Supplemental Data and Methods

Supplemental Methods:

GSEA/ GO Analysis

Gene sets from each pausing class were analyzed for overlap with curated data sets (C2, C4, C6, C7, H, MF) in MSigDB using the web interface available at http://www.broadinstitute.org/gsea/msigdb/ (Subramanian a. et al. 2005) and for functional annotation using the DAVID Bioinformatics Resource (http://david.abcc.ncifcrf.gov) (Dennis et al. 2003). CGI-associated TSSs in each class were annotated to the corresponding gene. Input gene lists (symbols) were created from genes ranked by pausing index (eg. Top 20% by PI) within each class or by the ratio of Proximal to Distal (or Distal to Proximal) pausing indexes (eg. Top 1000 ratio Prox:Dist, etc). For the silent class, all genes for which the average reads/kb in the gene body >0 was used as input. These data can be found in **Supplemental Files** MSigDB-GSEA.xls and DAVID_GO.xls, respectively.

Supplemental References:

Dennis G, Sherman, BT, Hosack, DA, Yang J, Baseler BW, Lane HC, Lempicki, RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* **4:** P3.

Ginno PA, Lim YW, Lott PL, Korf I, Chedin F. 2013. GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res* **23**: 1590-1600.

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

Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**: 950-953.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A.* **102**:15545-50.

Supplemental Figures:

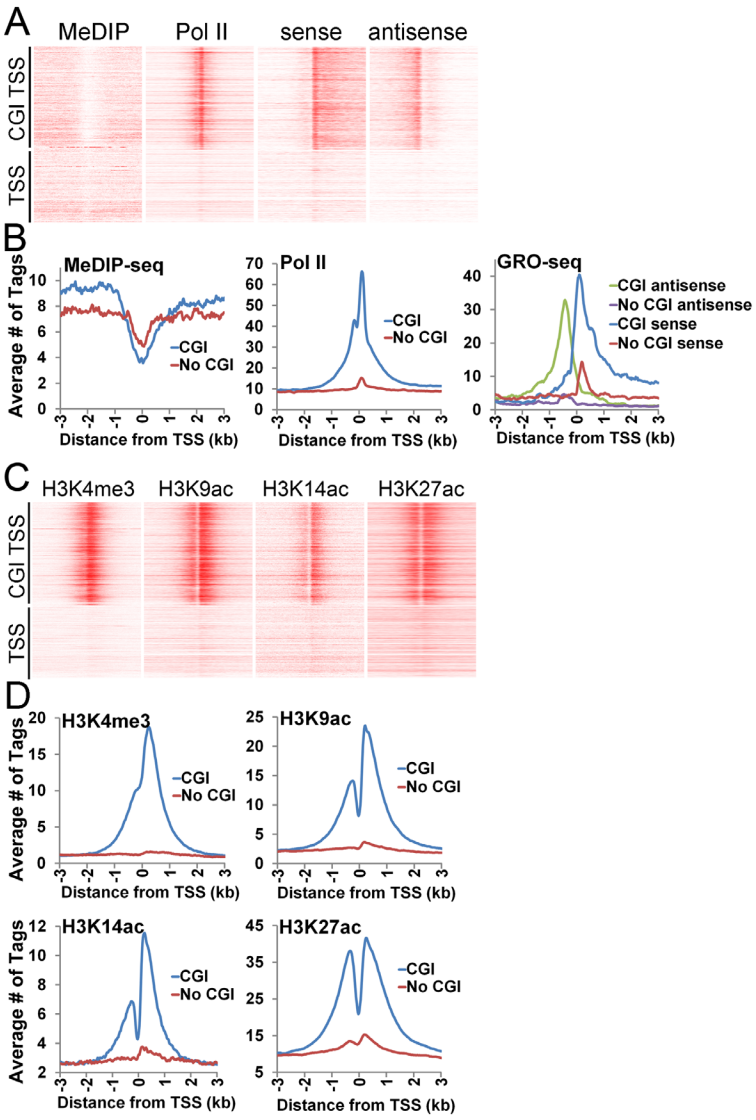

Figure S1. Promoter associated CGI are associated with higher levels of divergent transcription.

(A,C) Heat map representation of the tag densities (20bp bins) for DNA methylation (MeDIP-seq, white to red, 0-10), initiated Pol II (phospho–S5- Pol II ChIP-seq, white to red, 0-60), nascent transcription (GRO-seq; sense, antisense, white to red, 0-35); and (C) histone modifications: H3K4me3 (white to red, 0-20), H3K9ac (white to red, 0-25), H3K14ac (white to

red, 0-10), and H3K27ac (white to red, 0-40) for a 6 kb region surrounding the TSS (+/- 3kb) from MCF7 cells. TSSs were defined as CGI-associated if they were encompassed by an annotated CGI (CGI TSS; n=16,657) or not (No CGI TSS; n= 11,878). (B, D) Average tag densities of +/- 3 kb around for TSSs parsed by absence or presence of CGI from data in (A) and (C) respectively.

