**Supporting Information Appendix**

# Chromatin "stretch enhancer" states drive cell-specific gene regulation and harbor human disease risk variants

Stephen C. J. Parker[1,5], Michael L. Stitzel[1,5], D. Leland Taylor[1], Jose Miguel Orozco[1], Michael R. Erdos[1], Jennifer A. Akiyama[2], Kelly Lammerts van Bueren[4], Peter S. Chines[1], Narisu Narisu[1], NISC Comparative Sequencing Program[1], Brian L. Black[4], Axel Visel[2,3], Len A. Pennacchio[2,3], and Francis S. Collins[1,6]

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
[2]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[3]DOE Joint Genome Institute, Walnut Creek, CA, USA
[4]Cardiovascular Research Institute, University of California, San Francisco, CA, USA
[5]These authors contributed equally
[6]Author to whom correspondence should be addressed:
[6]Author to whom correspondence should be addressed:
Francis S. Collins, MD, PhD
50 South Drive, Room 5154
Bethesda, MD 20892-2152
Tel: 301-496-2433
Fax: 301-496-2700

**Table of Contents**

## Methods

### *Human Islets*

Fresh human pancreatic islets were obtained from the Islet Cell Resource (ICR) and National Disease Research Interchange (NDRI) in accordance with all Human Subjects research regulations. Islet viability and purity were determined by the distribution sites and are shown with organ donor information in Table S2. Upon receipt, islets were warmed to 37 C in CMRL shipping medium for 1-2 hours. After equilibration, islets were washed with calcium- and magnesium-free Dulbecco's phosphate-buffered saline (Invitrogen, Carlsbad, CA) prior to harvesting for RNA preparation and crosslinking. Approximately 4,000 islet equivalents (~4 million cells) were collected by centrifugation at 200 x g. The resulting pellet was flash frozen in liquid nitrogen and stored at -80 C. 12,000-16,000 islet equivalents (12-16 million cells) from the same sample were crosslinked in 1% formaldehyde for 20 minutes at room temperature, flash frozen in liquid nitrogen, and stored at -80 C.

### *Chromatin immunoprecipitation, RNA isolation, and sequencing library preparation*

Crosslinked chromatin was prepared and immunoprecipitated as described in(1). ChIP grade Abcam mouse anti-H3K4me3 (ab12209; lot 738622), rabbit anti-H3K27ac (ab4729, lots 961080, GR45851-1), rabbit anti-H3K36me3 (ab9050; lots 446805, 572220), and rabbit anti-GFP non-specific ChIP control antibody (ab290, lot C0411) were used to generate new islet ChIP-seq data used in this study. Total RNA was extracted and purified using Trizol (Invitrogen, Carlsbad, CA). ChIP-seq and non-directional RNA-seq libraries were prepared and sequenced using standard Illumina protocols for GAII and Hi-Seq2000.

### *ChIP-seq and RNA-seq read mapping*

Sources of unmapped reads from ChIP-seq and RNA-seq experiments used in the integrative analysis are found in Table S3. Non-islet ChIP-seq reads came from Ernst et al. (2) and non-islet RNA-seq reads came from Djebali et al. (3). Islet reads from multiple individuals' islets obtained in our previous study (H3K4me3, H3K4me1, and CTCF (1)), NIH Epigenome Roadmap (http://www.roadmapepigenomics.org/) (H3K27me3), or generated in this study (H3K36me3, H3K72ac) were merged to create a consensus human islet chromatin state map. Paired-end 100 bp islet H3K27ac ChIP-seq reads obtained in this study were truncated to single-end 36 bp fragments to match the read length of ENCODE and previous islet data sets and prevent artifacts caused by differential mappability. To ensure we used high-quality data, only reads that pass the Illumina chastity filter were considered. All ChIP-seq data was mapped using BWA (4) (version 0.5.8c) with default parameters to the hg19 version of the human genome. Non-directional RNA-seq reads from four islet samples (Table S3) were also compiled to create a reference islet transcriptome. Islet RNA-seq reads were trimmed to 76 bp paired-end fragments to match the read length of ENCODE data sets and prevent mappability issues. All

RNA-seq data was mapped using STAR (5) (version 2.1.1a_r109) with default parameters to the hg19 version of the human genome. Duplicate reads were removed using samtools (6) (version 0.1.18).

### *ChromHMM joint state learning across the ten cell types*

Chromatin states were learned jointly by applying the ChromHMM (version 1.02) hidden Markov model (HMM) algorithm at 200 bp resolution to seven data tracks (Input, CTCF, K27ac, K27me3, K36me3, K4me1, K4me3) from each of the ten cell types, as previously described (2). We ran ChromHMM with a range of possible states, and settled on a 12 state model as it accurately captured information from higher state models and provided sufficient resolution to identify biologically meaningful patterns in a reproducible way. To determine how our learned states relate to previously published states from the 9 cell types (2), we performed enrichment analyses comparing our states to the published states in each cell type (Fig. S1). We also performed gene body feature overlaps, and TSS proximity analyses, as previously described (2). This process led to a clear state assignment (Fig. S1), which we used for all subsequent analyses.

For the subsampled read segmentation we randomly selected reads so that each data track is equally represented across all cell types. This effectively normalizes all cell types to the lowest sampled cell type. We then repeated the state calls using the same model we learned with the full data sets. We note that the trends reported herein are consistently observed even when normalized read chromatin states are used.

We consider enhancers as any contiguous genomic region in a cell type marked by states 4-7 (enhancer states). To estimate the null expectation for the enhancer length distribution, we generated ten random chromatin segmentations using the transition parameters from the learned HMM. Of note, contiguous enhancer segments from the random model are not biased by unmappable regions like the observed data.

