters), and long skew (more than 700 bp; 2,160 promoters) (Figure 6A). Interestingly, CpG density across all three subclasses rose almost at the same point upstream of the TSS (5' boundary) but extended downstream of the TSS in a manner directly proportional to the length of GC skew (Figure 6B). A similar trend was also observed for GC content, indicating that longer GC skew downstream of the TSS correlates with the extension of the 3' boundaries of CGIs (Figure 6C). Given that high CpG densities (CpG o/e close to 1) are only observed in the human genome if the corresponding CpG sites are protected from DNA methylation (Illingworth and Bird 2009), it is not surprising that the length of the GC skew downstream of TSS was highly correlated with the length of the DNA methylation-free region across all three subclasses (Figure 6D). This

reinforces the notion that GC skew is a key sequence feature of unmethylated CGI promoters that correlates precisely with the 3' boundary of CGIs. A similar relationship was observed between the length of GC skew and the length of the peak of H3K4me3 (Figure 6E). Promoters with short GC skew showed the narrowest peak of H3K4me3 deposition. By contrast, the peak of H3K4me3 observed at promoters with long GC skew, while rising at a 5' boundary very similar to the previous subclass, extended further downstream. The correspondence between GC skew, DNA hypomethylation, and H3K4me3 deposition strengthens the hypothesis that GC skew, and thus co-transcriptional formation of R-loops, are highly predictive of an active, unmethylated epigenetic state at CGI promoters.

We computed the average length of the first exon of genes in each subclass (Figure 6F). Strikingly, the short promoter subclass had the shortest first exon with a median length of 178 bp. In contrast, promoters in the medium subclass had a median first exon length of 218 bp while the long subclass showed the longest median first exon length of 259 bp. This establishes that the length of the GC skew tract downstream of the TSS is also correlated with the length of the first exon. Taken together, these observations indicate that GC skew, and presumably R-loop formation, may be involved in setting a protective epigenetic landscape extending through the length of the first exon, into the first intron.

Given the strong association between GC skew and the protection against DNA methylation at promoters, we sought to determine if GC skew patterns at the 3' end of genes were also associated with any measure of local protection. For this, we graphed the percent methylation around the TTS of genes with 3' GC skew in comparison to those without 3' GC skew. Genes without 3' GC skew showed an increase in DNA methylation around the TTS (Figure 7). In contrast, genes with terminal GC skew showed a marked decrease in DNA

methylation around the TTS. The protection observed here, while not as significant as the one observed at promoter regions, reinforces the notion that GC skew and R-loop formation serve to shield a genomic locus from DNA methylation. Note that this reduction in DNA methylation, albeit reduced in amplitude, was still observed when tandem genes (TTS followed closely by a TSS) were filtered out of the gene set (Figure 7).

Discussion

We show here that GC skew enables the identification of four classes of promoters with distinct epigenetic profiles, genomic organizations, and gene ontologies. Class I promoters are strong CGIs that drive high transcriptional outputs and associate with clear housekeeping functions. These promoters are enriched on gene-poor chromosomes (Figure 3), suggesting that they are intrinsically self-sufficient in establishing and maintaining a DNA methylation-free state. R-loop profiling confirmed that these promoters are the main source of promoter R-loops (Figure 5), supporting the notion that GC skew and R-loop formation cooperate to enforce a protective barrier against DNA methylation (Ginno et al. 2012). The spread of GC skew downstream of the TSS correlates with the length of the first exon and the length of the CpG-dense, unmethylated, H3K4me3-marked region (Figure 6), suggesting that keeping the first exon free of DNA methylation is critical for transcriptional competence (Brenet et al. 2011). Likewise, these observations are consistent with a recent report indicating that the position of the first exon-intron boundary determines the length of the promoter-proximal H3K4me3 peak and that H3K4me3 levels are dependent upon splicing (Bieberstein et al. 2012). Altogether, this opens the possibility that R-loop formation may be involved in the recruitment of H3K4me3 and/or of the splicing machinery.

Class II promoters, in contrast to Class I, are ‘weak’ CGI promoters that tend to be enriched on gene-rich chromosomes. This suggests that Class II CGIs may not be entirely self-sufficient in promoting a DNA methylation-free state and may instead benefit from a shared protective environment established through neighboring genes (Figure 3C). The relative depletion of Class I and enrichment of Class II CGIs on the X chromosome is in accord with the idea that, while Class II islands are sufficiently protected to remain active on the active X, they can become efficiently DNA methylated and silenced on the inactive X owing to their intrinsically

weaker protection strength. Thus, the striking deviation observed for the distribution of Class I and II promoters on the X compared to autosomes may have been driven by constraints imposed through dosage compensation. It does not appear, however, that the ability of genes to escape X-inactivation correlates with the exact promoter type driving these genes (data not shown). This view is consistent with the fact that X-inactivation escape is likely to be mechanistically distinct from the epigenetic protection that operates at most CGI promoters (Berletch et al. 2011).

Class I and II CGI promoters are distinguished by the level of recruitment of the H4K20me1 and H3K79me3 histone marks (Figure 2). These two marks have been implicated both in transcriptional regulation and in aspects of DNA metabolism including DNA replication and DNA damage response (Nguyen and Zhang 2011; Beck et al. 2012). The dual nature of these marks is interesting in light of the fact that strong GC-skewed CGIs not only function as promoters but may also serve as DNA replication origins (Delgado et al. 1998; Sequeira-Mendes et al. 2009; Cayrou et al. 2011) and are associated with higher DNA recombination (Polak and Arndt 2009; Auton et al. 2012) and spontaneous mutation rates (Polak and Arndt 2008). R-loop formation likely underlies much of these varied effects (Aguilera and Garcia-Muse 2012). R-loops are well-suited to function as replication origins, as documented in Prokaryotes, bacteriophages, and mitochondria (Baker and Kornberg 1988; Carles-Kinch and Kreuzer 1997; Lee and Clayton 1998). R-loops also favor chromatin accessibility through a reduced affinity for histones (Dunn and Griffith 1980) and represent fragile regions that are more likely to break or undergo spontaneous deamination, particularly on the displaced strand (Beletskii and Bhagwat 1996). The preferential recruitment of H4K20me1 and H3K79me2 at R-loop-prone CGIs may therefore reflect the need to keep these loci under a broad DNA damage surveillance pathway.