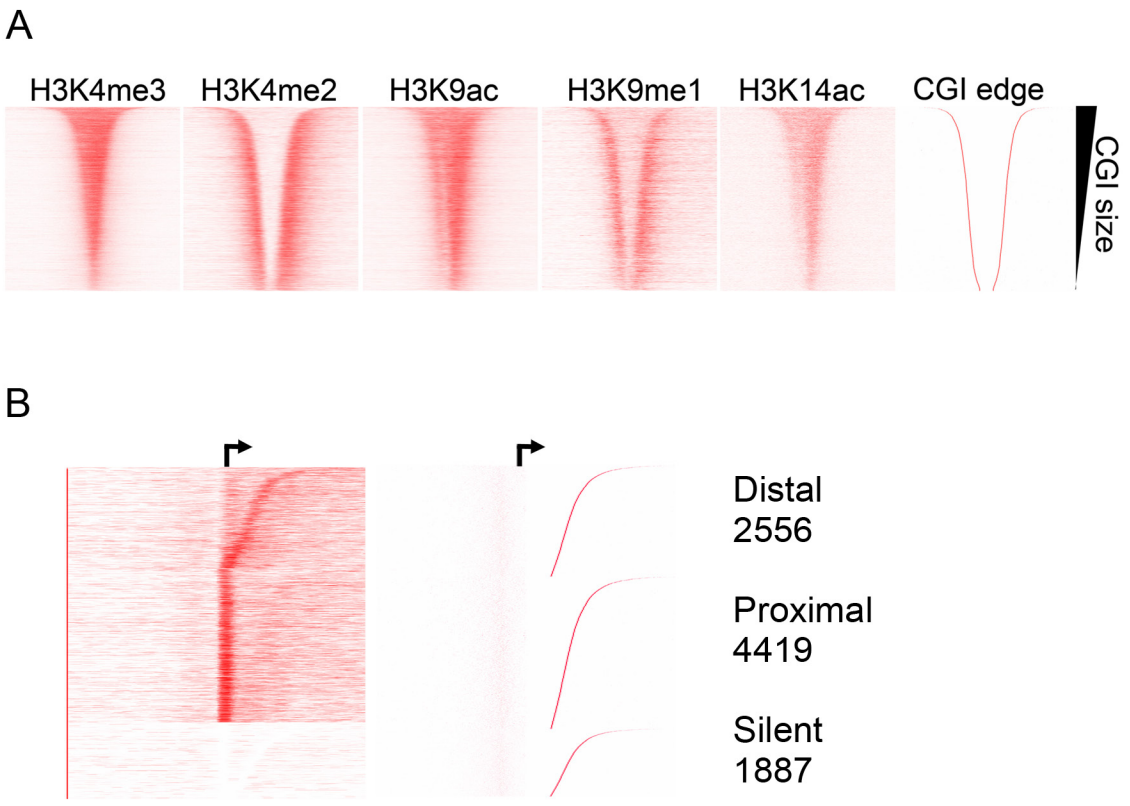

Figure S2. Promoter associated CGI maintain a transcriptionally permissive chromatin domain that is tightly constrained.

Heat map representation of ChIP-seq tag density (20 bp bins) of H3K4me3 (white to red, 0-20), H3K4me2 (white to red, 0-50), H3K9ac (white to red, 0-25), H3K9me1 (white to red, 0-25), and H3K14ac (white to red, 0-10) from MCF7 cells. CGI associated promoters (CGI TSS; n=16,657) were oriented to transcription and sorted by by descending CGI size. Plotted is the density around the midpoint of the CGI +/- 3kb. (B) Heatmap representation of MCF7 GRO-seq sense strand tag density (20 bp bins) across CGI containing a single annotated TSS (white to red, 0-35). CGI promoters are oriented to transcription and sorted within each pausing class by the distance from the TSS (arrow) to the 3' CGI edge using the same sort order described in Figure 1C. Exclusion of multi-TSS CGI does not influence the pausing patterns or class distinctions.

