Supplementary Information:

MIR retrotransposon sequences provide insulators to the human genome

Jianrong Wang, Cristina Vicente-García, Davide Seruggia, Eduardo Moltó, Ana Fernandez-Miñán, Ana Neto, Elbert Lee, José Luis Gómez-Skarmeta, Lluís Montoliu, Victoria V. Lunyak and I. King Jordan

Supplementary Methods

MIR-insulator prediction algorithm. RepeatMasker annotations were used to identify 590,373 MIR sequences in the human genome reference sequence (NCBI build 36.1, USCC version hg18). All human MIR sequences were analyzed in a series of steps to progressively narrow the list of candidate MIR-insulators to a final set of predictions (Fig. 1A). First, individual MIR sequences were compared to their constituent consensus sequences from Repbase (1) to identify MIRs that bear intact B-boxes (*n*=324,863). Then, B-box containing MIRs bound by RNA Pol III were identified by analyzing ChIP-seq data (2) (*n*=124,278). To do this, the number of ChIP-seq tags for RNA Pol III binding of each B-box containing MIR sequence was counted extending each candidate MIR sequence by 100 bp upstream and downstream to account for the potential underestimation of Pol III binding levels in MIRs. The tag count of every MIR sequence was then transformed into a *P*-value based on the Poisson distribution parameterized by the genomic average of the RNA Pol III binding tag counts: $P = \sum_{k=T}^{\infty} (\lambda^k / k!) e^{-\lambda}$ where T is the tag count of RNA Pol III binding and λ is the genomic average RNA Pol III binding tag counts. The *P*-value is then used to indicate the statistical significance of Pol III binding level. All B-box containing MIRs with Pol III binding levels more significant than the *P*-value threshold (0.05) were saved for further screening. Candidate MIR-insulators were then evaluated for their ability to segregate active versus repressive chromatin domains. To do this, 39 $CD4^+$ T cell histone modifications were classified as active (34 modifications) or repressive (5 modifications) based on their associations with transcribed or silent genes as previously described (3). Distributions of ChIP-seq counts for these active versus repressive histone modifications were analyzed for 100kb genomic regions centered on candidate MIR-insulators using a maximal segment algorithm that we previously developed to identify active and repressive chromatin domains (4). The maximal segment algorithm uses a probabilistic scoring scheme to identify a 'leading edge' that delineates adjacent contiguous regions that are enriched for active versus repressive histone modifications. The scoring scheme uses normalized ChIP-seq tag counts for histone modifications over 200bp non-overlapping windows in the regions upstream and downstream of the candidate MIR-insulators. A score for every bin is calculated as the logarithmic ratio of active modification tag densities over repressive modification tag densities. Thus, bins with more active histone modification tag densities have positive scores and bins with more repressive histone modification tag densities have negative scores. The sub-regions with maximal local cumulative positive scores detected by maximal-segment algorithm then represent contiguous regions enriched with active modifications. Contiguous regions with repressive modifications are found in the same way and are called repressive regions. MIRs

located between active and repressive regions, or MIRs located within active or repressive regions but close to the edge of the region, were selected for further consideration (*n*=22,620). The two sides of the screened MIRs are assigned as active side and repressive side separately based on the state of the most proximal region (i.e. active or repressive). Lastly, candidate MIR-insulators from this set that delineate expressed versus silent genomic regions were selected as the final set of predicted MIR-insulators. To do this, $CD4^+$ T cell RNA-seq levels (2) were compared against genomic background for 100kb windows centered on candidate MIR-insulators taking into consideration the locations of their previously determined active versus repressive domains. The MIRs that had RNA-seq levels above the genomic background, at $P<0.01$ determined using the Poisson distribution parameterized with the RNA-seq tag count per position genomic average, in the adjacent active domain and RNA-seq levels indistinguishable from genomic background in the adjacent repressive domain were taken for further consideration. Candidate MIRs that are located in intergenic regions and are distant from each other (>10kb, to reduce ambiguity) are selected as putative MIR-insulators (*n*=1,178).

MIR-insulator pipeline performance test. Having predicted MIR-insulators as described above, we performed a series multi-dimensional statistical analyses to test the performance of the pipeline of pinpointing specific MIR sequences that substantially segregate individual histone modifications and partition active and repressive modifications. To do this, ChIP-seq tag distributions were evaluated for all 39 $CD4^+$ T cell histone modifications upstream and downstream of the predicted MIR-insulators. The upstream and downstream 50kb regions of each predicted MIR-insulator were divided into non-overlapping 200bp bins, and the tags of each active and repressive histone modifications were counted. If a bin has more than 5 tags of an individual modification, then the bin is considered as a reliable modified site of the corresponding modification. The number of reliable modified sites is determined in this way for upstream and downstream regions of each MIR. Thus, for each individual histone modification, two arrays were obtained: 1) the upstream modified site number array, each element of which represents the number of reliable modified sites in the upstream region of a single MIR sequence and 2) the downstream modified site number array, which is the downstream counterpart of upstream modified site number array (Fig. S4). These arrays are denoted as follows: $u_i = (u_{i1}, u_{i2}, ..., u_{in})$ and $d_i = (d_{i1}, d_{i2}, ..., d_{in})$ where i indicates the ith histone modification, n indicates the total number of predicted MIR-insulators. u_i represents the upstream modified site number array for the ith histone modification, where u_{ij} is the number of modified sites of the ith histone modification in the upstream region of the jth MIR sequence. d_i represents the downstream modified site number array for the ith histone modification, where d_{ij} is the number of modified sites of the ith histone modification in the downstream region of the jth MIR (Fig. S4). If an individual histone modification is blocked and restricted to only one side or the other of the predicted MIR-insulators, then the upstream and downstream modified site number arrays are expected to be significantly negatively correlated (Fig. S4). Spearman correlation coefficients were computed for all 39 individual histone modifications to test this prediction.

For the next step of the computational validation, the ability of the predicted set of MIR-insulators to group active and repressive modifications together were computationally validated using correlation analysis of ChIP-seq data for the 39 $CD4^+$ T cell histone modifications. To do this, the upstream and downstream modified site number arrays for each histone modifications are joined into a single array. The joined arrays were then used to represent the distribution profiles of each histone modification across the predicted MIR-insulators: $h_i = (u_{i1}, u_{i2}, ..., u_{in}, d_{i1}, d_{i2}, ..., d_{in})$ (Fig. S4). Spearman correlation coefficients were calculated for each pair of histone modifications to test whether they are partitioned by the putative MIR-derived insulators globally since if the active and repressive modifications are partitioned by the MIRs, then their joined arrays are expected to be significantly negatively correlated (Fig. S4). Hierarchical clustering was employed on the joined arrays to show whether active and repressive modifications form distinct clusters, indicating whether their distributions across the predicted MIR-insulators are mutually exclusive to each other. As an additional validation step, principal component analysis was used to project the high dimensional joined array into the three dimensional space spanned by the first three principal components, and to visualize the relative distances of active and repressive modification arrays (Fig. S6).

References

1. Jurka J*, et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110(1-4):462-467.
2. Barski A*, et al.* (2010) Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat Struct Mol Biol* 17(5):629-634.
3. Wang Z*, et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 40(7):897-903.
4. Wang J, Lunyak VV, & Jordan IK (2012) Genome-wide prediction and analysis of human chromatin boundary elements. *Nucleic Acids Res* 40(2):511-529.