We identify a third class of CGIs solely based on their unusual “reverse” GC skew (Class IV, Figure 1). These promoters represent an interesting gene set highly enriched for developmental regulators including numerous Polycomb group targets. In agreement, Class IV CGI promoters are enriched for the repressive H3K27me3 mark and for members of the PRC1 and PRC2 complexes (Figure 2). Interestingly, Class IV CGIs are still strongly protected from DNA methylation despite the relative absence of GC skew downstream of their TSS. This indicates that the protection observed at Class IV CGIs may be mechanistically distinct from Class I promoters. Multiple studies have shown that H3K27me3 and DNA methylation are mutually exclusive (Lindroth et al. 2008; Bartke et al. 2010) and that H3K27me3-marked regions are usually hypomethylated (Tanay et al. 2007). Therefore, recruitment of Polycomb complexes could be sufficient to protect these promoters. This suggestion is compatible with evidence that PRC targets are nonetheless more susceptible to DNA methylation both during development and in disease states such as cancer (Ohm et al. 2007; Mohn et al. 2008). Class IV CGIs are unlikely to benefit from an R-loop-mediated protection mechanism due to the absence of GC skew downstream of the TSS. Loss of Polycomb binding and/or H3K27me3 marking is therefore likely to render these loci particularly susceptible to the action of *de novo* DNA methyltransferases, thereby initiating an epigenetic switch to a long-term silent state. Our observation of a “reverse” GC skew upstream of the TSS of Class IV CGIs raises questions as to its biological significance. It is possible that PRC recruitment is favored by specific C-rich motifs upstream of the TSS on the non-template strand. The observation that the CXXC domain-containing, PRC1-associated, KDM2B protein is particularly enriched at a subset of PRC-targeted CGIs (Farcas et al. 2012; Wu et al. 2013) suggests that KDM2B, or other members of the PRC complexes, possess sequence specificity beyond CpG sites. It is also possible that the decreased propensity for R-loop formation downstream of the TSS might enable PRC recruitment. Alternatively, R-loop formation driven by antisense transcription

upstream of the TSS may target PRC complexes to mediate dynamic gene silencing. These possibilities remain to be investigated.

By increasing the sensitivity of our SkewR algorithm, we showed that GC skew occurs at the 3' ends of ~2000 human genes and that these genes tend to be located in gene-dense neighborhoods. This observation is significant given that R-loops were suggested to mediate efficient transcription termination (Skourti-Stathaki et al. 2011). Closely arranged genes might require an efficient termination mechanism to avoid transcriptional read-through. In agreement, we show here that closely arranged transcription units, most particularly those arranged in tandem orientation, are highly enriched for terminal GC skew. Our DRIP-seq data experimentally establishes that R-loop formation at the 3'-end of human genes is in fact observed at a thousand genes at least in Ntera2 cells and that R-loop formation in the vast majority of cases is driven by GC skew. This suggests that co-transcriptional R-loop formation may be broadly used to enable transcription termination in human cells. The exact mechanism by which R-loops mediate termination remains to be determined.

Altogether, our work establishes GC skew as an important DNA sequence feature that offers insights into the classification, function, and organization of human promoters. Our data reinforce the notion that R-loops may play a critical role in establishing a DNA methylation-free state at strong CGI promoters and indicates that sequence-driven DNA structures may represent a new layer of control for the deposition of epigenetic marks. At the same time, we uncovered a broad potential role for R-loop formation at the 3'-end of genes in mediating efficient transcription termination. While some epigenetic protection can also be detected at the 3'-end of terminally skewed genes (Figure 7), it is likely that epigenetic protection is not the main function of R-loop formation there. Thus, R-loops appear to function in two separate processes

depending on which gene end is being considered. Studying how R-loops are formed, sensed, and resolved is likely to provide new insights into important biological mechanisms in mammalian cells.

METHODS

Definition of gene sets for analysis

The human RefSeq gene set was filtered to remove any gene smaller than 2 kb. This resulted in 19,737 unique genes, 25.8% of which had at least 2 annotated promoters (24,836 promoters in total). These additional promoters were used in overlap analyses to allow for a more comprehensive sampling. The numbers of unique genes in each class are reported in Supplementary Table 2. The GC skew annotation of gene promoters and ends was performed by overlapping the output of SkewR under high stringency parameters (Ginno et al. 2012) with a -500 to +1500 bp window surrounding the TSS or TTS, respectively. Class I promoters are CGI promoters (as annotated by the UCSC Genome Browser) that overlapped with co-oriented GC skew peaks. Class II promoters are CGI promoters that did not overlap with any GC skew peak. Class III promoters fall outside of the CGI class and do not overlap with a SkewR peak. Class IV promoters correspond to CGI promoters that overlap with a “reverse” GC skew peaks (i.e. C-skew on the + strand or G-skew on the - strand). Annotated bidirectional promoters were filtered out of Class IV loci to avoid confounding effects. Additionally, Class IV promoters that overlapped with both co-oriented and reverse GC skew were recorded as Class I.

Meta-analysis of histone modification, DNA methylation, and gene expression profiles.

The average density profile for each epigenetic mark was determined over sets of loci (promoter or 3’-end) using appropriate coordinates files. Loci were all aligned and co-oriented at their TSS for promoters or TTS for gene ends. All datasets were for hESCs. Datasets for histone modifications and variants were obtained from ENCODE (Rosenbloom et al. 2010). DNA methylation datasets were from (Laurent et al. 2010). Human ES cell RING1B and EZH2 ChIP-seq datasets were from (Ku et al. 2008). The average signal for each mark was extracted from

the aligned wig files and a 50 or 100 bp smoothing window was applied. Each panel in Figure 2 represents the aggregate signal color-coded for each skew class. RNA-seq data from two independent studies of the hESC transcriptome were used (Lister et al. 2009; Rada-Iglesias et al. 2011). RPKM values for each gene were extracted and used in calculating average expression levels for each class. P-values were calculated by first determining that the spread of data was normal via the D'Agostini K-squared test, and then ANOVA was applied to determine the significance between the means.

New SkewR Annotations

The stringency of the SkewR algorithm was decreased by modifying the GC-rich state of the HMM. In this case, 7,500 copies of a set of GC-rich (non-skewed) DNA sequences were added to train this state as opposed to 1 million copies in the high stringency version. This increased the number of GC skew peaks from 19,864 to 66,282. All other parameters for SkewR were as described (Ginno et al. 2012).

Genome-wide R-loop profiling using DRIP-seq

DRIP-seq was performed on genomic DNA from human pluripotent Ntera2 cells as previously described (Ginno et al. 2012) except that the DNA was either fragmented using HindIII, EcoRI, BsrGI, XbaI and SspI (DRIP 1) or BamHI, NcoI, ApaLI, NheI and PvuII (DRIP 2, two technical replicates). Input DNA was also fragmented with each restriction enzyme cocktail and sequenced on the Illumina HiSeq platform. Sequencing reads were mapped to the hg19 genome using BWA 0.6.1 (Li and Durbin 2009), resulting in 20.4 and 53.5 million mapped reads for each sample. Peak calling was first performed by MACS 1.4.2 (Zhang et al. 2008) using the matching input library as control. DRIP peaks were further assigned onto restriction fragments

using custom Java and Perl scripts. Regions common to DRIP 1 and DRIP 2 were considered consensus DRIP-seq peaks. Overlap of DRIP peaks with TSS and TTS was performed using BEDTools (Quinlan and Hall 2010).