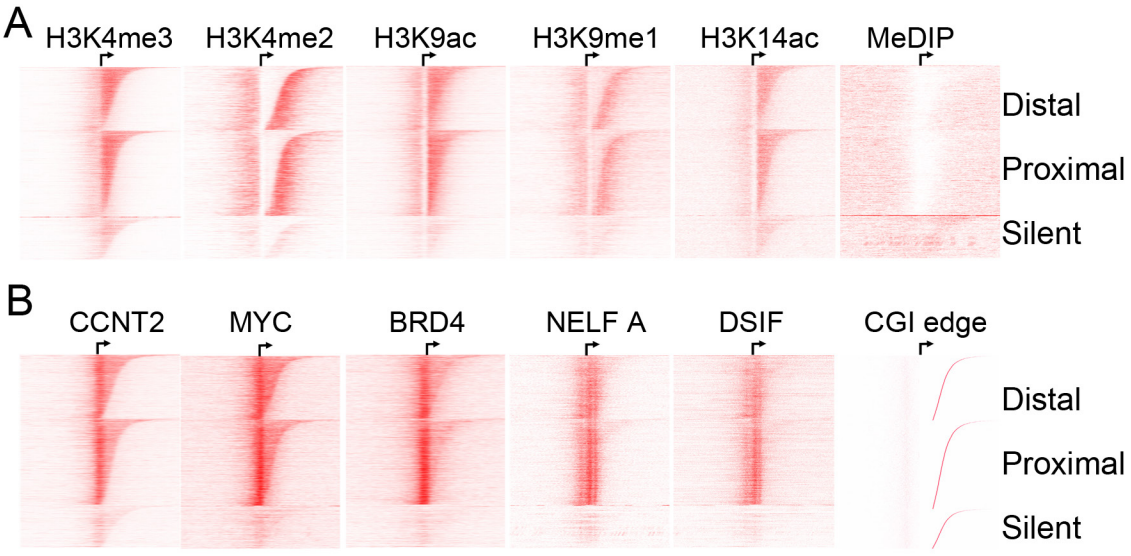

Figure S3. Neither promoter-associated histone modifications nor known pausing factors correlate with Pol II pausing at the distal edge.

(A,B) Heatmap representations of the density (20 bp bins) of (A) histone modifications and DNA methylation and (B) pausing factors for CGI promoters in the three pausing classes. CGI promoters are oriented to transcription and sorted within each class by the distance from the TSS (arrow) to the 3' CGI edge using the same sort order described in Figure 1C. (A) H3K4me3 (white to red, 0-20), H3K4me2 (white to red, 0-50), H3K9ac (white to red, 0-25), H3K9me1 (white to red, 0-25), H3K14ac (white to red, 0-10) ChIP-seq data and DNA methylation (MeDIP-seq) (white to red, 0-10) data are from MCF7 cells. (B) MYC (white to red, 0-100) and BRD4 (white to red, 0-25) from MCF7 cells; NELFA (white to red, 0-3) and the SUPT5H component of the DSIF complex (white to red, 0-10) from HeLa cells; and CCNT2 (white to red, 0-90) from K562 cells.