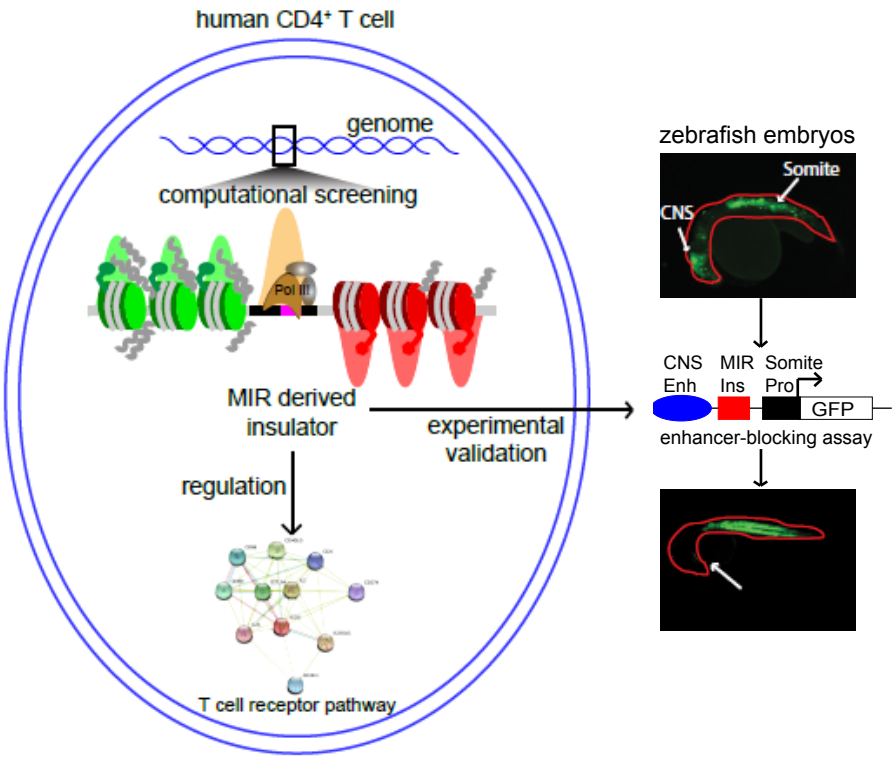


Figure S1: Summary of the study design for prediction and validation of putative MIR-derived insulators. A computational screening pipeline is developed and applied on genomic and functional datasets of human CD4+ T cells. MIR retrotransposon sequences with intact B-box (purple) and RNA Pol III binding (yellow) are further screened for their ability to partition active (green) and repressive histone modifications (red), along with the partition of active and repressive transcriptions (gray curves). The putative MIR derived insulators are validated using enhancer-blocking assay (EBA). For zebrafish EBA, the MIR-insulators (MIR Ins) are transiently transfected between a central nervous system enhancer (CNS Enh) and a somite promoter (Somite Pro) for GFP expression. GFP expressions in CNS are expected to be suppressed if the putative MIR-insulator can block the interactions between the enhancer and promoter. Putative MIR-insulators are further analyzed for their potential ability to regulate T cell related functional pathways. Genes of T cell receptor pathway are found to be enriched in proximal flanking regions of putative MIR-insulators

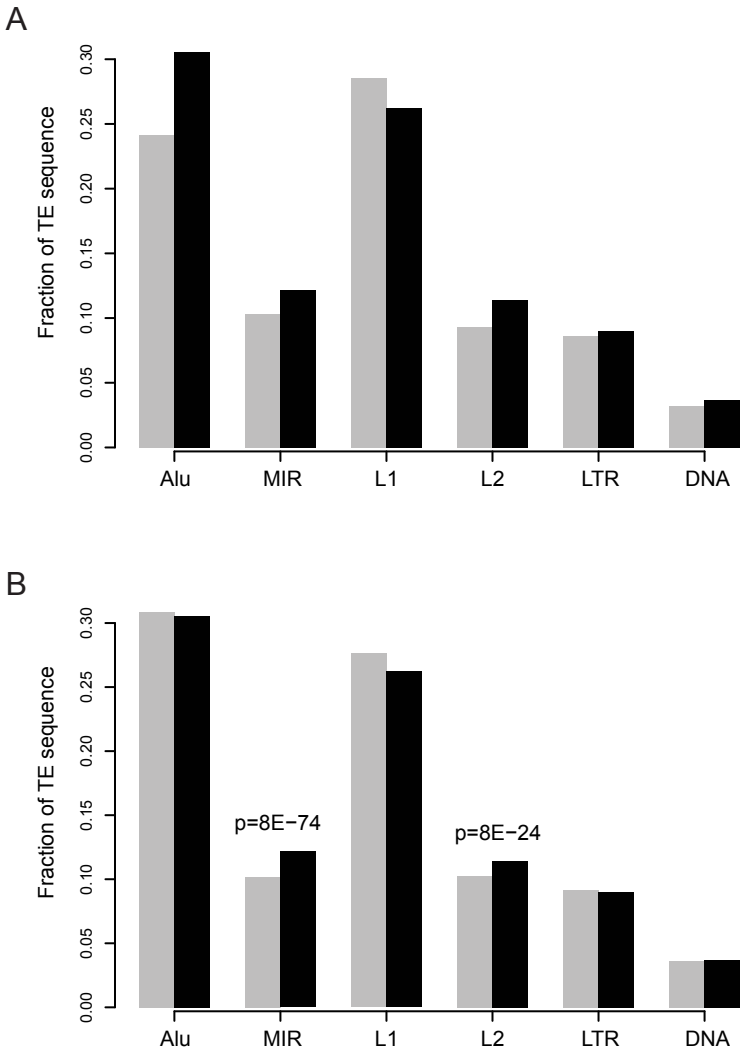


Figure S2: Enrichment of different transposable elements (TE) in transition regions between repressive and active chromatin domains. A) Fractions of transition region sequences derived from TE families (black bars) compared to fractions of the whole genome derived from those TE families (grey bars). B) Fractions of transition region sequences derived from TE families (black bars) compared to fractions of the flanking sequences around transition regions derived from those TE families (grey bars).

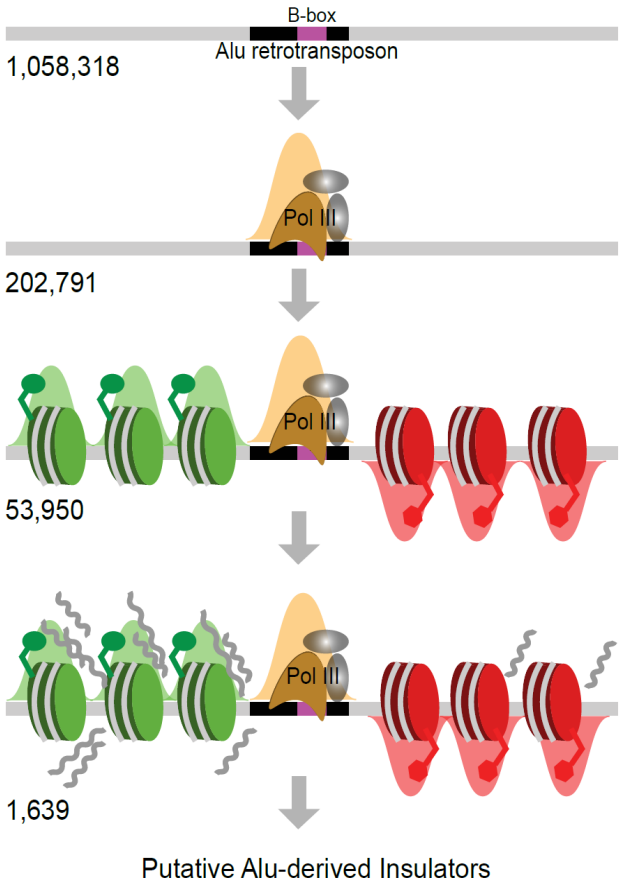


Figure S3: Computational screen for insulator-like Alu elements. The same pipeline was applied on Alu retrotransposons, and the numbers of Alu elements after each step are listed.

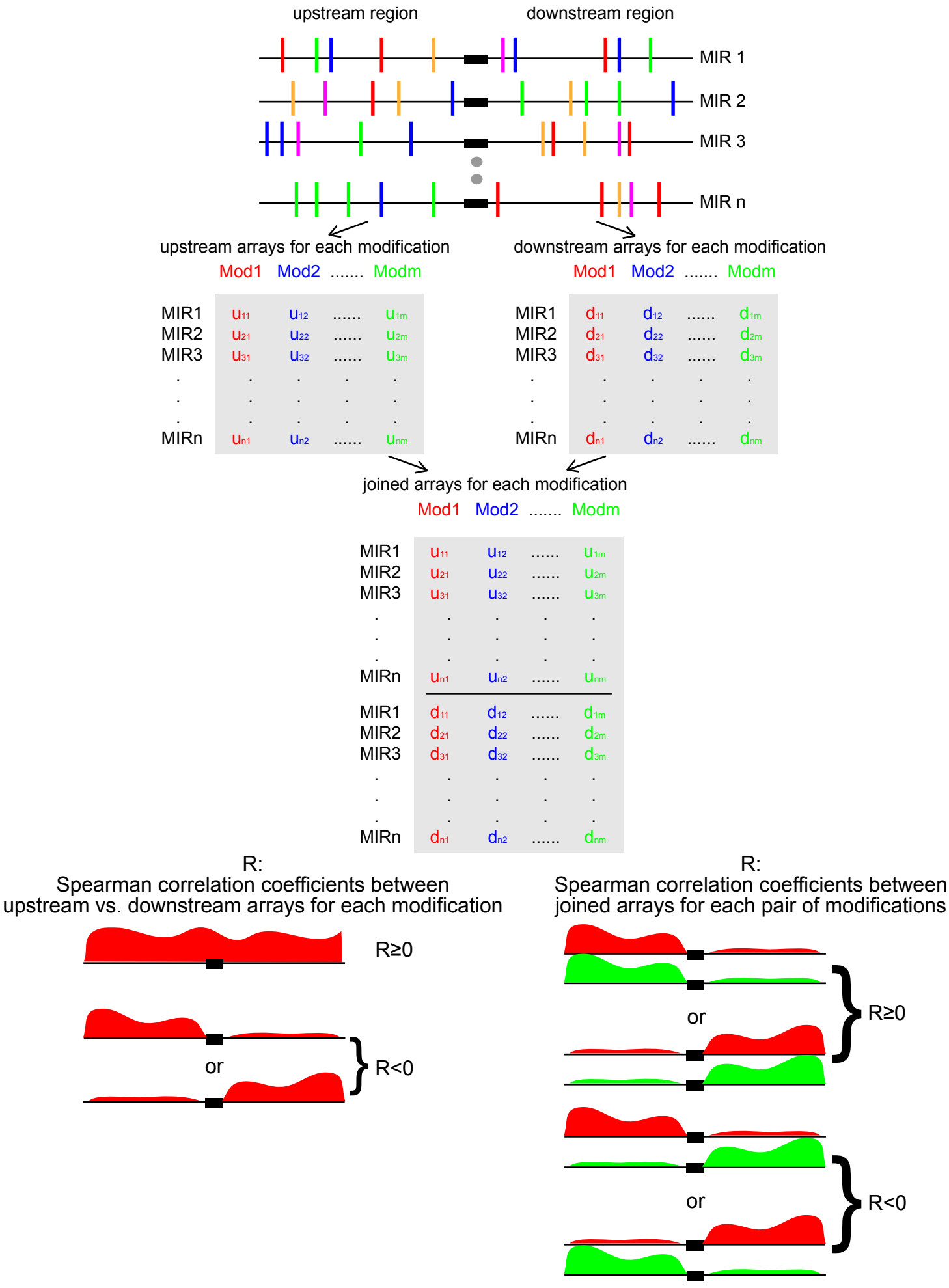


Figure S4: Scheme illustrating the performance evaluation procedure (described in detail in the Supplementary Methods).