Our random expectation is conservative with respect to enhancer lengths, which leads to an underestimate in the size distribution shift when comparing the real enhancers to the null set. Although this trend is clear, the difference does not reach statistical significance. We suspect that read mappability differences contribute to decreased enhancer state sizes in the observed data and inflated enhancer state sizes in the null distribution. Specifically, large enhancers may be split by poorly mapping regions into smaller enhancers in the real data. In the best case scenario, 36-bp reads used for the ChIP-seq analyses can uniquely map to 80% of the reference genome (7). The null distribution is not affected by mappability, so large enhancers will not be broken up into smaller enhancers in the same manner. Therefore, the shift in enhancer size we report is a conservative estimate of the real shift.

### *Gene expression quantification*

We calculated gene expression in reads per kilobase per million mapped reads (RPKM). In order to only quantify uniquely mapping reads, we filtered all RNA-seq data for primary mapped reads with a mapping quality score of 255. Using ensembl 68 annotations, we counted the number of reads that overlapped each feature with HTSeq-count (http://www-

huber.embl.de/users/anders/HTSeq/), changing the default parameters to "--stranded=no." Using the default parameters for non-quantile normalization, we calculated standard RPKMs (not quantile normalized RPKMs, though we note this does not change our results) with the Conditional Quantile Normalization (CQN) R library (8), providing the length and read counts per gene as well as the total counts for each sample (including the counts of reads that did not overlap any feature). In order to calculate the expression of genes surrounding enhancers, we filtered the resulting RPKMs for protein-coding genes that were also included in the Gene Ontology Database in order to maintain consistency with GO analyses (Figs. S12, S13).

For browser display purposes we created RNA-seq signal density tracks by calculating RPM/bp across the genome and displaying these tracks on a scale from zero to two.

To define housekeeping and cell-specific genes, we first filter for genes expressed at RPKM > 3 in any cell type. Next we calculate normalized information content, as previously described (9)**,** using RPKMs across the ten cell types. We consider cell-specific genes as those having a normalized information content greater than 0.75 and housekeeping genes as those having a normalized information content less than 0.25. We note that similar results are observed when using different RPKM and normalized information content thresholds. We included the RPKM values for all ensembl 68 protein coding genes for all cell types in dataset S1.

All processed ChIP-seq and RNA-seq results are browsable and downloadable at: http://research.nhgri.nih.gov/manuscripts/Collins/islet_chromatin/

### *GWAS variant enrichment in enhancer states*

We filtered the NHGRI GWAS catalog (http://www.genome.gov/gwastudies/;downloaded on December 10, 2012) for genome-wide significant ($P<5x10^{-8}$) SNPs and then collapsed all SNPs for a given trait into one unique set. SNPs in linkage disequilibrium (LD) with the lead SNP were defined as those with $r^2 \geq 0.8$. To find LD SNPs we used 1000 Genomes SNPs in the CEU population (http://www.1000genomes.org/; phase 1, v3).

To calculate enrichment we performed a permutation test that measures SNP and enhancer overlaps as previously described (2), except that we include LD SNPs. We run 10,000 iterations of the permutation test and estimate the maximal P-value as the number of permutations equal to or greater than the observed overlap value plus one divided by the number of iterations plus one (10,001). Overlap enrichment values were calculated as the observed overlap count divided by the mean of all permutations. Notably, our enrichment strategy does not shuffle enhancer territory, which preserves enhancer length distributions. Thus, there should be no enhancer length-associated bias. Finally, because analyses are performed simultaneously for all 10 cell types, the other 9 cell types serve as internal (negative) controls. Furthermore, all enrichment analyses are performed across all ten cell types so that any trend observed in a single cell type is controlled by the other nine cell types.

### *Cohesin complex component RAD21 enrichment in enhancer states*

We downloaded ENCODE ChIP-seq data for the cohesin complex factor RAD21. To focus on enhancer associated regions, we first removed any RAD21 peaks that also overlap CTCF sites

in the same cell type. Next, we compared overlap enrichment across cell types using enhancers less than or equal to the median (0.8 kb) size, enhancers greater than or equal to the stretch enhancer threshold of 3 kb (90$^{th}$ percentile), enhancers greater than or equal to 4.2 kb (95$^{th}$ percentile), and enhancers greater than or equal to 6.2 kb (99$^{th}$ percentile). Non-CTCF RAD21 peak enrichment relative to enhancers was calculated in a similar manner as GWAS SNP enrichment (see above). We, again, used 10,000 randomizations to estimate the null expectation. These results are reported in Supplementary Figure 11.

### *GO term enrichment analyses*

We calculated GO enrichments (GO database: http://www.geneontology.org/; accessed December 20, 2012) by assigning enhancers to genes whose TSS, based on all known isoforms, is within 125 kb. We discarded enhancers that could not be assigned to a gene using the defined criteria. We used GO::TermFinder (10) (v.0.86) to calculate GO term enrichment in genes linked to enhancers greater than or equal to any given length threshold (0.2 kb to 6.2 kb in 0.2 kb increments) and reported the Bonferroni corrected *P*-value (Fig. 3A). We calculated the information content of the top 10 most enriched, significant (Bonferroni corrected $P < 0.05$) GO terms across all enhancers in each cell type using the GOSim (11) (v.1.2.7.7) R library. We use the calculated information content as a GO term specificity score. As a null model, we shuffled the genomic coordinates of enhancers 100 times along the same chromosome. For each shuffle, we assigned enhancers to nearby genes and calculated significantly enriched (Bonferroni corrected $P < 0.05$) GO terms. We computed the specificity scores for the top 10 most enriched, significant GO terms for each cell type across all shuffles and enhancer lengths (Fig. S13). We compared the distribution of the specificity scores between shuffled and observed data sets for enhancers ≥ 3 kb in length using a Wilcoxon rank sum test. We also compared the specificity score distribution of GO terms associated with enhancers ≤ 0.8 kb in length (median enhancer length) to enhancers ≥ 3 kb in length using a Wilcoxon rank sum test.