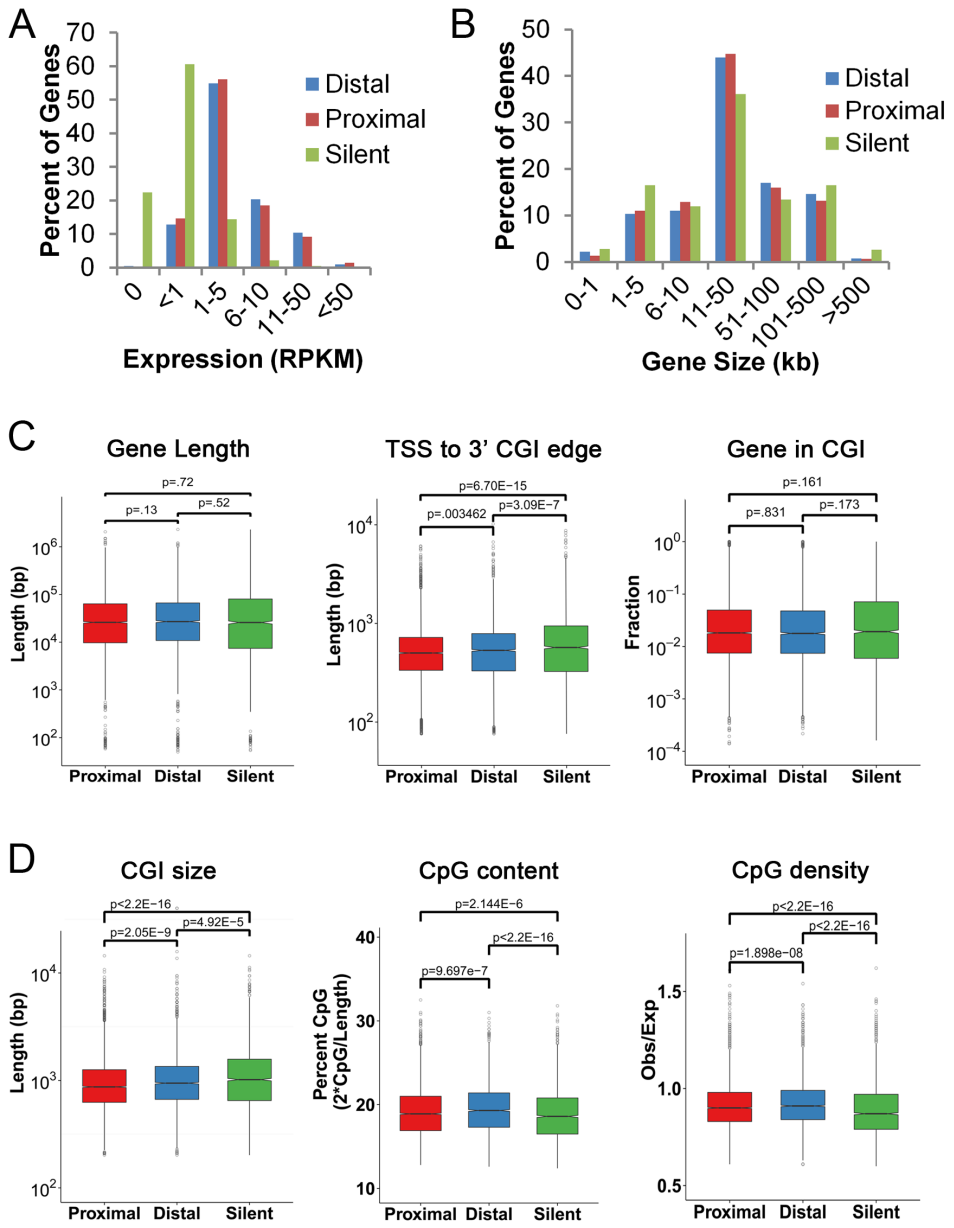

Figure S4. Distally and proximally paused genes do not differ in gene expression levels nor gene characteristics, but show modest differences in CGI features.

(A) Gene expression was measured as RPKM of gene body GRO-seq sense strand tags from MCF7 cells. Shown is the percent of genes in each class falling into the indicated expression levels. (B) Distribution of gene size (distance from the TSS to the TES) is plotted for genes in

each of the three Pol II pausing classes. (C) Box plots of CGI size, the distance from the TSS to the 3' CGI edge and the percent of gene covered by the CGI of genes in each pausing class. Median is indicated by a line, box is the 1st and 3rd quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. Significance (p-value) was assessed by the Mann-Whitney *U* test. (D) Box plots of CpG content and CpG density among genes in the three pausing classes. Median is indicated by a line, box is the 1st and 3rd quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. Significance (p-value) was assessed by the Mann-Whitney *U* test.

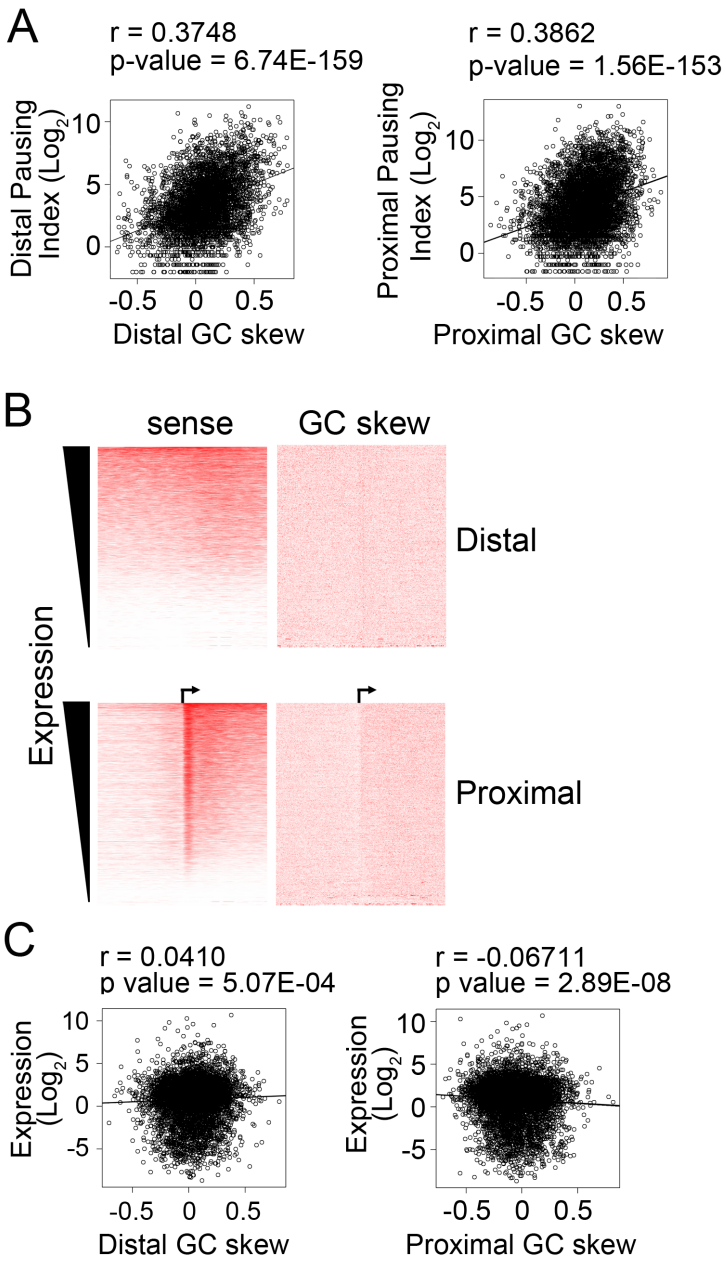

Figure S5. Pol II pausing correlates better with GC skew than with expression levels.