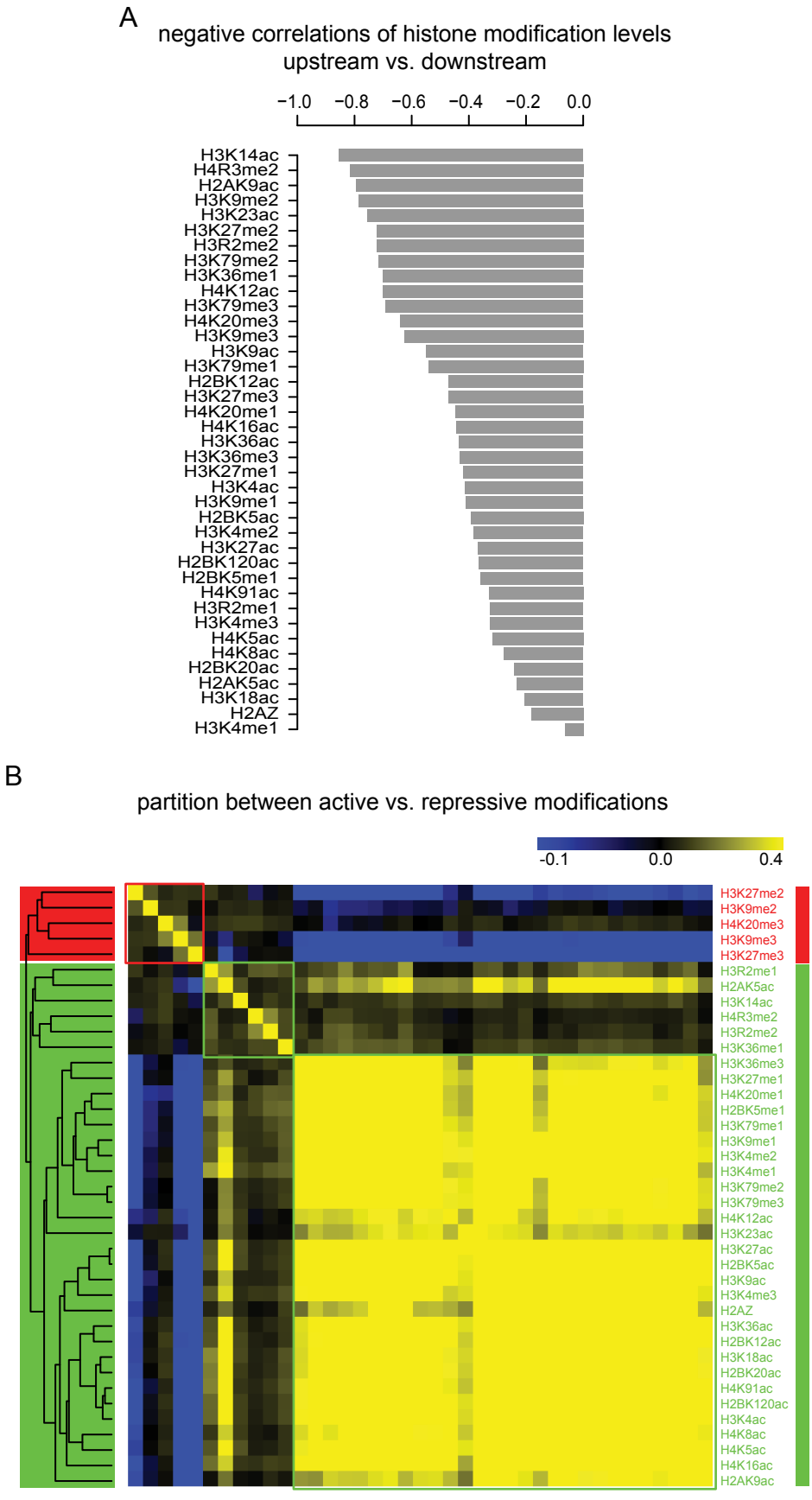


Figure S5: Partition of active and repressive marks across MIR-insulators. (A) Spearman correlations for individual histone modification profiles upstream versus downstream of predicted MIR-insulators. (B) Heatmap showing Spearman correlations for pairs of histone modification profiles upstream versus downstream of predicted MIR-insulators. Hierarchical clustering based on the correlation matrix groups repressive (red) and active (green) histone modifications.

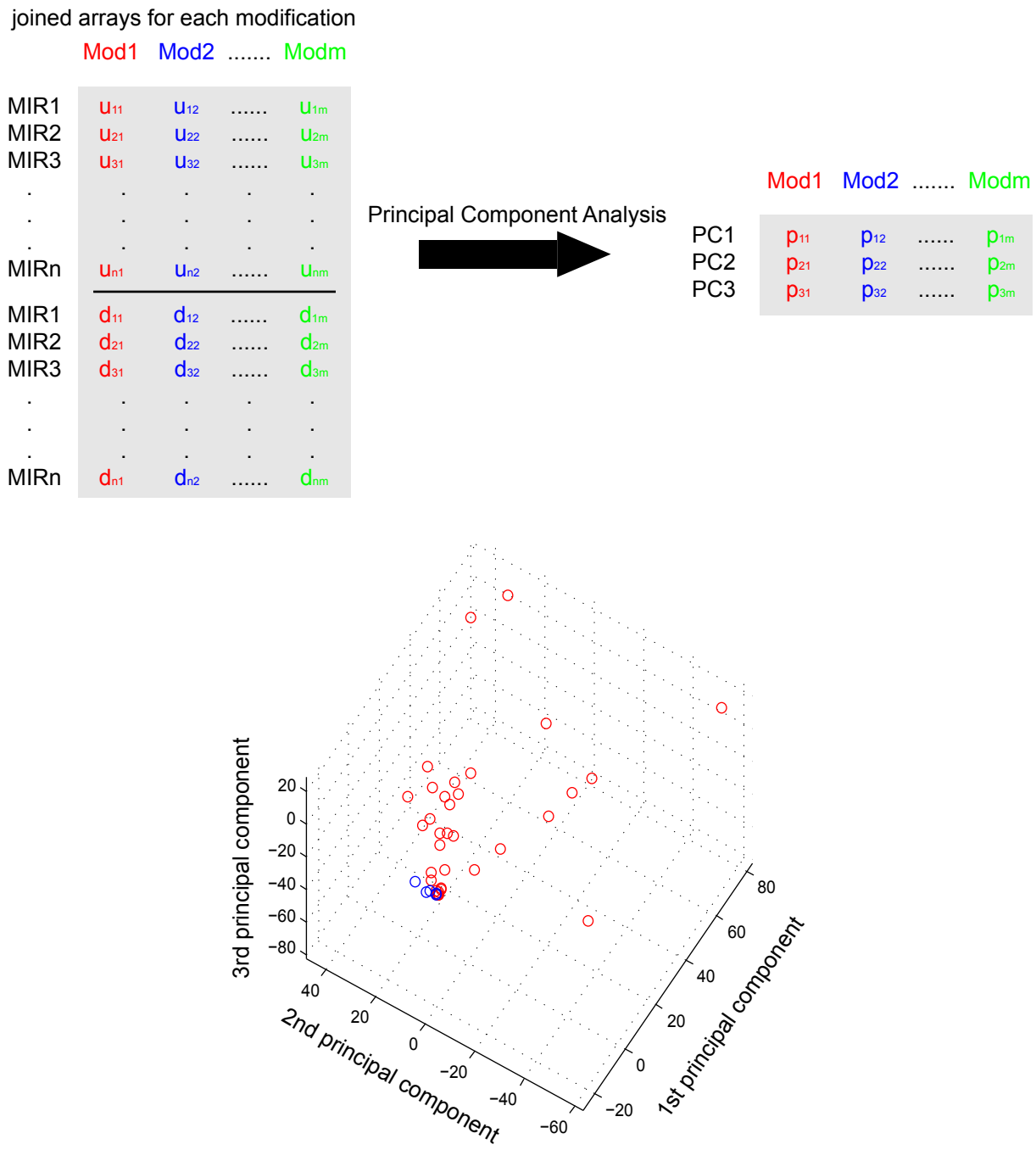


Figure S6: Scheme and results of the principal components analysis (PCA) applied on histone modification distributions around predicted MIR-insulators. Upper, the joined histone modification arrays projected onto the first three principal components from the PCA analysis. Lower, a three-dimensional plot showing the locations of individual active (red) and repressive (blue) histone modifications in the principal component space.