Clustering of Class I promoters according to GC skew length

8,332 Class I promoters were clustered into 3 subclasses based on GC skew peak length downstream of TSS using a custom Perl script. The length cutoffs for the GC skew peaks were less than 300 bp (1,288 promoters; short), between 300 and 700 bp (4,674 promoters; medium), and more than 700 bp (2,160 promoters; long), respectively. The total number of promoters is slightly less than 8,332 because some of the promoters in Class I have GC skew upstream of TSS and were not included. GC skew, CpG obs/exp, GC content, DNA methylation and H3K4me3 profiles were plotted for each subclass as described earlier. The lengths of first exons were extracted from corresponding RefSeq entries.

Data access

DRIP-seq datasets are available in the Gene Expression Omnibus (GEO) database, under the accession number GSE45530.

Figure Legends

Figure 1. GC skew distinguishes four promoter classes in the human genome. (A-D) Metaplots of GC skew (red line), GC percent (green line), and CpG observed over expected ratio (o/e; blue line) were determined for each class of promoters over a 5 kb window centered around the TSS. (E) Expression levels (RPKM) for each GC skew promoter class, as determined in H1 hESCs. (F) Top gene ontology hits for Class IV genes. The x-axis represents the p-value of enrichment after Bonferroni correction.

Figure 2. Promoter classes present distinct epigenetic signatures in hESCs. (A) DNA methylation metaplots for each of the four promoter classes over a 10 kb window centered on the TSS. The numbers of promoters in each class were: Class I = 8,332, Class 2 = 5,799, Class 3 = 7,968, Class 4 = 2,099. (B-D) Histone modification metaplots for each promoter class for H4K20me1 (B), H3K79me2 (C) and H3K27me3 (D). (E-F) Average binding profiles for each promoter class for EZH2 (E) and RING1B (F). Class-specific color codes are all identical and indicated in panel B. The y-axes in panels B-F represent arbitrary units.

Figure 3. Gene density strongly affects the distribution of Class I and Class II genes and the X chromosome represents an exception to the autosomal trends. (A-B) The distribution of Class I and Class II genes on individual chromosomes is represented as a percentage of total RefSeq genes on that chromosome (y-axis) plotted against a measure of gene density (x-axis; CGI/Mb – a set of 10,279 high confidence promoter CGIs (Bock et al. 2007) was used). The X chromosome is shown in blue; autosomes are in red – a few relevant chromosomes are indicated. The data was fit to a linear regression shown here with the corresponding 95% confidence interval. (C) Schematic representation of the manner by which a gene-rich region

may enable a shared epigenetic state (arrows) between neighboring genes while a gene-poor may not. CGI promoters are shown by green boxes; peaks of G-skew or C-skew are shown by red and blue boxes, respectively. (D) The distribution of Class I (left) and Class II (right) genes is represented as a percentage of total RefSeq genes calculated over each X-chromosome evolutionary strata (PAR1 (0-2.8 Mb), XAR (2.8-46.8 Mb), S2a (46.8-60 Mb), XCR (60-148.6 Mb) and S2b (148.6-154.8 Mb)). The expected percentage of Class I and Class II genes based on their autosomal distributions is shown by a straight line together with standard deviation (dotted lines). The X-inactivation efficiency across each strata is color-coded and was determined from Carrel and Willard (2005).

Figure 4. Terminal GC skew is a novel feature of a subset of human genes that correlates with high gene density. (A) GC skew metaplot for genes with co-oriented terminal GC skew. The window is centered on the 3' end of each gene (as defined by RefSeq annotation) and calculated using a 100 bp sliding window. The box whisker plot represents the distribution of GC skew peak starts. (B) Chromosomal distribution of genes with 3' GC skew. Symbols are as in Figure 4. (C) Schematic representation of the arrangement of genes with terminal GC skew relative to their closest neighbor (focusing on neighbors located less than 2 kb away).

Figure 5. DRIP-seq illustrates R-loop formation at the 5' and 3' ends of human genes. (A to D) DRIP-seq profiles. The SkewR track shows regions of GC skew with red indicating G-rich blocks and blue C-rich blocks. DRIP 1 and DRIP 2 correspond to DRIP-seq experiments for which the genome was fragmented with two distinct cocktails of restriction enzymes (cut sites are indicated below each DRIP dataset). The DRIP peak track indicates consensus DRIP signal. Figures (A) and (B) show an R-loop at the TSS and TTS of a gene, respectively. (C) R-loop forms at the TTSs of two convergent genes. (D) The *PODXL2* gene shows both TSS and

TTS R-loops. Note that the TTS is followed closely by the TSS of the neighboring *ABTB1* gene. (E) Distribution of DRIP-seq peaks over TSS and TTS classes.

Figure 6. Clustering of Class I promoters reveals new correlations between the genetic and epigenetic landscapes of CGI promoters. (A) Average GC skew profiles for the three main Class I promoter clusters. Each panel represents relevant genetic and epigenetic profiles for each cluster, including: (B) CpG density; (C) GC content ; (D) DNA methylation profiles; (E) H3K4me3 profiles; and (F) first exon length (represented in a boxplot format). Color codes are as indicated.

Figure 7. Terminal GC skew also confers a measure of protection against DNA methylation. The graph represents the average DNA methylation profiles of genes with and without terminal GC skew. All genes were aligned at their TTS and DNA methylation in hESCs was from Laurent *et al.* (2011). Genes whose TSS was located within 2 kb or closer to the nearest downstream promoter were also filtered out (tandem filter) to remove any confounding effects due to the presence of a nearby protected promoter.

Acknowledgements

This work was supported by research grants from the National Institutes of Health (NIH 1R01GM094299 to F.C.). P.A.G. and Y. W. L. were supported in part by a predoctoral NIH Training Grant (5T32GM007377).

REFERENCES

Aguilera A, Garcia-Muse T. 2012. R loops: from transcription byproducts to threats to genome stability. *Molecular cell* **46**(2): 115-124.

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular cell* **31**(6): 785-799.

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Segurel L, Street T, Leffler EM, Bowden R, Aneas I, Broxholme J et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**(6078): 193-198.

Baker TA, Kornberg A. 1988. Transcriptional activation of initiation of replication from the E. coli chromosomal origin: an RNA-DNA hybrid near oriC. *Cell* **55**(1): 113-123.

Bartke T, Vermeulen M, Xhemalce B, Robson SC, Mann M, Kouzarides T. 2010. Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143**(3): 470-484.

Beck DB, Oda H, Shen SS, Reinberg D. 2012. PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. *Genes Dev* **26**(4): 325-337.