(A) Relationship between GC skew and pausing index at promoter-associated CGI. GC skew was calculated for the 100 bp underlying the distal pause (+20 to +120 bp from the 3' CGI edge, left panel) or the proximal pause (+40 to +140 bp from TSS, right panel) and plotted against log2

of either the distal (left panel) or proximal (right panel) pausing index. Pearsons correlation coefficients are provided. (B) Heat map representation of GRO-seq sense strand tag density (left, white to red, 0-35 tags/20bp) and GC skew (right, white to red, 0-1)) for 6kb centered on the 3' edge of the CGI (top) or the TSS (arrow, bottom). Promoter associated CGI (n= 16,657) were sorted by expression levels as measured by RPKM of GRO-seq sense strand tags across the gene body (3' edge of CGI to TES). (C) Relationship between gene expression and GC skew at the distal (left) or proximal (right) pause. GC skew at the distal (left) or proximal (right) pause was determined for all promoter associated CGI as in A and plotted against log2 of gene expression as measured by the RPKM across the gene body.

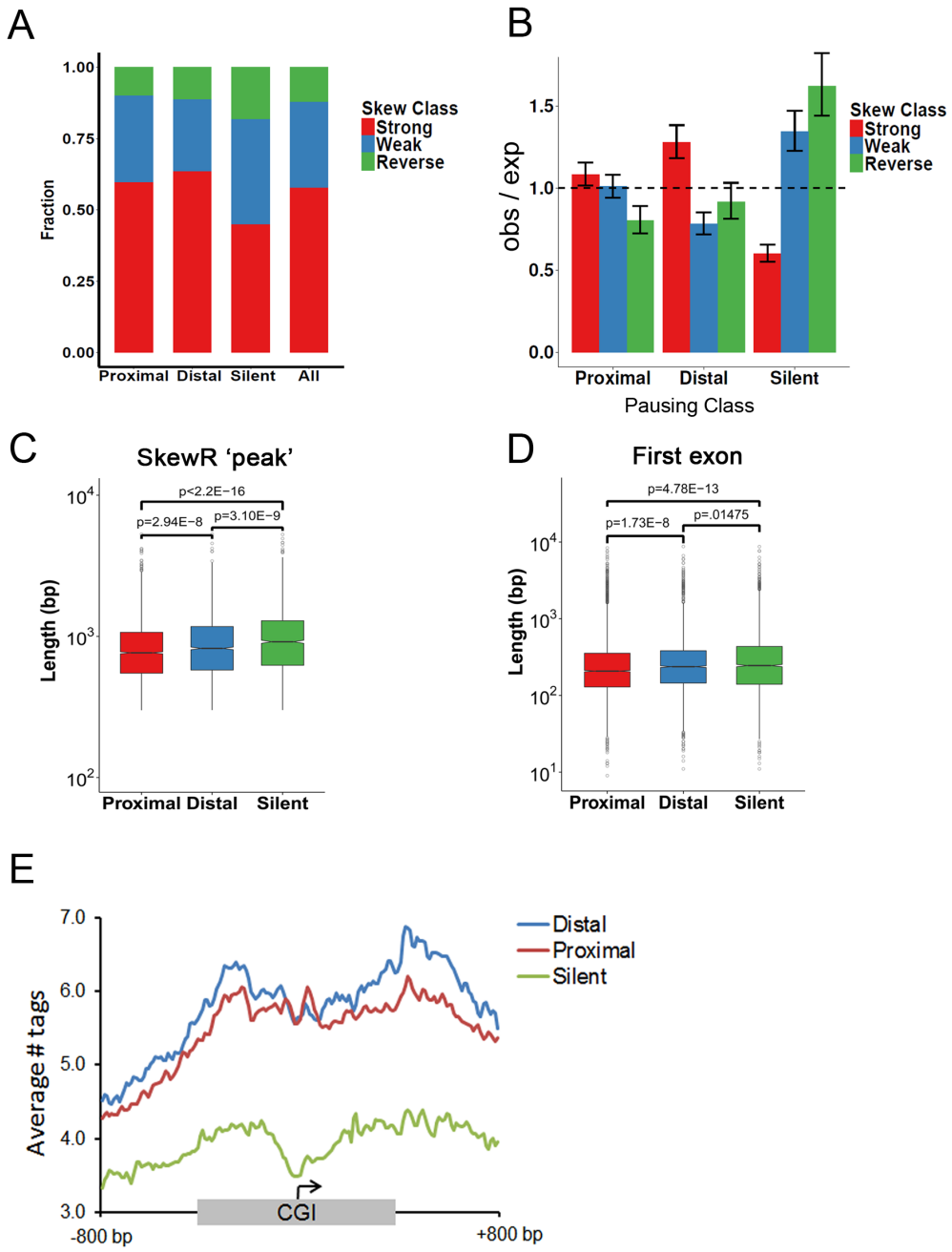

Figure S6: Distally paused CGI genes are associated with longer GC skew regions and are enriched in R-loops downstream of the TSS.

(A) Relationship between pausing class and skew "class" as defined by Ginno et al. (Ginno et al., 2012; Ginno et al., 2013). Shown is the fraction of CGI-associated TSSs in each 'skew' class across the 3 pausing classes (Proximal, Distal, Silent) relative to that of unique CGI- associated