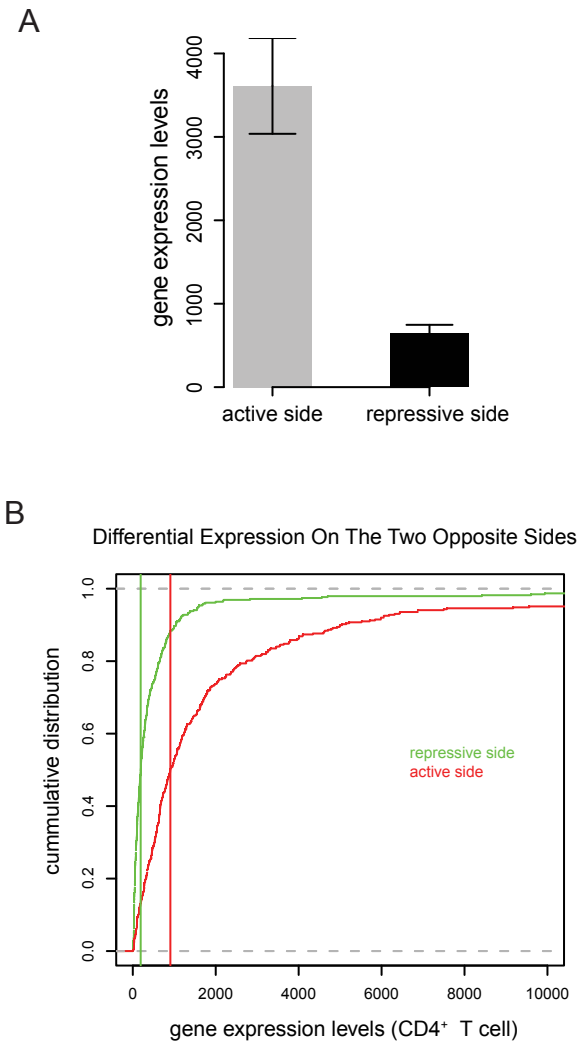


Figure S7: Differential gene expression on opposite sides of MIR-insulators. A) Average (± standard error) CD4+ T cell expression levels (Affymetrix signal intensity values) of proximal genes from the active (grey) and repressive (black) sides of predicted MIR-insulators. B) Cumulative distributions of the CD4+ T cell gene expression levels for MIR-insulator proximal genes located on the repressive (green) and active (red) domain sides.

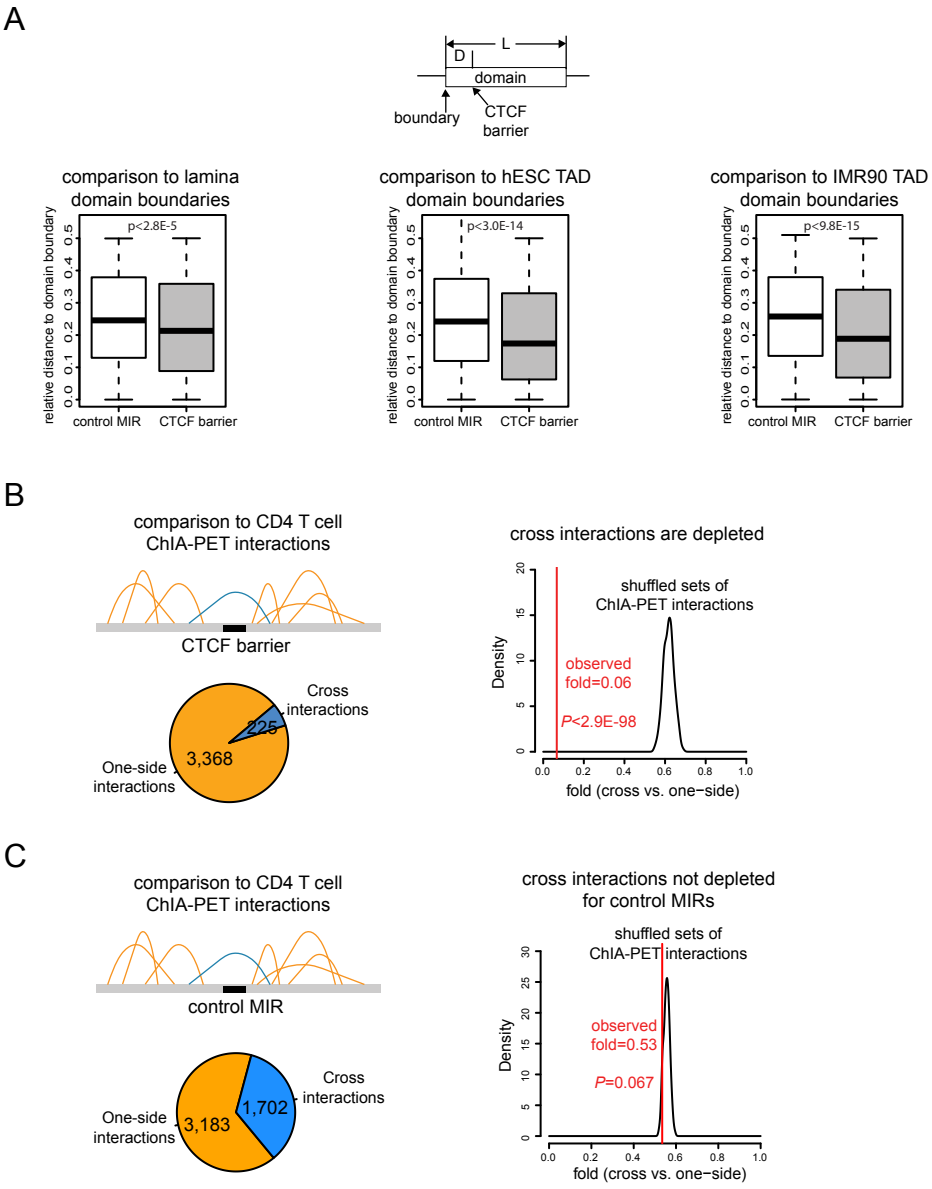


Figure S8: Insulator function evaluation for CTCF barriers and control MIRs. A) Distribution of relative distances to boundaries of lamina domains, hESC Hi-C topological domains and IMR90 Hi-C topological domains, compared to the distributions of random B-box containing MIR elements. B) Depletion of CD4+ T cell ChIA-PET interactions across CTCF barriers. C) Evaluation of cross-interactions around random B-box containing MIR sequences. Random B-box containing MIR sequences with ChIA-PET interactions in their local regions (within +/-500kb) are selected as negative controls. Cross-interactions around control MIRs are compared to one-side interactions. Calculations based on shuffled ChIA-PET interactions are also carried out to demonstrate the background distribution of folds for those control MIRs.

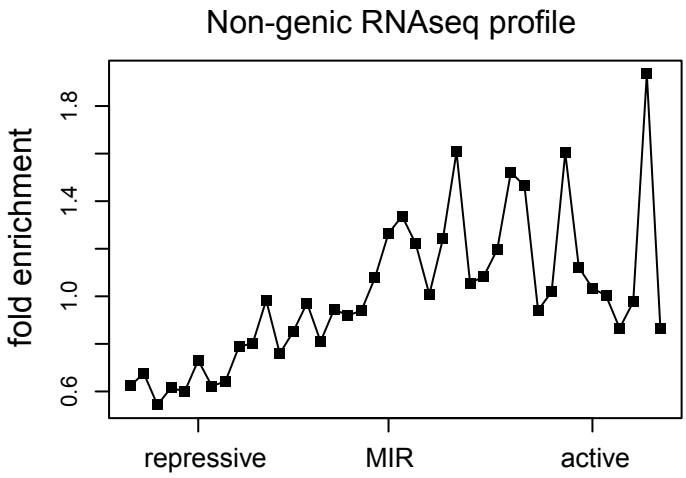


Figure S9: Non-genic RNAseq signal profile around MIR-insulators. MIR-insulators are extended by 4kb on each side, and aligned and oriented with the MIR elements in the middle and repressive chromatin side on the left. CD4+ T cell RNAseq data are purified, and only non-genic RNAseq signals are used. Fold enrichment is computed by comparing to the genomic average RNAseq signal.