### *Enhancer sub-clustering*

In order identify cell type specific and ubiquitous enhancers, we sub-clustered all 200 bp windows identified as an enhancer in any cell type based on the ChromHMM-defined posterior probability of being in any enhancer state across all ten cell types using *k*-means clustering, as previously described (2). We tried clustering solutions with 10-30 different clusters and found that 20 clusters was a good fit for the data based on minimizing the variance across clusters.

### *Stretch enhancer sequence cloning*

Genomic DNA from islet sample ULI102 (Table S2) or from K562 cells, respectively, was used to amplify randomly selected sequences underlying islet-specific (cluster 17) and K562-specific (cluster 19) stretch enhancers and cloned into Gateway-modified luciferase (GW-pGL4.23) and *hsp68-lacZ* reporter plasmids using the Gateway cloning system (Invitrogen, Carlsbad, CA). Primers were designed using PrimerTile (http://research.nhgri.nih.gov/tools). Amplicons and primer information is located in Table S4.

### *Transfection and Dual Luciferase Assays*

Cells were seeded in 96-well plates (MIN6: 60,000 cells/well; K562: 25,000 cells/well) and cotransfected with 0.072 pmol of firefly test plasmid (GW-pGL4.23; Promega, Madison, WI) and 2 ng of pRL-TK renilla luciferase control plasmid using Lipofectamine 2000 (MIN6) or Lipofectamine LTX (K562) (Invitrogen). Two plasmid preparations for each insert orientation were tested. Transfections were performed in triplicate. Cells were lysed in 1x Passive Lysis Buffer (Promega) 30-36 hours after transfection. Luciferase activity was determined using the Dual Luciferase Reporter kit (Promega; Madison, WI) and measured on a Centro/Centro XS3 Microplate Luminometer LB 960 (Berthold; Bad Wildbad, Germany). Firefly measurements were normalized to Renilla measurements to control for variation in cell number or transfection efficiency.

### *Transgenic mouse enhancer assay*

The enhancer activity of an intragenic (*ABCC8*) and a gene desert intergenic (Islet 32) islet stretch enhancer were tested for *in vivo* spatiotemporal activity by Gateway cloning into the *Hsp68-promoter-lacZ* reporter vector as previously described (12, 13). Genomic coordinates for these regions are found in **Table S4.** Transgenic mouse embryos were generated as previously described (13). Patterns observed in a minimum of three different embryos resulting from independent transgenic integration events of the same construct were considered reproducible (14). For histological analyses, X-gal stained embryos were embedded in paraffin and 12 micrometer transverse sections were counterstained using Neutral Fast Red (15). All animal work was performed in accordance with protocols reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee.

## References


1. Stitzel ML et al. (2010) Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab* 12:443–455.

2. Ernst J et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473:43–49.

3. Djebali S et al. (2012) Landscape of transcription in human cells. *Nature* 489:101–108.

4. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl* 26:589–595.

5. Dobin A et al. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinforma Oxf Engl* 29:15–21.

6. Li H et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma Oxf Engl* 25:2078–2079.

7. Derrien T et al. (2012) Fast Computation and Applications of Genome Mappability. *PLoS ONE* 7:e30377.

8. Hansen KD, Irizarry RA, Wu Z (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostat Oxf Engl* 13:204–216.

9. Greenbaum JA, Parker SCJ, Tullius TD (2007) Detection of DNA structural motifs in functional genomic elements. *Genome Res* 17:940–946.

10. Boyle EI et al. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinforma Oxf Engl* 20:3710–3715.

11. Fröhlich H, Speer N, Poustka A, Beissbarth T (2007) GOSim--an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics* 8:166.

12. Kothary R et al. (1988) A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube. *Nature* 335:435–437.

13. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. *Science* 302:413.

14. Visel A, Minovitsky S, Dubchak I, Pennacchio LA (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* 35:D88–92.

15. Dodou E, Verzi MP, Anderson JP, Xu S-M, Black BL (2004) Mef2c is a direct transcriptional target of ISL1 and GATA factors in the anterior heart field during mouse embryonic development. *Dev Camb Engl* 131:3931–3942.

16. Mutskov V, Felsenfeld G (2009) The human insulin gene is part of a large open chromatin domain specific for human islets. *Proc Natl Acad Sci U S A* 106:17419–17424.

17. Sanyal A, Lajoie BR, Jain G, Dekker J (2012) The long-range interaction landscape of gene promoters. *Nature* 489:109–113.

18. Higgs DR et al. (1990) A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes Dev* 4:1588–1601.

19. Aronow BJ et al. (1992) Functional analysis of the human adenosine deaminase gene thymic regulatory region and its ability to generate position-independent transgene expression. *Mol Cell Biol* 12:4170–4185.

20. Dang Q et al. (1995) Structure of the hepatic control region of the human apolipoprotein E/C-I gene locus. *J Biol Chem* 270:22577–22585.

21. Camper SA, Godbout R, Tilghman SM (1989) The developmental regulation of albumin and alpha-fetoprotein gene expression. *Prog Nucleic Acid Res Mol Biol* 36:131–143.

22. Kalos M, Fournier RE (1995) Position-independent transgene expression mediated by boundary elements from the apolipoprotein B chromatin domain. *Mol Cell Biol* 15:198–207.

23. Neznanov NS, Oshima RG (1993) cis regulation of the keratin 18 gene in transgenic mice. *Mol Cell Biol* 13:1815–1823.