TSS (All) for which both classes could be called (n=11552) (B) Enrichment of CGI associated TSS with various degrees of skew among the different pausing classes. Bars represent the odds ratio (Fisher's exact, observed/expected) and whiskers represent the 95% confidence interval. (C,D) Box plots comparing the SkewR peak lengths (C) and length of the first exon (D) among CpG island associated TSSs in each pausing class. CGI associated TSSs were intersected with SkewR 'peaks', an HMM that predicts R-loop forming regions based on the degree of GC skew (available from https://www.mcb.ucdavis.edu/faculty-labs/chedin/Resources.html). Approximately 80% of the TSS in each pausing class were found within a SkewR peak (Proximal N=4348, Distal N=2730, Silent N=2641). Median is indicated by a line, box denotes the first and third quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. Significance (p-value) was assessed by the Mann-Whitney *U* test. (E) Average tag densities of RNA:DNA hybrid analysis (DRIP-seq) for promoters in the three pausing classes. CGI associated promoters were oriented to transcription and the distance from the TSS to the upstream and downstream CGI edge independently scaled and anchored to the TSS (arrow). The average tags per 20 bp bin for 800 bp to either side of the CGI (unscaled) is included. Consistent with a lack of transcription, silent genes are depleted for R-loops relative to the other two classes.

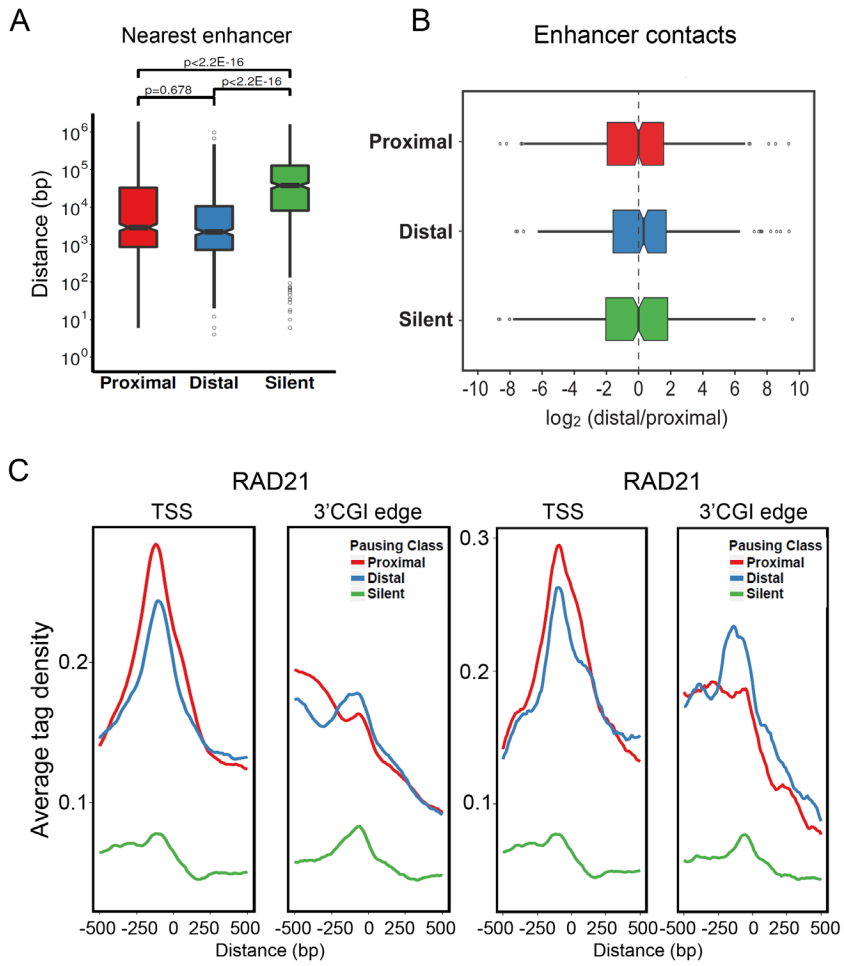

Figure S7. Spatial relationship between distal enhancer contacts and pausing class