Beletskii A, Bhagwat AS. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. *Proc Natl Acad Sci U S A* **93**(24): 13919-13924.

Berletch JB, Yang F, Xu J, Carrel L, Disteche CM. 2011. Genes that escape from X inactivation. *Hum Genet* **130**(2): 237-245.

Bieberstein NI, Carrillo Oesterreich F, Straube K, Neugebauer KM. 2012. First exon length controls active chromatin signatures and transcription. *Cell Rep* **2**(1): 62-68.

Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS genetics* **2**(3): e26.

Bock C, Walter J, Paulsen M, Lengauer T. 2007. CpG island mapping by epigenome prediction. *PLoS Comput Biol* **3**(6): 110.

Brenet F, Moh M, Funk P, Feierstein E, Viale AJ, Socci ND, Scandura JM. 2011. DNA methylation of the first exon is tightly linked to transcriptional silencing. *PloS one* **6**(1): e14524.

Brenner C, Fuks F. 2006. DNA methyltransferases: facts, clues, mysteries. *Curr Top Microbiol Immunol* **301**: 45-66.

C. David Allis TJ, Danny Reinberg. 2007. *Epigenetics*. John Inglis, Cold Spring Harbor.

Carles-Kinch K, Kreuzer KN. 1997. RNA-DNA hybrid formation at a bacteriophage T4 replication origin. *J Mol Biol* **266**(5): 915-926.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**(7031): 400-404.

Cayrou C, Coulombe P, Vigneron A, Stanojcic S, Ganier O, Peiffer I, Rivals E, Puy A, Laurent-Chabalier S, Desprat R et al. 2011. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res* **21**(9): 1438-1449.

Delgado S, Gomez M, Bird A, Antequera F. 1998. Initiation of DNA replication at CpG islands in mammalian chromosomes. *Embo J* **17**(8): 2426-2435.

Dunn K, Griffith JD. 1980. The presence of RNA in a double helix inhibits its interaction with histone protein. *Nucleic acids research* **8**(3): 555-566.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**(7345): 43-49.

Farcas AM, Blackledge NP, Sudbery I, Long HK, McGouran JF, Rose NR, Lee S, Sims D, Cerase A, Sheahan TW et al. 2012. KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. *Elife* **1**: e00205.

Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. 2003. Predicting aberrant CpG island methylation. *Proc Natl Acad Sci U S A* **100**(21): 12253-12258.

Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. 2006. DNA motifs associated with aberrant CpG island methylation. *Genomics* **87**(5): 572-579.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ et al. 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**(2): 212-224.

Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. 2012. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Molecular cell* **45**(6): 814-825.

Hervouet E, Vallette FM, Cartron PF. 2009. Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. *Epigenetics* **4**(7): 487-499.

Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, Park J, Butler J, Rafii S, McCombie WR et al. 2011. Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Molecular cell* **44**(1): 17-28.

Illingworth RS, Bird AP. 2009. CpG islands--'a rough guide'. *FEBS Lett* **583**(11): 1713-1720.

Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. 2010. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**(7): 2926-2931.

Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS et al. 2008. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS genetics* **4**(10): e1000242.

Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J et al. 2010. Dynamic changes in the human methylome during differentiation. *Genome Res* **20**(3): 320-331.

Lee DY, Clayton DA. 1998. Initiation of mitochondrial DNA replication by transcription and R-loop processing. *J Biol Chem* **273**(46): 30614-30621.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schubeler D. 2011. Identification of genetic elements that autonomously determine DNA methylation states. *Nature genetics* **43**(11): 1091-1097.

Lin IG, Tomzynski TJ, Ou Q, Hsieh CL. 2000. Modulation of DNA binding protein affinity directly affects target site demethylation. *Molecular and cellular biology* **20**(7): 2343-2349.

Lindroth AM, Park YJ, McLean CM, Dokshin GA, Persson JM, Herman H, Pasini D, Miro X, Donohoe ME, Lee JT et al. 2008. Antagonism between DNA and H3K27 methylation at the imprinted Rasgrf1 locus. *PLoS genetics* **4**(8): e1000145.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**(7271): 315-322.

Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schubeler D. 2008. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular cell* **30**(6): 755-766.

Nguyen AT, Zhang Y. 2011. The diverse functions of Dot1 and H3K79 methylation. *Genes Dev* **25**(13): 1345-1358.

Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu W et al. 2007. A stem cell-like chromatin pattern may predispose tumor

suppressor genes to DNA hypermethylation and heritable silencing. *Nature genetics* **39**(2): 237-242.
Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD et al. 2007. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**(7154): 714-717.
Polak P, Arndt PF. 2008. Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res* **18**(8): 1216-1223.
Polak P, Arndt PF. 2009. Long-range bidirectional strand asymmetries originate at CpG islands in the human genome. *Genome Biol Evol* **1**: 189-197.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.
Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**(7333): 279-283.
Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ, Wang T, Hinrichs AS, Zweig AS et al. 2010. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic acids research* **38**(Database issue): D620-625.
Roudier F, Ahmed I, Berard C, Sarazin A, Mary-Huard T, Cortijo S, Bouyer D, Caillieux E, Duvernois-Berthet E, Al-Shikhley L et al. 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *EMBO J* **30**(10): 1928-1938.
Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gomez M. 2009. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS genetics* **5**(4): e1000446.
Skourti-Stathaki K, Proudfoot NJ, Gromak N. 2011. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Molecular cell* **42**(6): 794-805.
Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, Simon I, Yakhini Z, Cedar H. 2009. Developmental programming of CpG island methylation profiles in the human genome. *Nat Struct Mol Biol* **16**(5): 564-571.
Tanay A, O'Donnell AH, Damelin M, Bestor TH. 2007. Hyperconserved CpG domains underlie Polycomb-binding sites. *Proc Natl Acad Sci U S A* **104**(13): 5521-5526.
Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD et al. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**(7291): 1082-1086.
Vavouri T, Lehner B. 2012. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Genome Biol* **13**(11): R110.
Wassenegger M, Heimes S, Riedel L, Sanger HL. 1994. RNA-directed de novo methylation of genomic sequences in plants. *Cell* **76**(3): 567-576.
Wienholz BL, Kareta MS, Moarefi AH, Gordon CA, Ginno PA, Chedin F. 2010. DNMT3L modulates significant and distinct flanking sequence preference for DNA methylation by DNMT3A and DNMT3B in vivo. *PLoS genetics* **6**(9).
Wu X, Johansen JV, Helin K. 2013. Fbxl10/Kdm2b Recruits Polycomb Repressive Complex 1 to CpG Islands and Regulates H2A Ubiquitylation. *Molecular cell*.
Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**(4): 816-831.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.