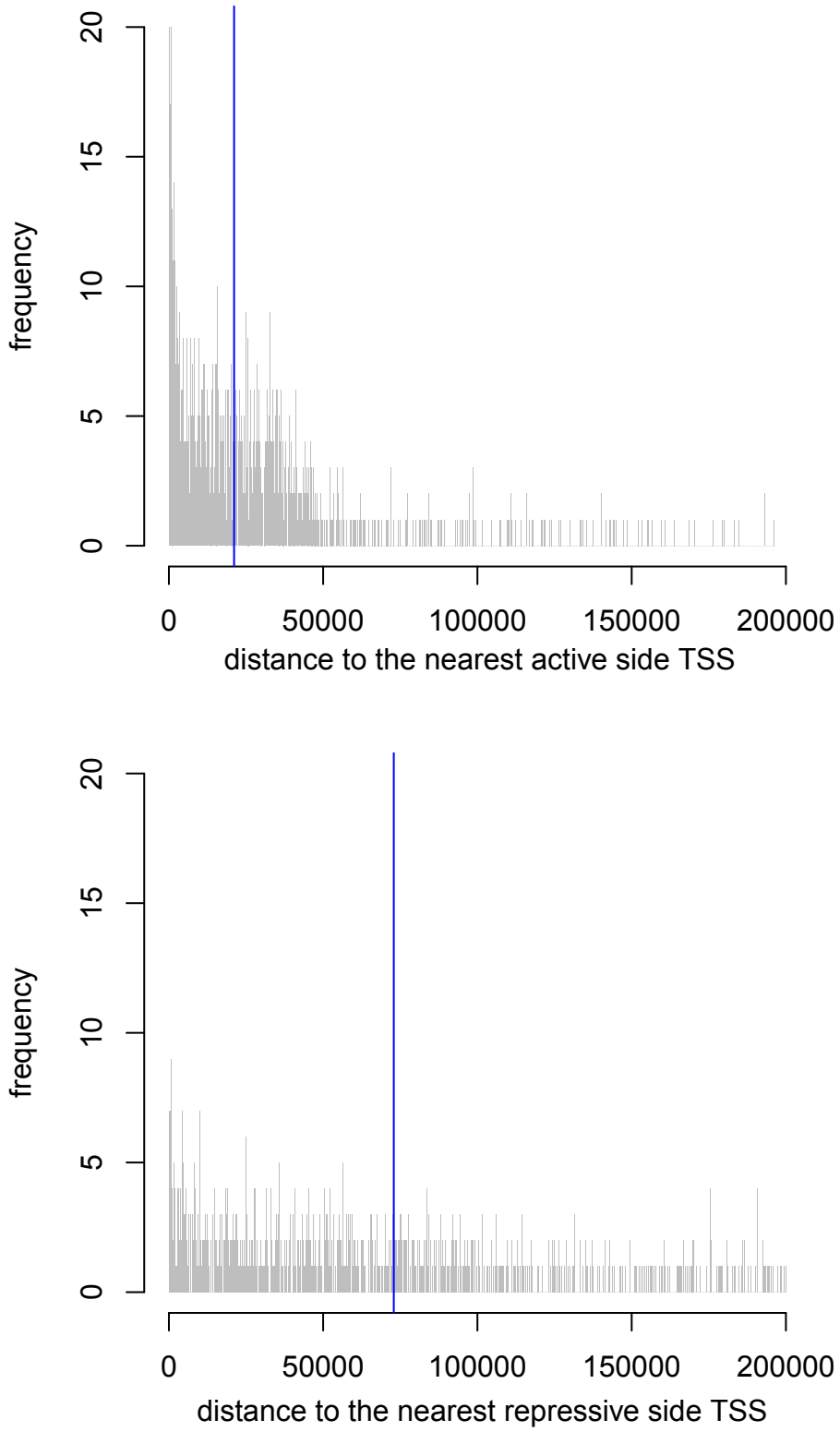


Figure S10: Distributions showing the distances between predicted MIR-insulators and the nearest gene TSS on the active domain side (upper), and the nearest gene TSS on the repressive domain side (lower). Median values of the distributions are shown in blue on each plot.

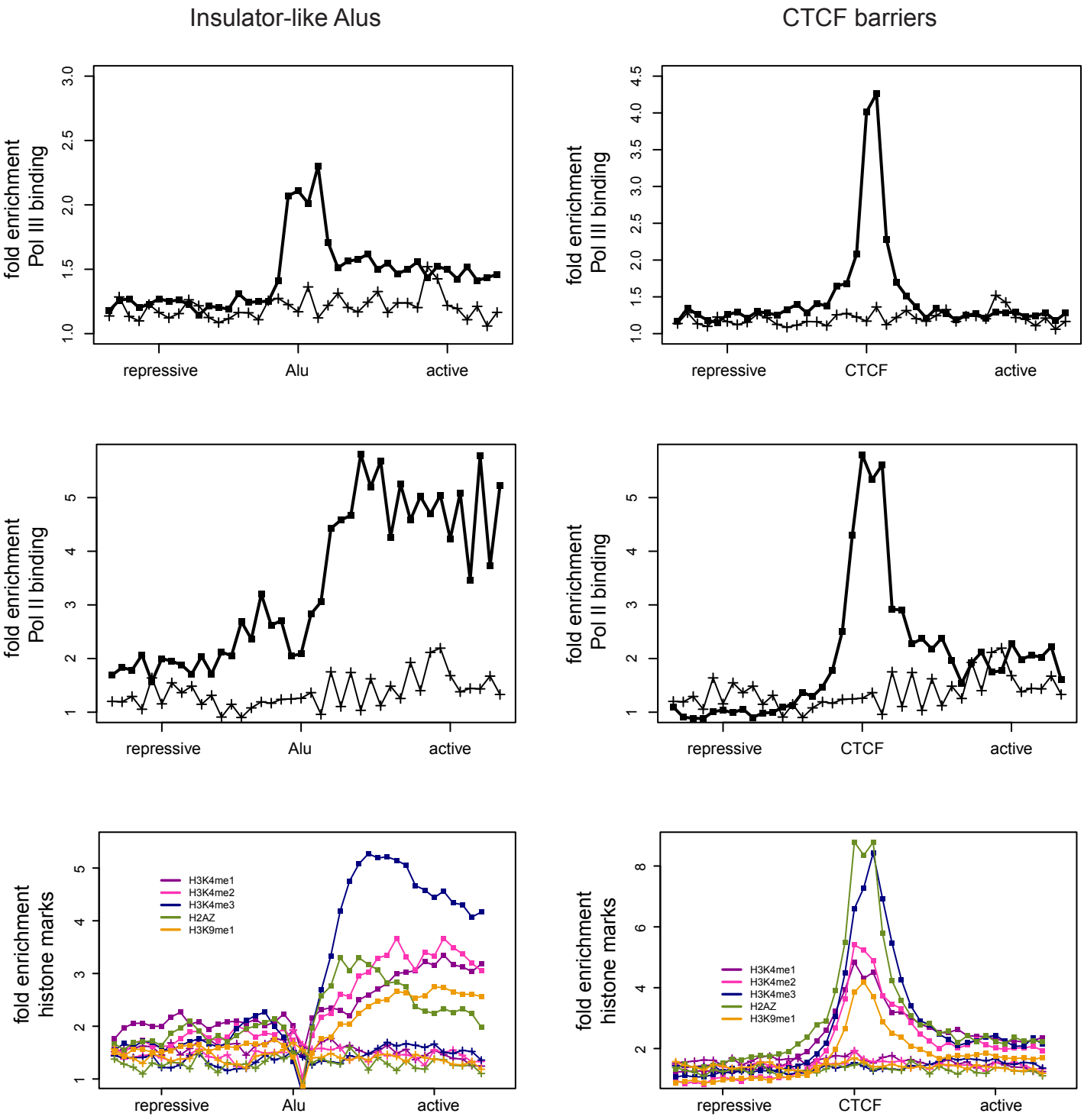


Figure S11: Signature profiles around insulator-like Alu elements (left) and CTCF barriers (right). Random B-box containing MIR sequences are used to show the background profiles for each plot.