24. Tam JLY et al. (2006) The human desmin locus: gene organization and LCR-mediated transcriptional control. *Genomics* 87:733–746.

25. Tuupanen S et al. (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* 41:885–890.

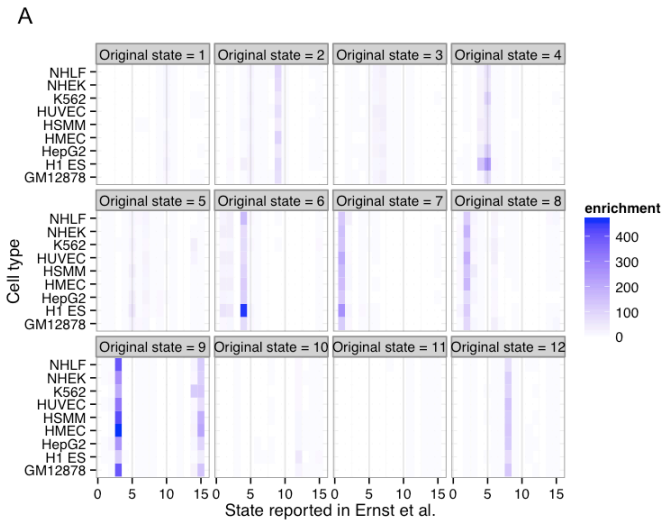
26. Sur IK et al. (2012) Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* 338:1360–1363.

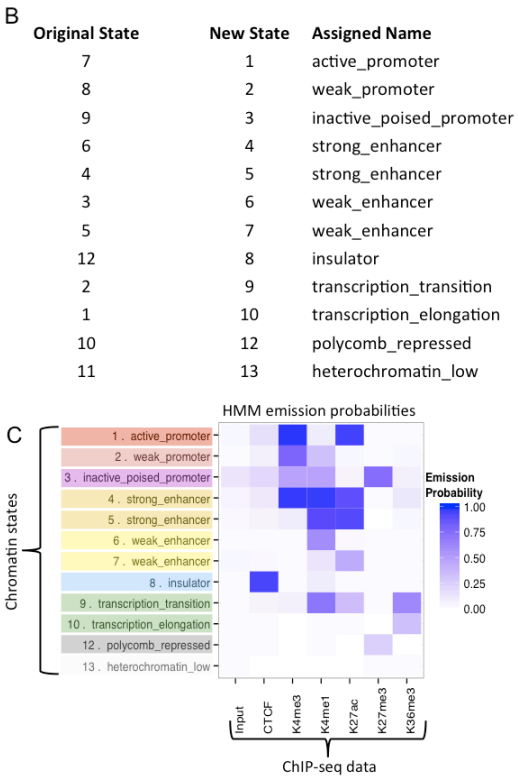
27. Musunuru K et al. (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466:714–719.

28. Grosveld F, van Assendelft GB, Greaves DR, Kollias G (1987) Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* 51:975–985.

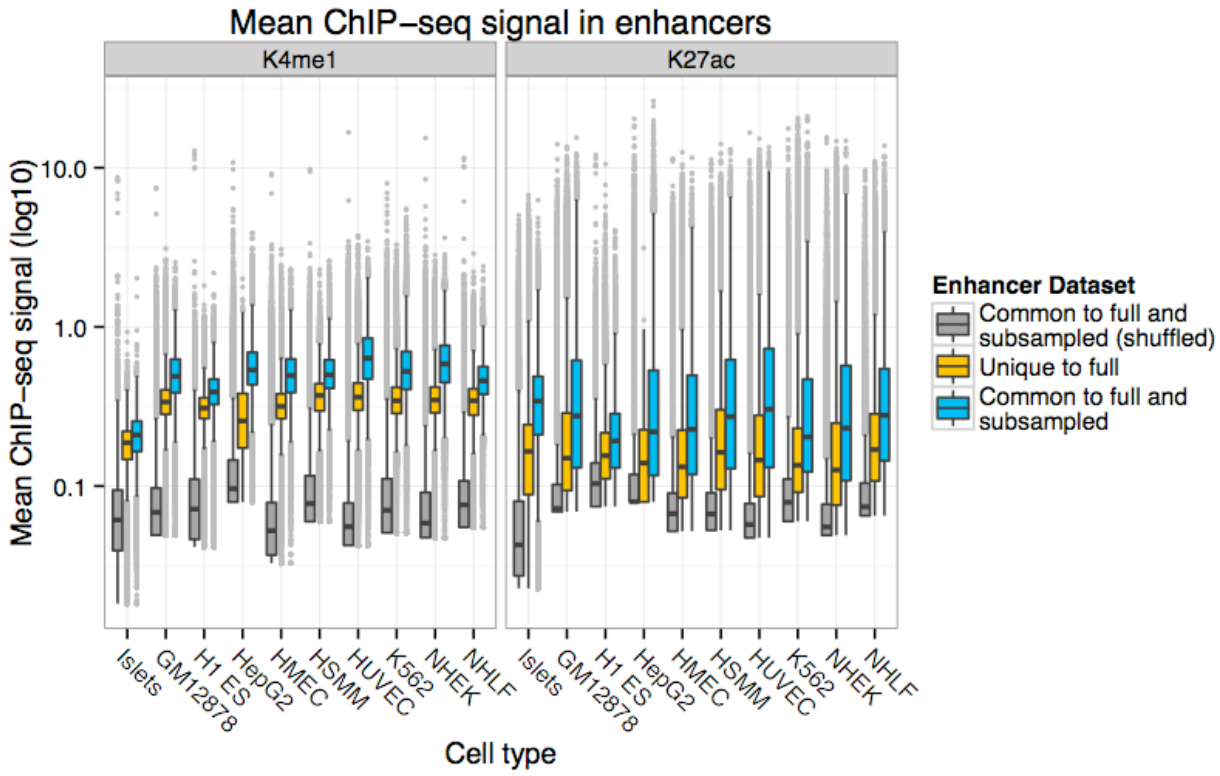
29. Li G et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148:84–98.