Figure1

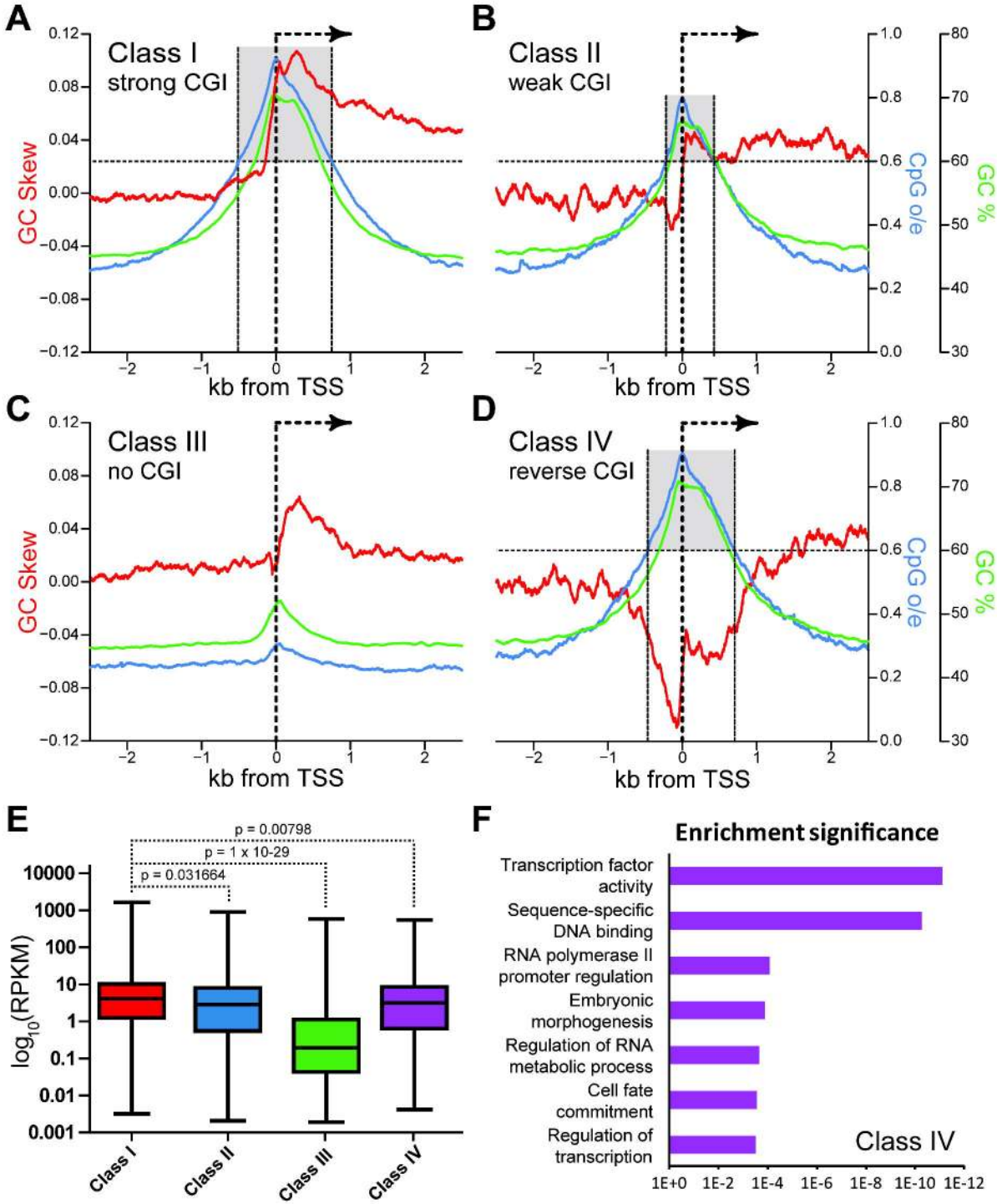

A
Class I
strong CGI
GC Skew
kb from TSS
B
Class II
weak CGI
CpG o/e
GC %
kb from TSS
C
Class III
no CGI
GC Skew
kb from TSS
D
Class IV
reverse CGI
CpG o/e
GC %
kb from TSS
E
p = 0.00798
p = 1 x 10-29
p = 0.031664
log10(RPKM)
Class I
Class II
Class III
Class IV
F
Enrichment significance
Transcription factor activity
Sequence-specific DNA binding
RNA polymerase II promoter regulation
Embryonic morphogenesis
Regulation of RNA metabolic process
Cell fate commitment
Regulation of transcription
Class IV
1E+0
1E-2
1E-4
1E-6
1E-8
1E-10
1E-12