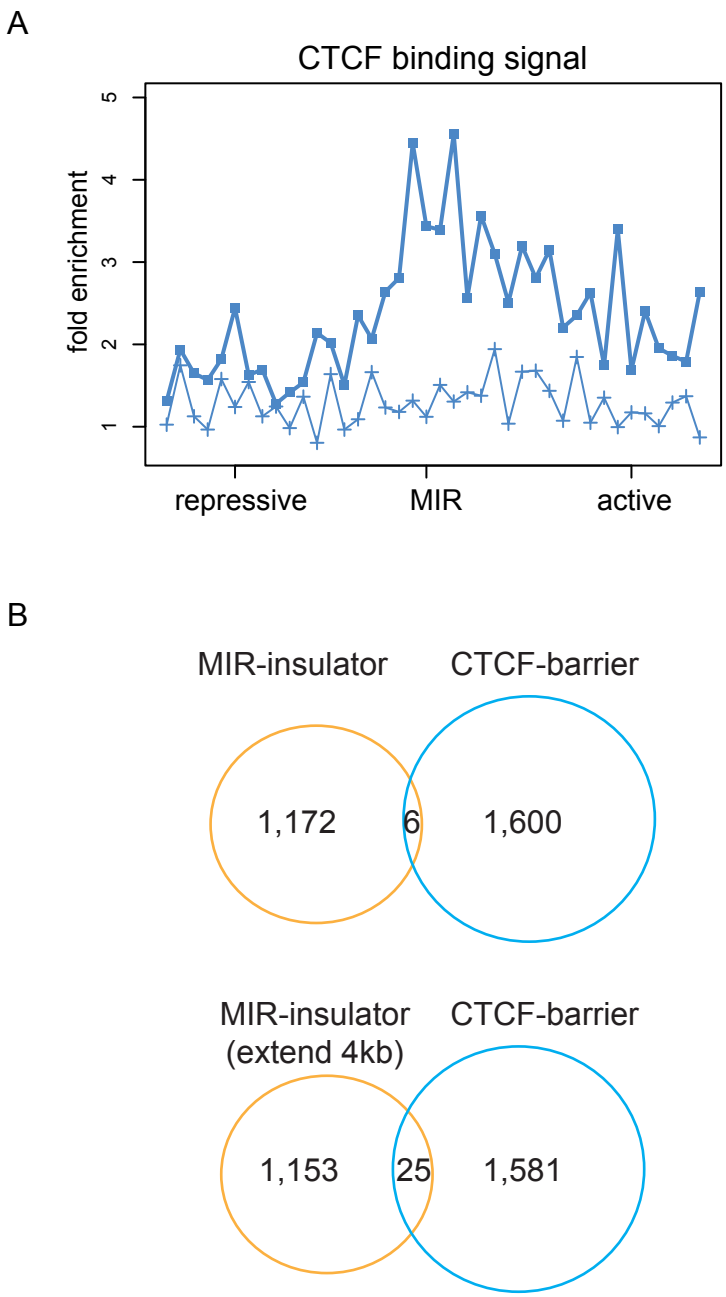


Figure S12: Comparison of MIR-insulators with CTCF. A) CTCF binding signal profile around MIR-insulators (blue curve). Fold enrichment is computed by comparing to genomic average CTCF signal. MIR-insulators are extended by 4kb on each side, and aligned and oriented with MIRs in the middle and the repressive chromatin side on the left. The signal profile around random B-box containing MIR sequences (light blue curve) is used as a comparison. B) Overlap of MIR-insulators with CTCF barrier elements. MIR elements (upper) or extended MIRs (lower) are compared to CTCF barriers in CD4+ T cells.

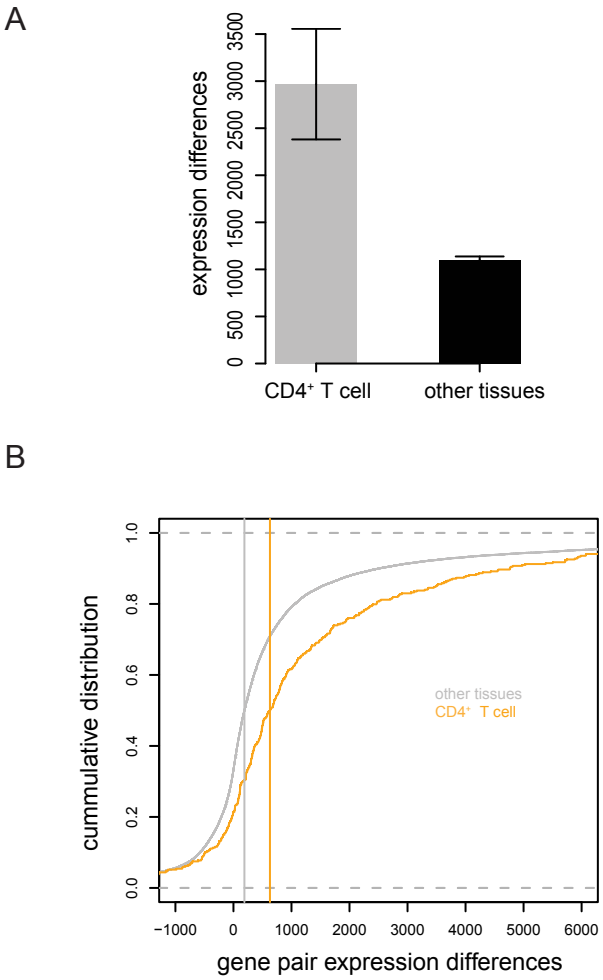
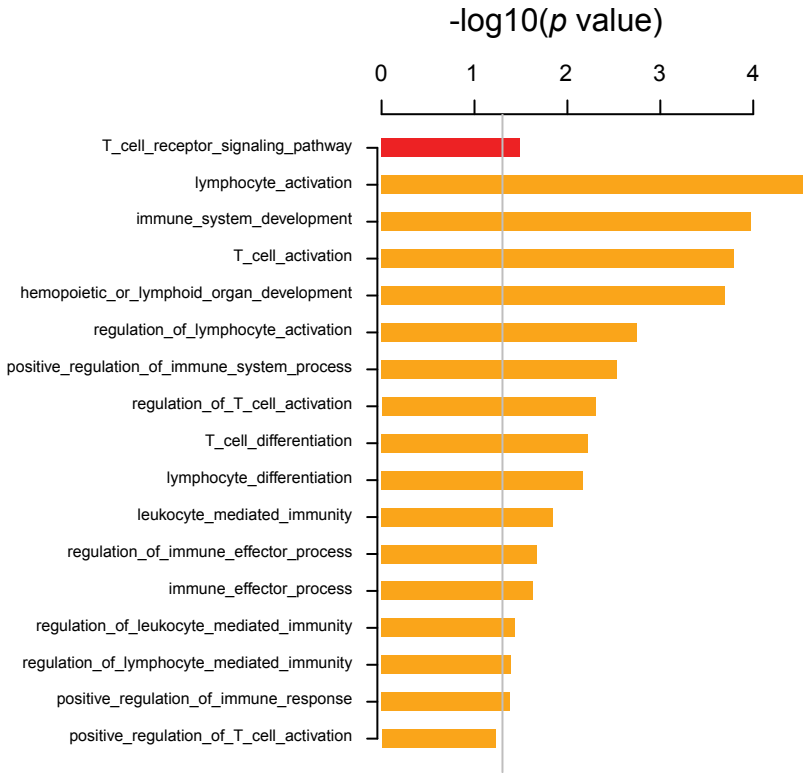


Figure S13: Tissue specific differential gene expression across MIR insulators. A) Average (± standard error) differences in gene expression levels (Affymetrix signal intensity values) for genes located on the opposite sides of individual predicted MIR-insulators. Differences are shown for CD4+ T cell expression levels compared to average differences for expression levels across 78 different tissues. B) Cumulative distributions of the differences in the gene expression levels for genes located on the opposite sides of individual predicted MIR-insulators. Difference distributions are shown for CD4+ T cell expression levels (orange) and for expression levels across 78 different tissues (grey).

A



B

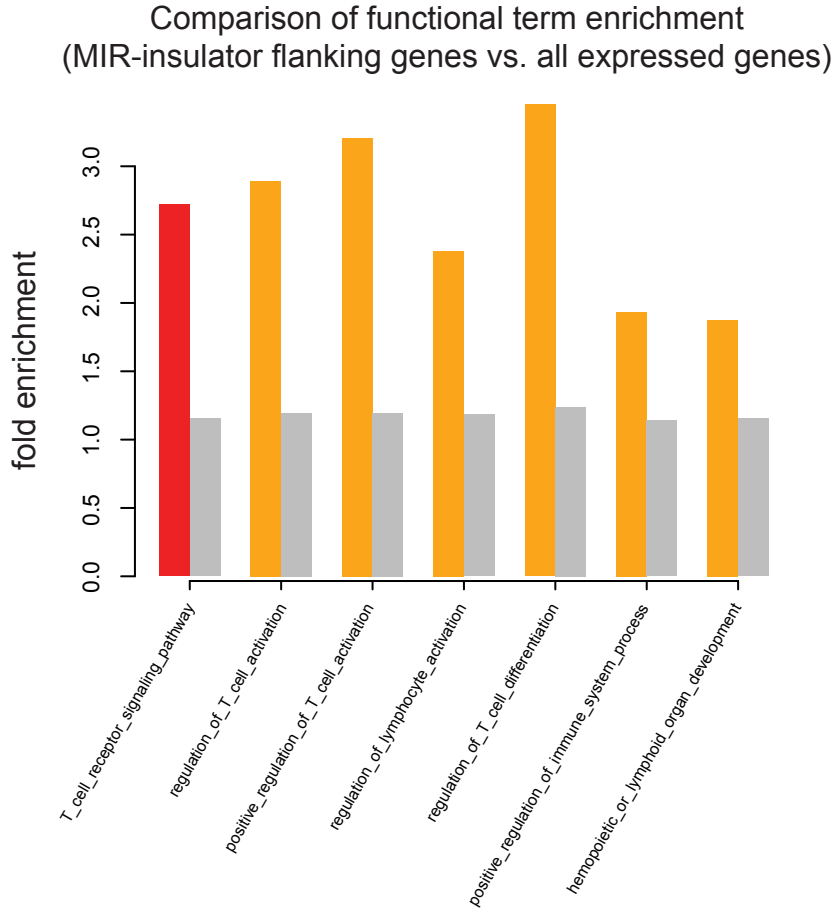


Figure S14: Comparison of pathway enrichment. A) Pathway enrichment for all expressed genes in CD4+ T cells. B) Comparison of pathway enrichment (fold enrichment) for genes close to MIR-insulators vs. all expressed genes in CD4+ T cells (grey bars).