(A) CGI associated TSSs in each pausing class were annotated to their nearest enhancer, defined as regions of overlapping H3K4me1 and H3K27ac peaks from GM12878 cells. Box plot representation of the distance from the CGI associated TSS to its nearest enhancer for genes in each class. Median is indicated by a line, box denotes the first and third quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. (B) Skewing of the frequency of promoter-enhancer contacts towards the 3'CGI edge in distally paused genes. Hi-C contact matrices at 1kb resolution from GM12878 cells (GSE63525) were used to determine the frequency of contacts between the nearest enhancer and the proximal (TSS+200bp) versus distal

(3'CGI edge +/- 100bp) pausing site for each gene. Given the resolution of the Hi-C data, the analysis was limited to those genes for which the TSS and 3'CGI edge are at least 1 kb apart, and that had at least one contact between each pause site and the nearest enhancer (Proximal n=527, Distal n=390, Silent n=308). Shown are box plots of the log2 ratio of enhancer-distal site contacts to enhancer-proximal site contacts among the genes in each pausing class. Median is indicated by a line, box denotes the first and third quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. Relative to CGI in the proximal or silent classes, which showed equivalent frequencies contacts at the TSS and distal site, distally-paused genes exhibit a greater frequency of interaction at the distal site (median ratio 1.25 (distally-paused class) vs. 1.0 (proximally paused class), p=.027, Mann-Whitney *U* test). (C) Average tag densities for the cohesin subunit RAD21 across CGI in the three pausing classes. RAD21 ChIP-seq from MCF7 cells (GSM101079) was used to determine the average tag density per 20 bp bin for +/- 500bp anchored at the TSS or the 3' edge of the CGI, for all promoters in each class (proximal, n=5889; distal n=3663) (left), or the top 20% in each class ranked by pausing index (proximal, n=1289; distal n=738) (right). In both cases, all CGI in the silent class were used (n= 2724).

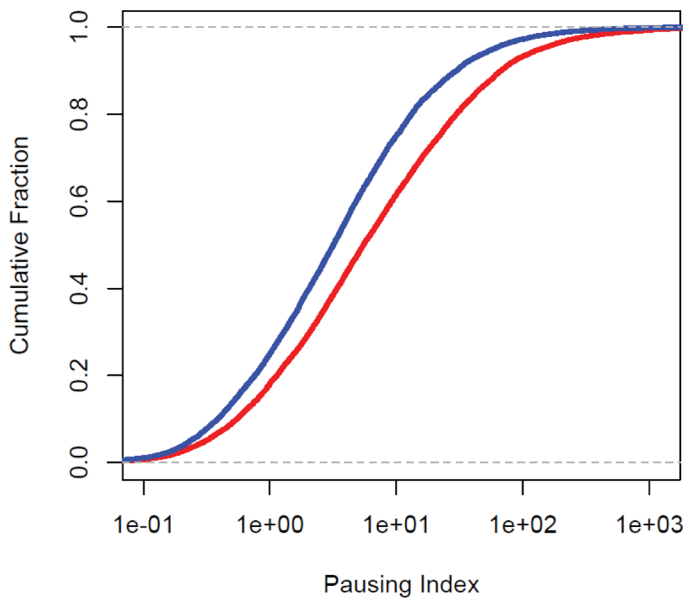

Figure. S8. Relative intensity of the proximal and distal pause.

Cumulative distribution plots showing the pausing index at the proximal (red) or distal (blue) pause points. Pausing index was calculated from GRO-seq tag density in the 100bp underlying the proximal (+40 to +140 bp from TSS) and distal (+20 to +120 bp from the 3' CGI edge) peaks relative to that in the gene body (+120 bp from 3' CGI edge to TES).