Figure 2

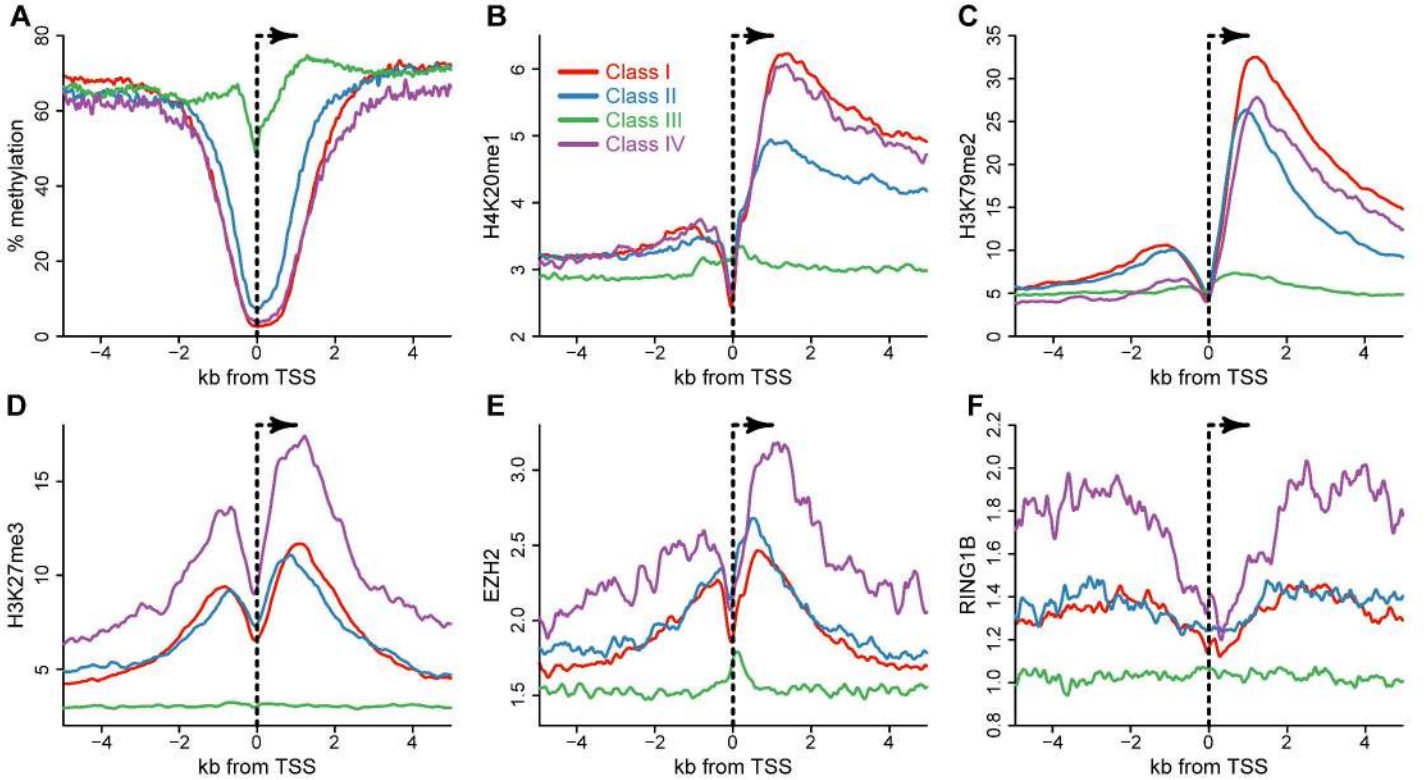

A
% methylation
kb from TSS
B
H4K20me1
Class I
Class II
Class III
Class IV
kb from TSS
C
H3K79me2
kb from TSS
D
H3K27me3
kb from TSS
E
EZH2
kb from TSS
F
RING1B
kb from TSS

Figure 3

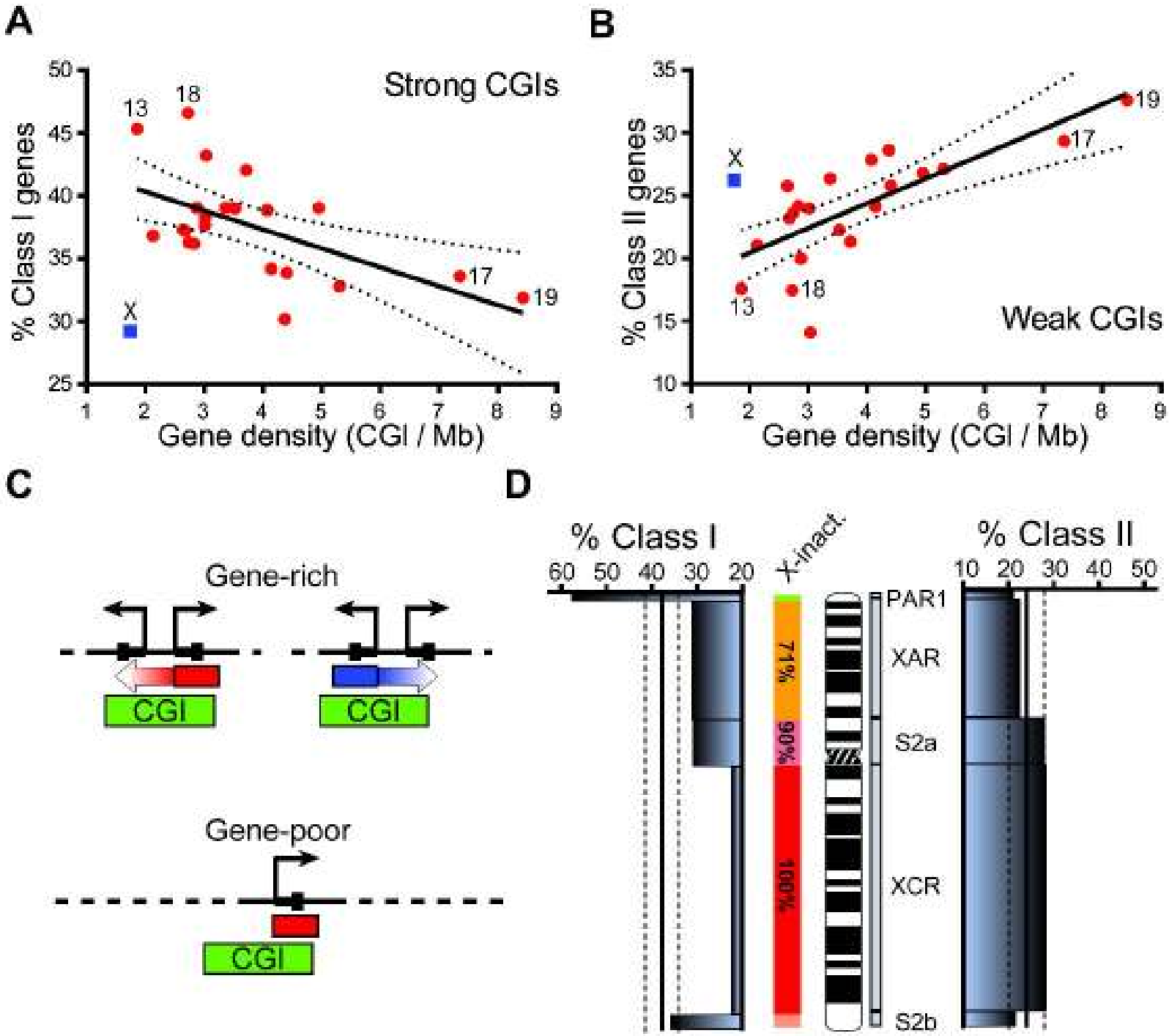
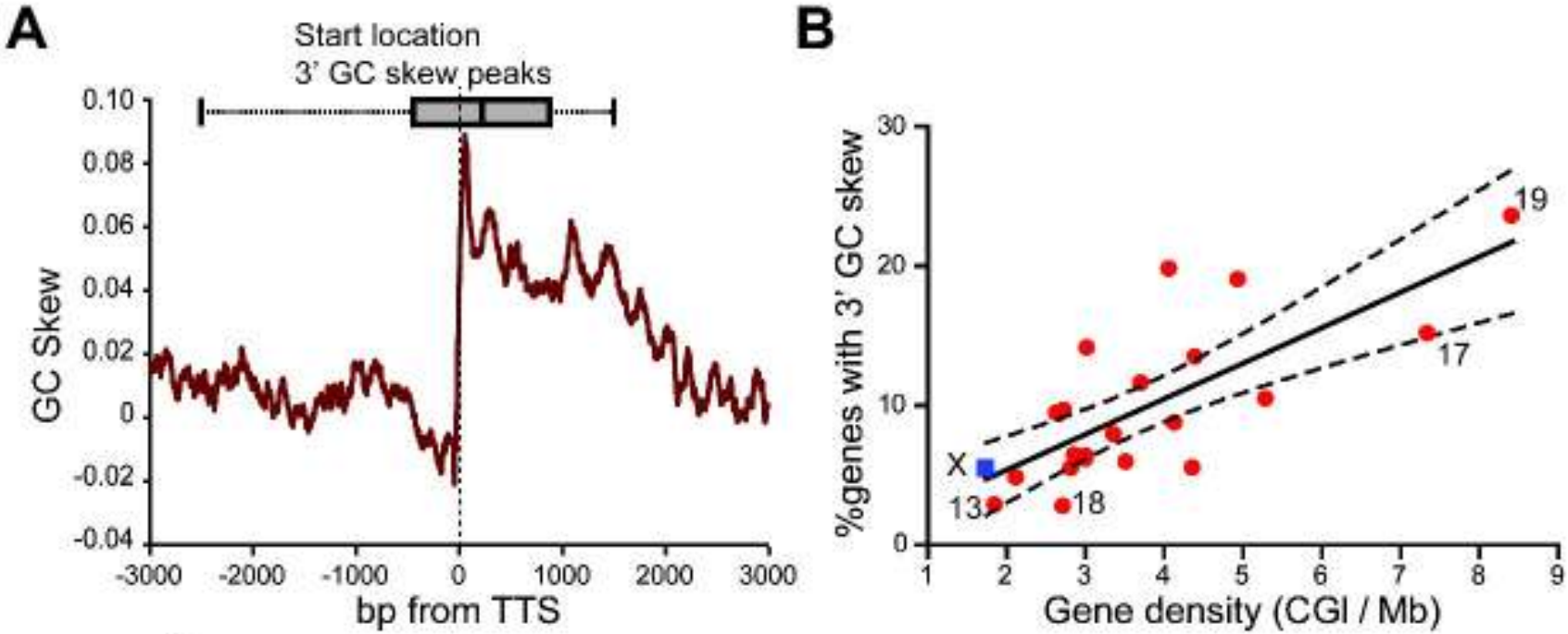