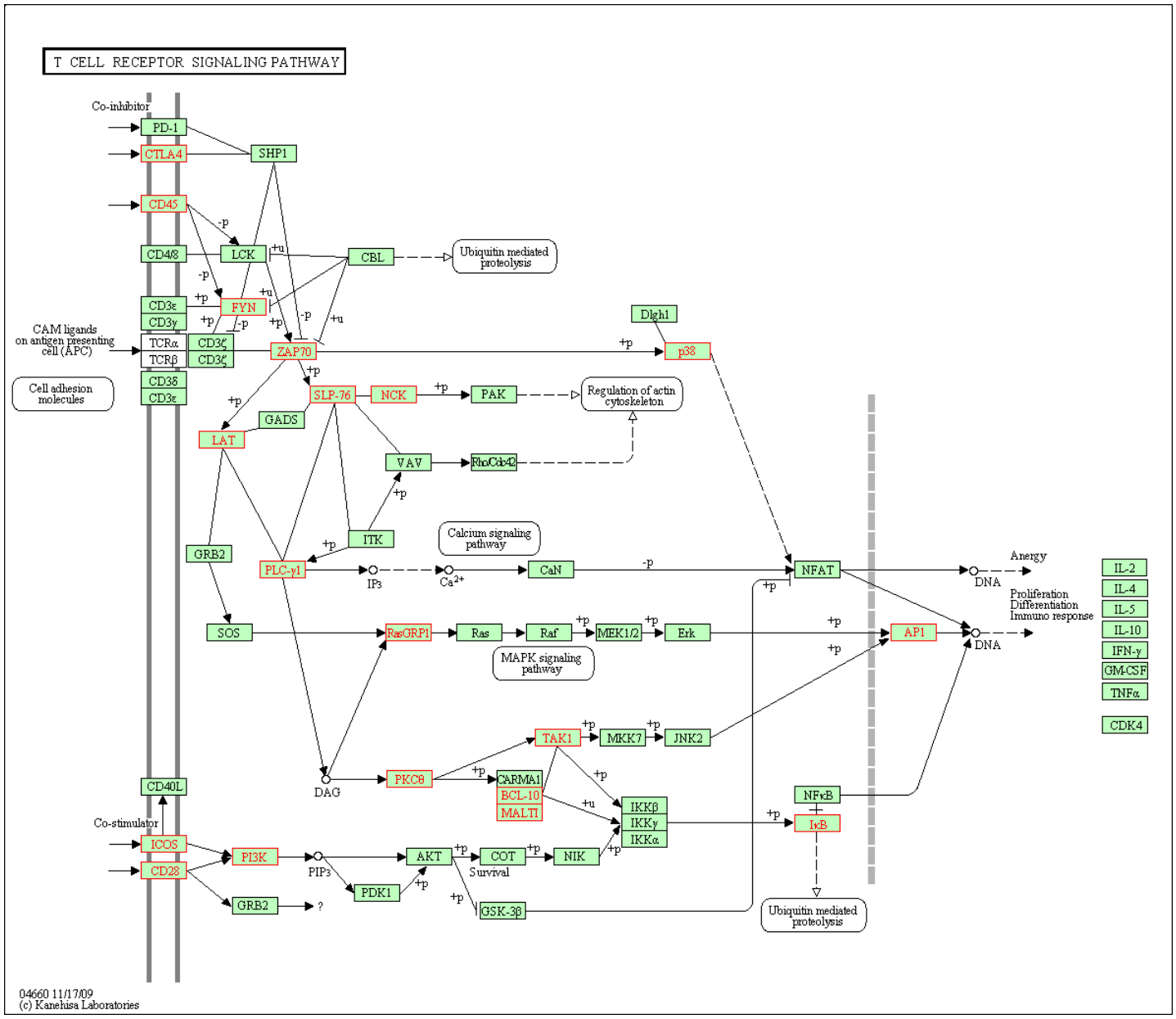


Figure S15: T cell receptor pathway illustration from the KEGG database (hsa04660). Genes located proximal to MIR-insulators, on the active domain side, are highlighted in red.

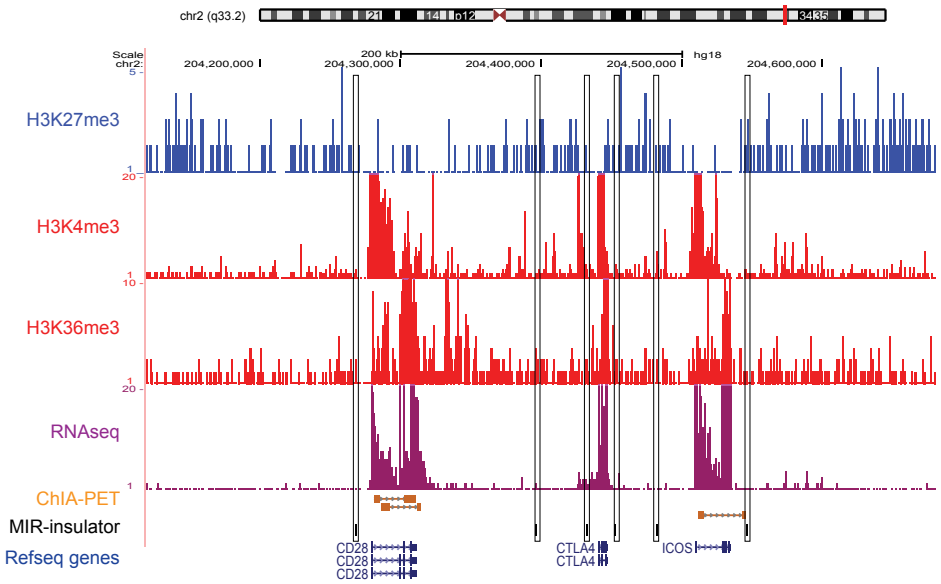


Figure S16: Chromatin environment around three TCR genes. The same genomic locus is shown as Figure 4D. ChIP-seq signals of three histone marks and RNAseq data (purple) are shown along with CD4+ T cell ChIA-PET interactions (orange). The three pairs of MIR-insulators are highlighted by boxes.