## Supplementary figures

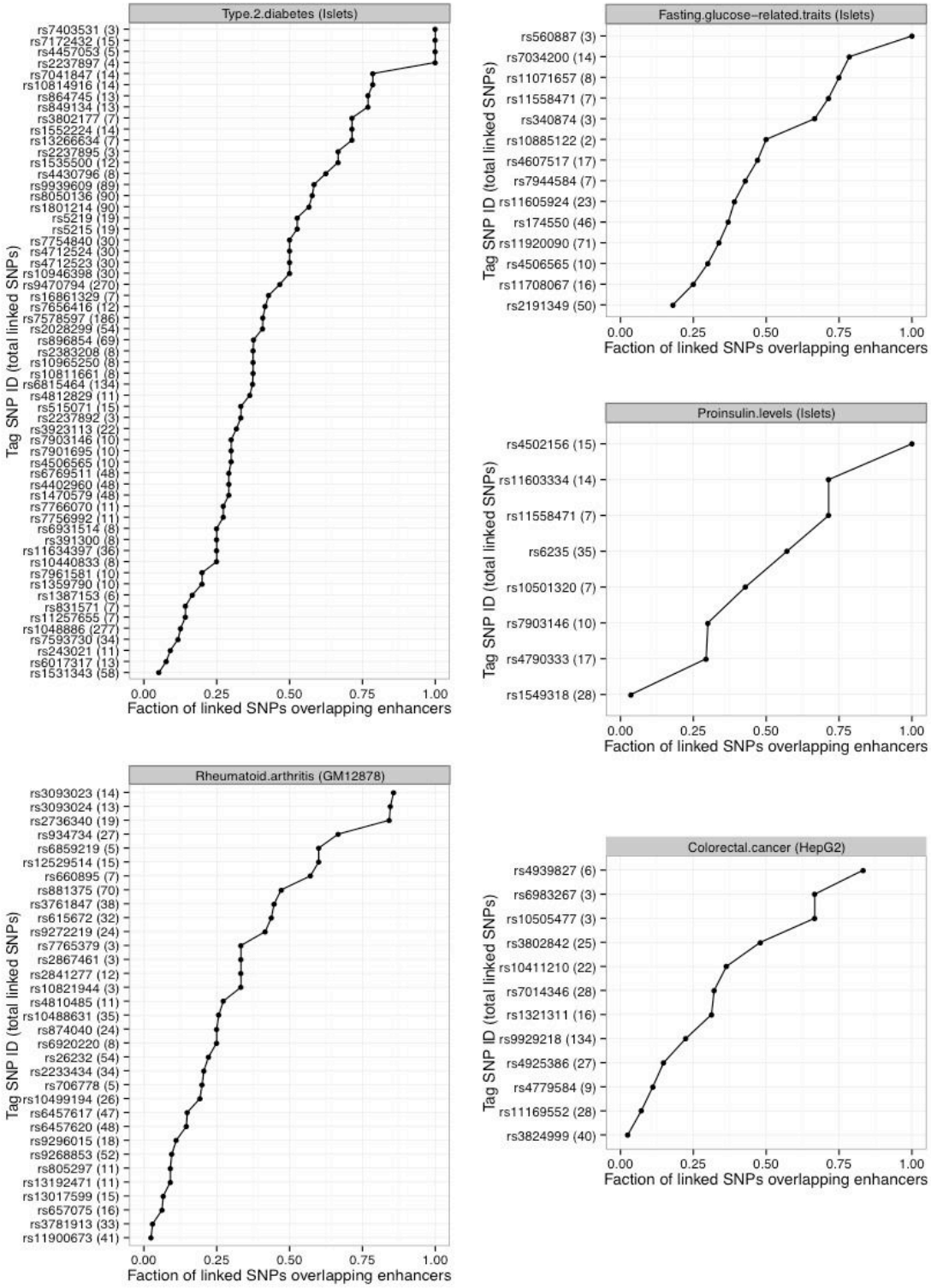


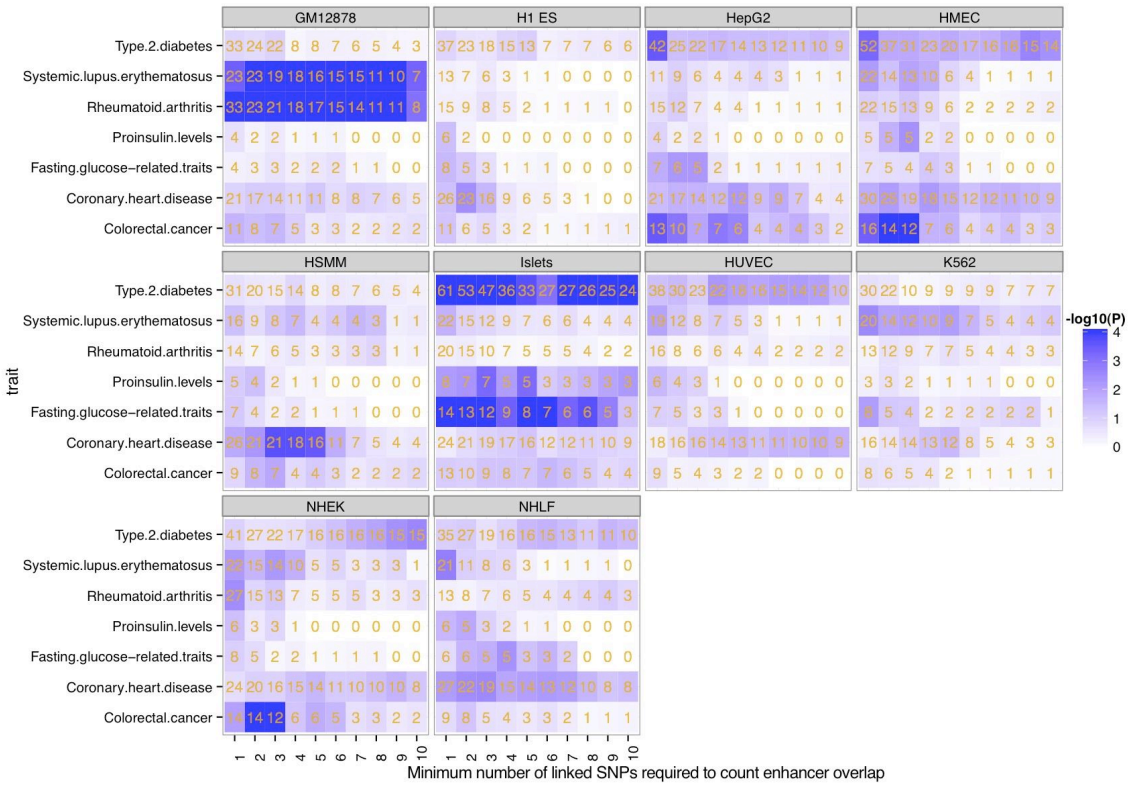

*Supplementary Figure 1*. Original state assignments from our ChromHMM run were compared to chromatin state assignments in the same cell types published by Ernst et al. (2) (A). Based on enrichment with previously published chromatin state calls, we re-assigned our states (B). The emission probabilities of our re-assigned chromatin state model (C).



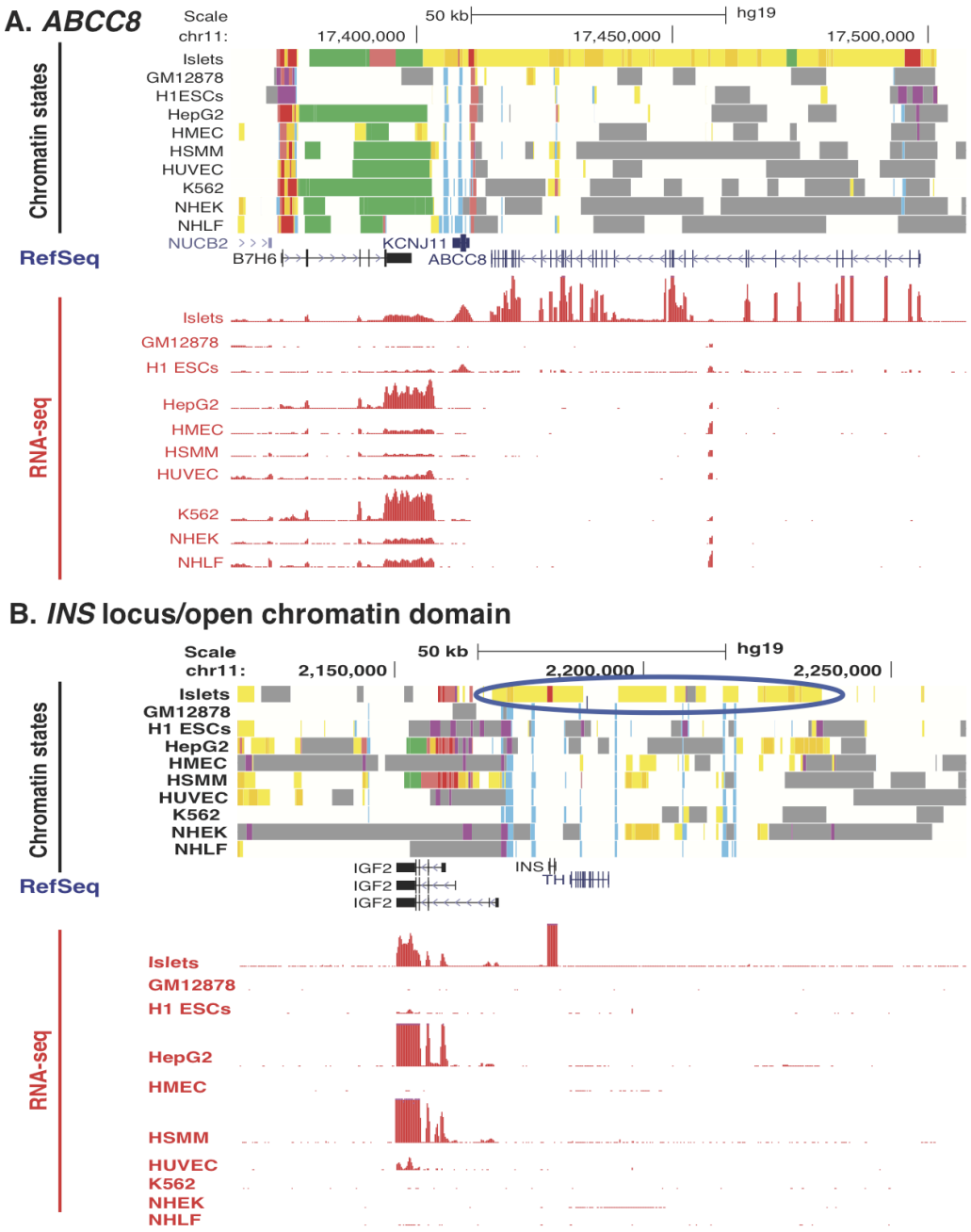

*Supplementary Figure 2*. Mean ChIP-seq signal for enhancers identified in normalized-read and full-read chromatin segmentations. Signal is calculated by extending reads 200 bp and calculating the mean reads per million mapped reads (RPM) per bp. This procedure effectively normalizes each ChIP-seq experiment to the number of mapped reads. We use the mean RPM/bp over an enhancer as the signal for that enhancer. Enhancers uniquely identified using the full read segmentation are shown in yellow. Enhancers common to full and normalized read segmentations are shown in blue. Control enhancers were generated by randomly shuffling the common enhancers along the same chromosome and are shown in grey.