A
Strong CGIs
% Class I genes
Gene density (CGI / Mb)
13
18
17
19
X
B
Weak CGIs
% Class II genes
Gene density (CGI / Mb)
C
Gene-rich
CGI
Gene-poor
D
% Class I
% Class II
X-inact.
71%
90%
100%
PAR1
XAR
S2a
XCR
S2b

Figure 4



3' GC skew statistics	TTS **with** GC skew	Percent of Total	TTS **w/o** GC skew	Percent of Total
Total	2044	100%	19146	100%
Tandem (< 2 kb)	529	25.88%	387	2.02%
Converging (< 2 kb)	303	14.82%	1303	6.81%
3' and 5' GC skew	1388	68%	na	na

Figure 5

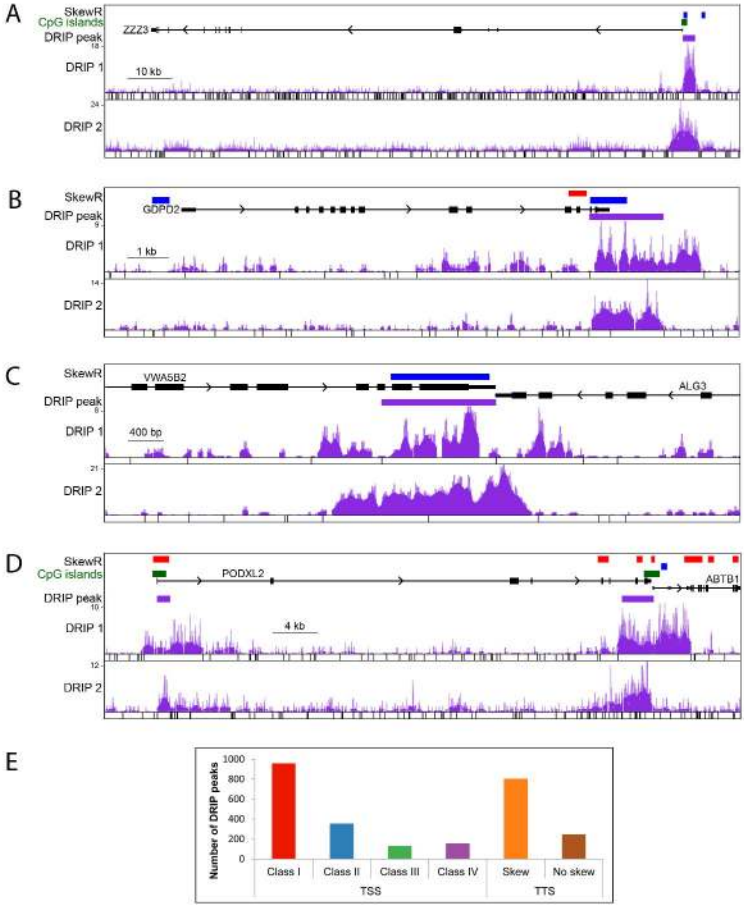

A
SkewR
CpG islands
DRIP peak
ZZZ3
DRIP 1
10 kb
DRIP 2
B
SkewR
DRIP peak
GDPD2
DRIP 1
1 kb
DRIP 2
C
SkewR
VWA5B2
ALG3
DRIP peak
DRIP 1
400 bp
DRIP 2
D
SkewR
CpG islands
PODXL2
ABTB1
DRIP peak
DRIP 1
4 kb
DRIP 2
E
Number of DRIP peaks
1000
800
600
400
200
0
Class I
Class II
Class III
Class IV
Skew
No skew
TSS
TTS