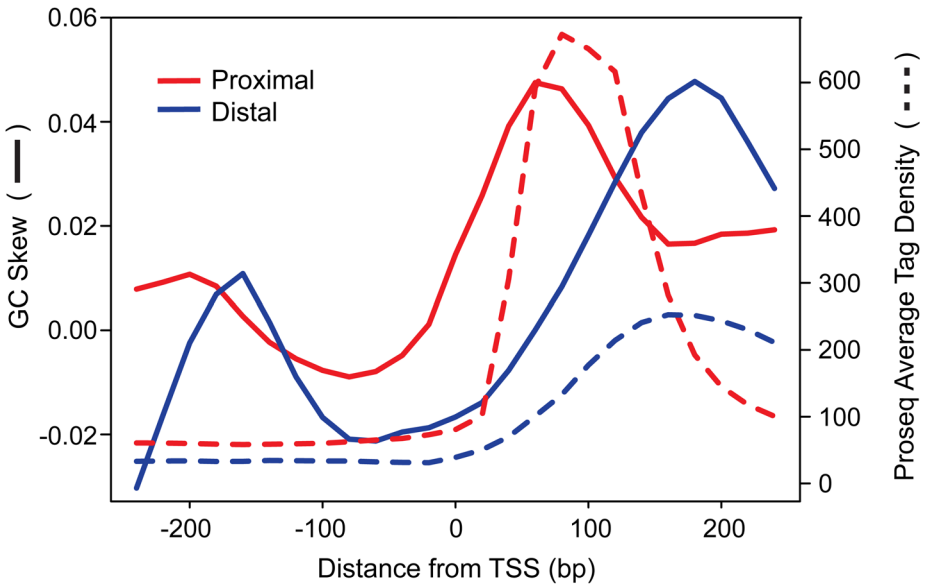

Figure S9. Relationship between GC skew and pausing class in *Drosophila*

Average GC skew (solid lines) and Pro-seq tag densities (dotted lines) were computed for +/-250 bp surrounding the TSS of *Drosophila* genes defined as proximally (red) versus distally (blue) paused by Kwak et al. (Kwak et al., 2013). Note the tight relationship between the position of the peak of GC skew and the Pol II pause in each pausing class.

Supplemental Table I – Primers used for 3C Analysis

Gene/Fragment	Sequence
SIAH2_Constant	TCAGATATTAATTGGGTGCCTAGGT
F22_SIAH2	ATAGAAACGACAGCCCTGGGA
F21_SIAH2	CCAACACTTCTTTGGCCCCTA
F20_SIAH2	CTGTGTTCGGAGTCAGGACG
F19_SIAH2	TGCTCCTGAAGGTTTCCACG
F18_SIAH2	CCAAGTTTTACTGGTGCGGC
F17_SIAH2	TTGGTTACACACCAGGTGCC
F16_SIAH2	GTGGTGCTGCGGGGAC
F15_SIAH2	CTTCCTGCTCGGGCTGC
F14_SIAH2	TCAGGACGAGAAGCATTGGG
F7_SIAH2	TGAACCGCATGTCCAAATGT
F1_SIAH2	ATGGCGTAAGAGCCCAGAAG
SIAH2 Taqman Probe	6FAM-CCAGGGACTTCATTG-MGBNFQ
MYC_constant	TGGCTGTTTACCTGGGATCCT
F1_MYC	TGCTCTCTCCTCTGCCGAAA
F4_MYC	GTGACTCACACTGGCAAATTCT
F9_MYC	TATTACCTCCACTACCTGGGGC
F14_MYC	TCCTCCCTGATAGAAGCTCCA
F22_MYC	TACAGCACTTCAAAGCCTCCC
F41_MYC	CATCATCTACAGGGGAGCAGC
F51_MYC	AAACAACCAAGGGTGAGCTACT
F52_MYC	GGGGAAGGGACAACACTAAGC
F53_MYC	TACTGGGCTGGGGTATCAGG
F54_MYC	ACGGAAGTAATACTCCTCTCCTC
F55_MYC	ACTCAGTCTGGGTGGAAGGT
F56_MYC	GACTCTTGATCAAAGCGCGG
F57_MYC	TACTGCGACGAGGAGGAGAA
F58_MYC	CTCCACCTCCAGCTTGTACC
F59_MYC	AGAAATGTCCTGAGCAATCACCT
F60_MYC	ACTTAGAGAGCTCACAGCTTGG
Myc Taqman Probe	6FAMA-CAATGTGTTGCAAGAGT-MGBNFQ
P2RY2_constant	TTGCCCAGGCTGCAATG
F1_P2RY2	CACTGGCCTGGAGATTCAAC
F2_P2RY2	CCTTGGCTGCTTGGTTCCAG
F3_P2RY2	CAGTCAGCTGATATGGAGCCC
F4_P2RY2	CCAGCTCCCTTCTAGCGTG
F5_P2RY2	CAGACACGCTGACCCCG
F6_P2RY2	CTTCGGGGTTGGGGAACAG
F7_P2RY2	GCACCCTGAGAGGAGAAGC

F9_P2RY2	CCAGACTGGCGCAGGTG
F10_P2RY2	GCCAGAAAGGACAGTTAAGCC
F11_P2RY2	AGAAACAGAGCAGTGGCGTG
F12_P2RY2	GCGCTTCCTCTTCTACACCA
F13_P2RY2	CTGCCGCTGCTGGTCTATTA
F14_P2RY2	CTGGATAATGCCGAGTGGCT
F15_P2RY2	ACCTCAGTGAAGGCACAACC
P2RY2 Taqman Probe	6FAM-TGGCACAATCTCGG-MGBNFQ