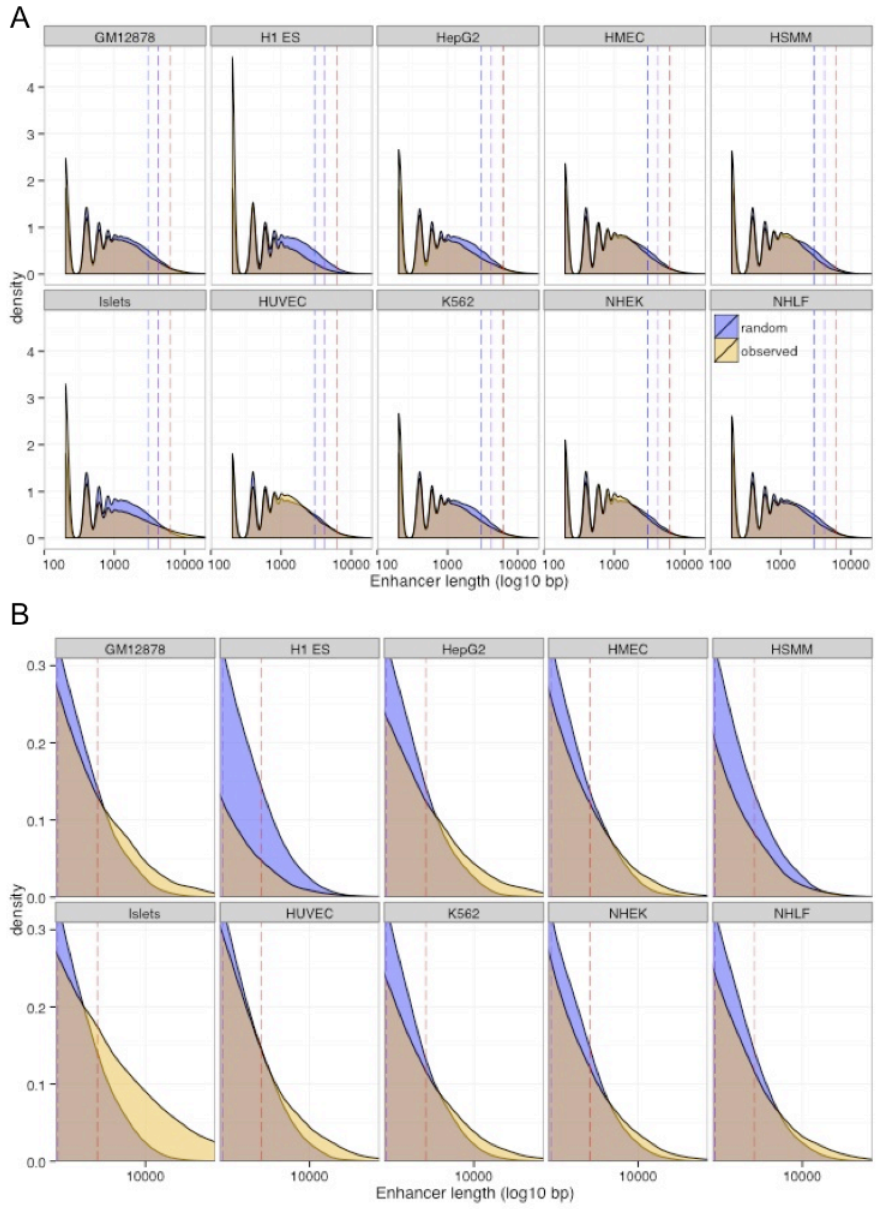

*Supplementary Figure 3*. Fraction of linked SNPs overlapping enhancers (in a given cell type).