Figure 6

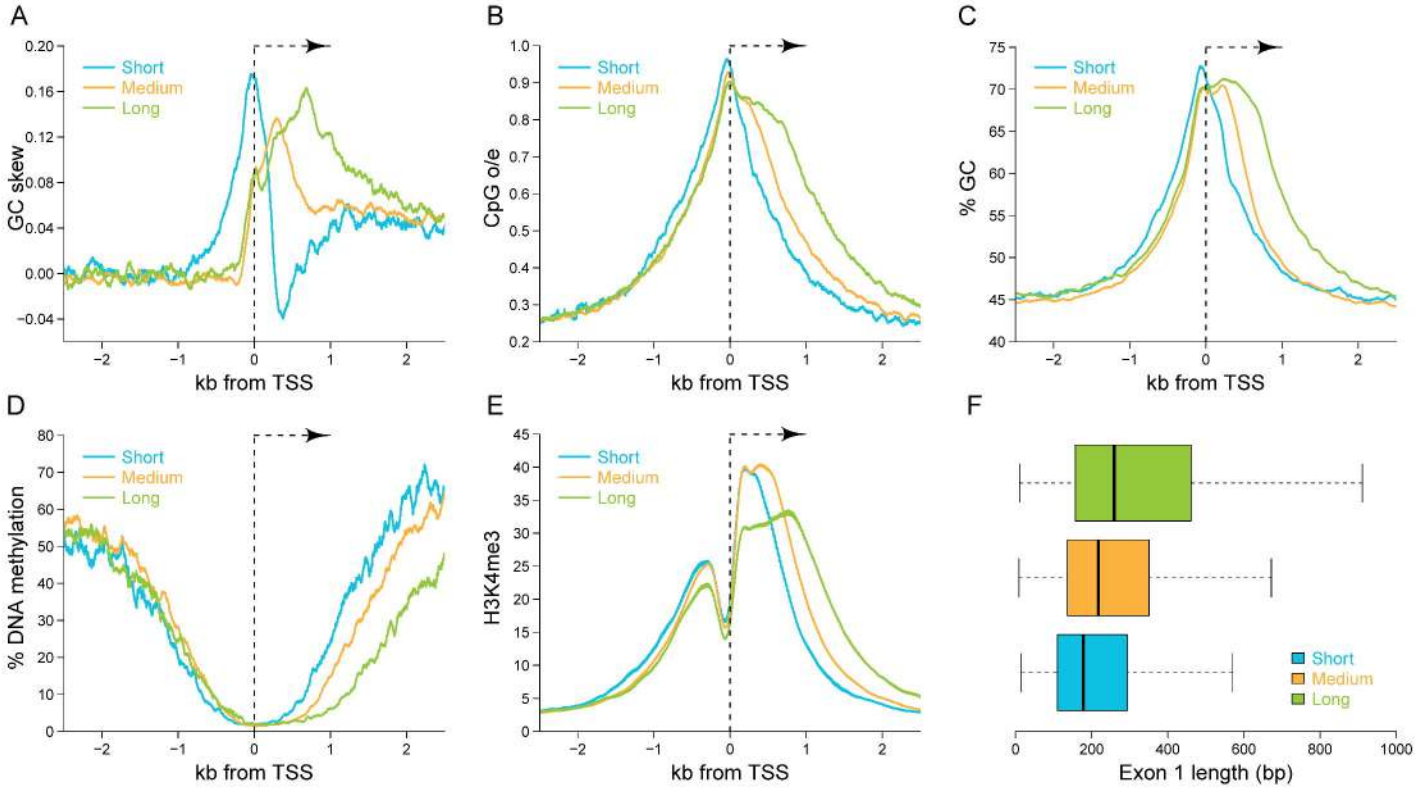
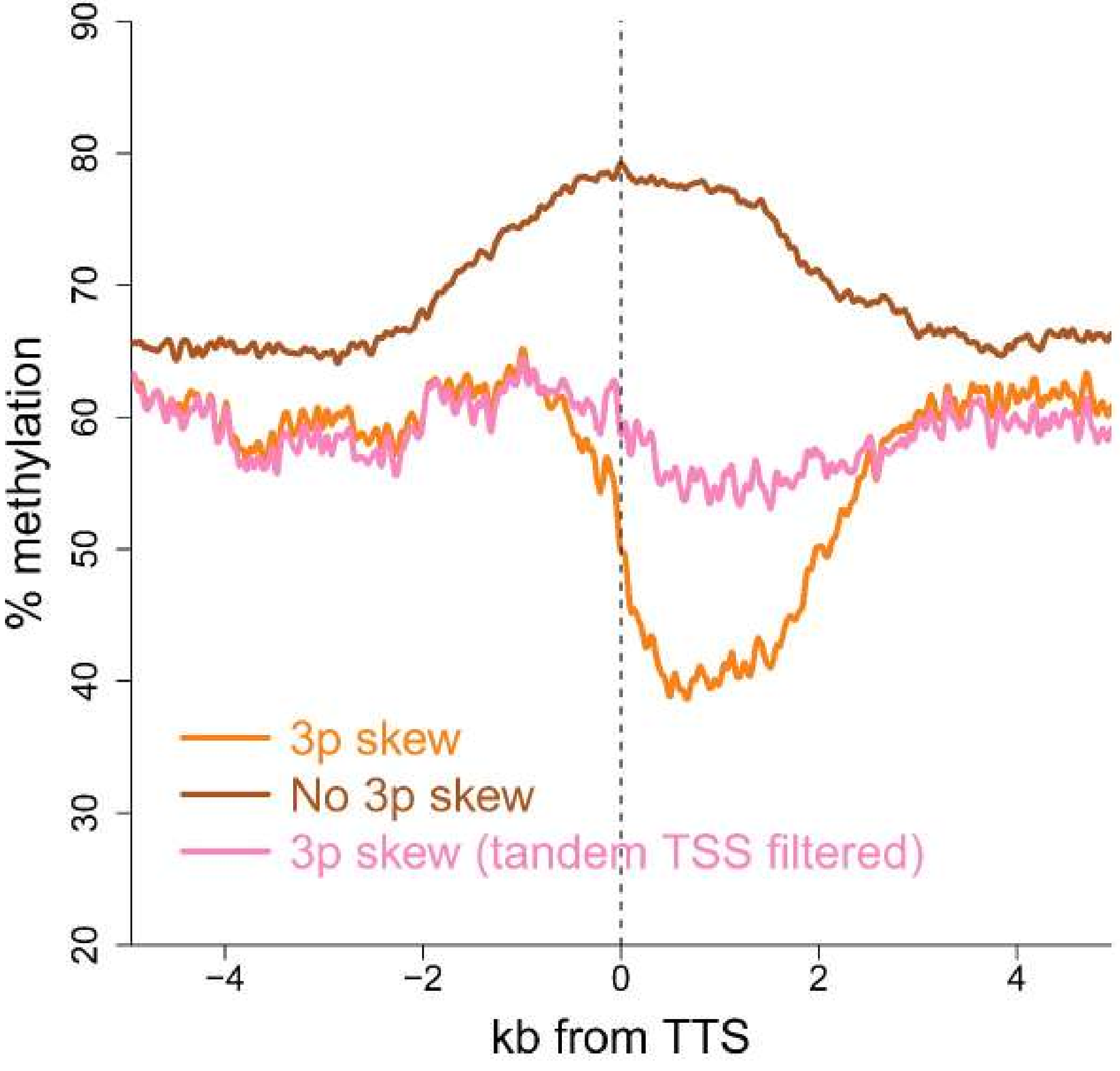

A
GC skew
Short
Medium
Long
kb from TSS
B
CpG o/e
Short
Medium
Long
kb from TSS
C
% GC
Short
Medium
Long
kb from TSS
D
% DNA methylation
Short
Medium
Long
kb from TSS
E
H3K4me3
Short
Medium
Long
kb from TSS
F
Short
Medium
Long
Exon 1 length (bp)

Figure 7


% methylation
90
80
70
60
50
40
30
20
−4
−2
0
2
4
kb from TTS
3p skew
No 3p skew
3p skew (tandem TSS filtered)