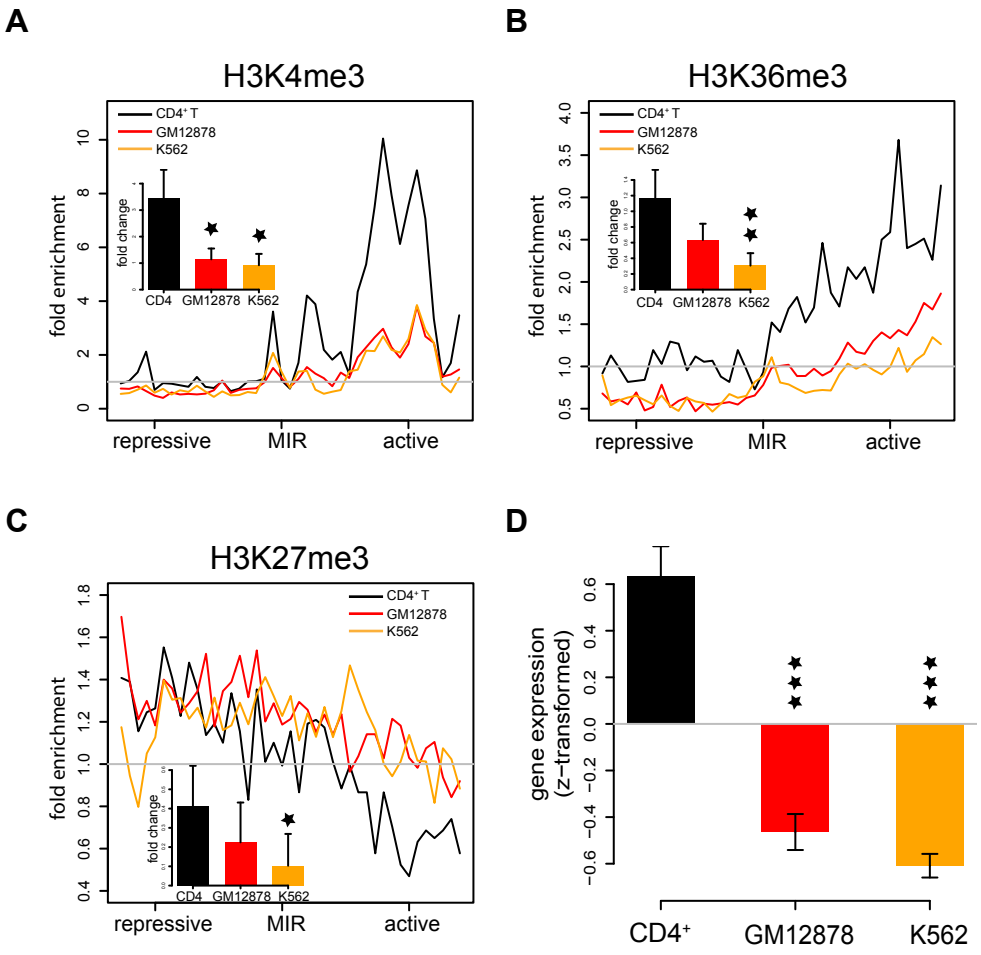


Figure S17: Cell type-specific chromatin barrier activity and gene regulation by MIR-insulators from the T cell receptor pathway. ChIP-seq fold enrichment levels around MIR-insulators proximal to the 21 T cell receptor genes are shown for H3K4me3, H3K36me3 and H3K27me3 in CD4+ T cells (black), GM12878 cells (red) and K562 (orange) cells. Insets show the average differences (± standard error) between the active versus repressive domains surrounding MIR-insulators for the marks and cells. Significance of the differences between CD4+ T cells and other cells are indicated as * P<0.05 ** P<0.01 *** P<0.001. Average gene expression levels (± standard error) are shown for genes located in the active domain side proximal to MIR-insulators at the 21 T cell receptor genes. Gene expression levels are z-transformed within each cell-type.

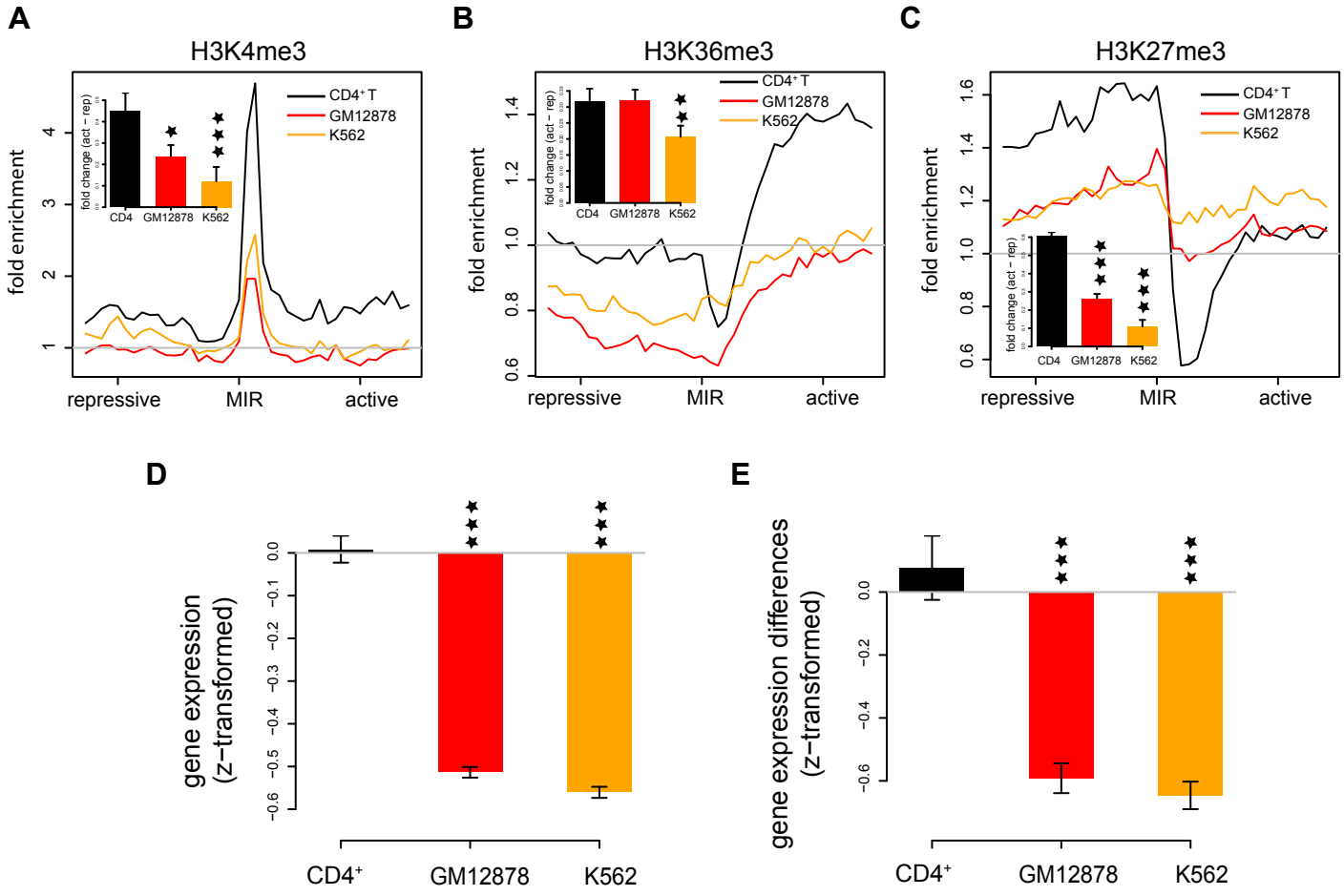


Figure S18: Cell-type specific chromatin barrier activity and gene regulation by CTCF barriers. ChIP-seq fold enrichment levels around tissue-specific CTCF barriers are shown for (A) H3K4me3, (B) H3K36me3 and (C) H3K27me3 in CD4+ T cells (black), GM12878 cells (red) and K562 (orange) cells. Insets show the average differences (± standard error) between the active versus repressive domains surrounding CTCF barriers for the marks and cells. (D) Average gene expression levels (± standard error) are shown for genes located in the active domain side proximal to CTCF barriers. Gene expression levels are z-transformed within each cell-type. (E) Average (± standard error) differences in the gene expression levels for genes located on the opposite sides of individual CTCF barriers. Gene expression difference values are z-transformed within each cell-type. For all bar plots, significance of the differences between CD4+ T cells and other cells are indicated as * P<0.05 ** P<0.01 *** P<0.001.

Table S1: Spearman correlations between upstream and downstream histone modification levels across putative MIR-insulators.

Histone modifications	Spearman correlations	*P* value
H2AK5ac	-0.23	2.3E-8
H2AK9ac	-0.79	3.1E-81
H2BK5ac	-0.39	1.2E-41
H2BK12ac	-0.47	2.7E-55
H2BK20ac	-0.24	3.7E-16
H2BK120ac	-0.36	5.9E-36
H3K4ac	-0.41	1.4E-44
H3K9ac	-0.55	1.3E-82
H3K14ac	-0.85	1.6E-21
H3K18ac	-0.20	3.1E-12
H3K23ac	-0.76	3.6E-61
H3K27ac	-0.37	2.3E-37
H3K36ac	-0.43	4.9E-49
H4K5ac	-0.32	1.7E-27
H4K8ac	-0.28	9.6E-21
H4K12ac	-0.70	8.5E-84
H4K16ac	-0.44	5.7E-45
H4K91ac	-0.33	6.8E-29
H2AZ	-0.18	10.0E-10
H2BK5me1	-0.36	3.7E-34
H3K4me1	-0.06	1.5E-2
H3K4me2	-0.38	4.1E-41
H3K4me3	-0.32	1.0E-29
H3K9me1	-0.41	5.0E-49
H3K9me2	-0.78	1.6E-37
H3K9me3	-0.63	6.5E-48
H3K27me1	-0.42	3.4E-47
H3K27me2	-0.72	7.2E-66
H3K27me3	-0.47	1.0E-31
H3K36me1	-0.70	1.0E-56
H3K36me3	-0.43	3.7E-52
H3K79me1	-0.54	3.2E-80
H3K79me2	-0.72	3.5E-146
H3K79me3	-0.69	3.4E-139
H3R2me1	-0.33	6.6E-20
H3R2me2	-0.72	1.2E-12
H4K20me1	-0.45	1.3E-55
H4K20me3	-0.64	1.2E-17
H4R3me2	-0.82	1.6E-29

Table S2: Genomic coordinates (hg18) of tested MIR-insulators and their corresponding primers for EBA validations.

ID	MIR element locations	Type	Coordinates of tested sequences	Size (bp)	Primer ID	Primer Sequences
MIR1	chr1:23555914-23556047	MIR	chr1:23555859-23556088	230	1	ATACACTCGAGATGCATGATATGGCCCAGTGATGGTC
					2	ATACACTCGAGATGCATAGTCATGCCCATACCACCTC
MIR2	chr2:97999554-97999807	MIR	chr2:97999495-97999868	374	3	ATACACTCGAGCTGCAGTGAACATAGGAGGGGAGGTG
					4	ATACACTCGAGCTGCAGAAGATGATCCACCCTGCAAT
MIR3	chr11:82289556-82289817	MIRb	chr11:82289550-82289843	294	5	ATACACTCGAGATGCATAACGGCAATAACAGCTACCA
					6	ATACACTCGAGATGCATTAGGGAGTGGTTAGGCTCCA