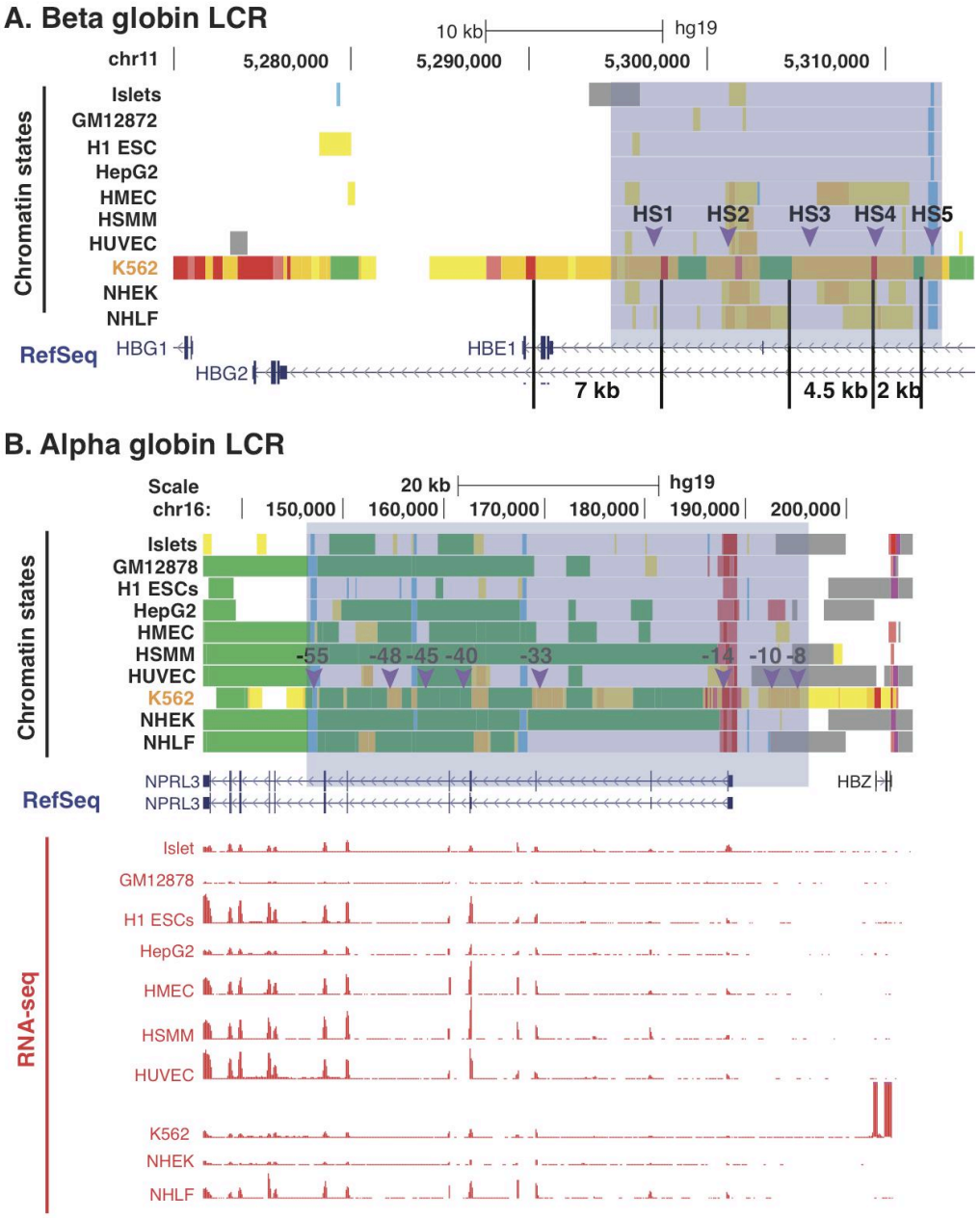

*Supplementary Figure 4*. Enrichment for multiple LD SNPs to overlap enhancers. blue coloration corresponds to significance of the permutation overlap test whereby a minimum number of linked SNP (x-axis) enhancer overlaps are required to count the overlap. Orange numbers represent the total number of tag loci that meet this threshold. For example, there are 24 T2D loci that each contain at least 10 linked SNPs that all overlap islet enhancers (see top right cell in the “Islets” panel).




*Supplementary Figure 5*. UCSC Genome browser view of *ABCC8* (A) and *INS* (B) loci. Chromatin state colors are assigned as in Figure 1A. Locations of annotated genes are indicated in the RefSeq portion of the view. RNA-seq data (red) demonstrates the robust islet-specific expression of *ABCC8* (A) and *INS* (B) in the same islet samples used to define the chromatin states. The blue oval in panel (B) highlights the *INS/IGF2/TH* open chromatin domain (16) that contains multiple islet-specific stretch enhancers.

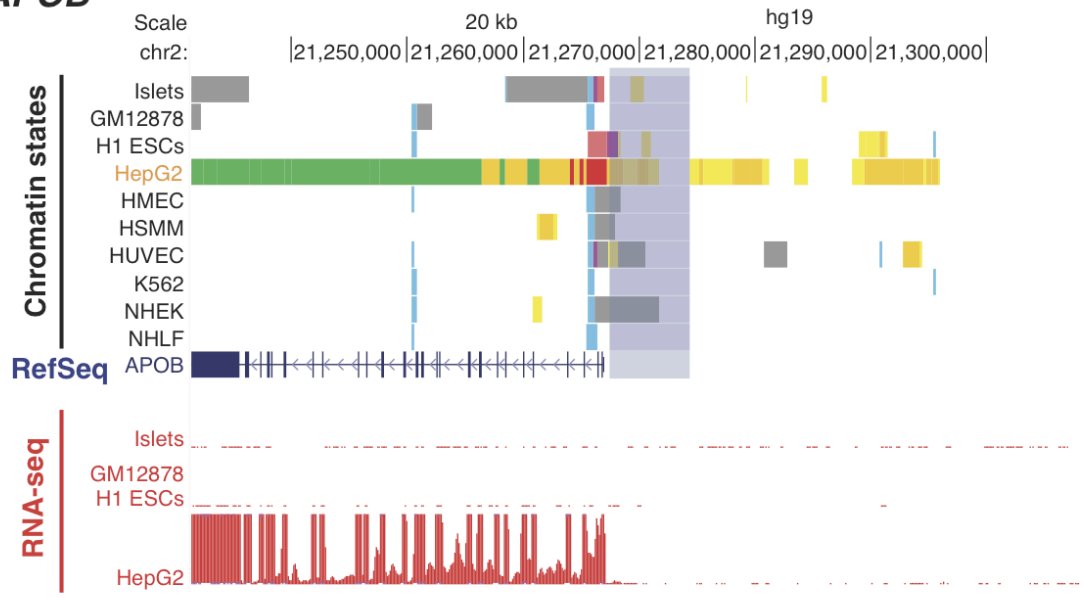



*Supplementary Figure 6*. Length distribution of enhancers for each cell type (A). Enhancer sizes for the 90th (blue dashed line), 95th (purple dashed line), and 99th (red dashed line) percentile of the random distribution are indicated for reference. A zoomed in view of the tail of the length distributions (B).

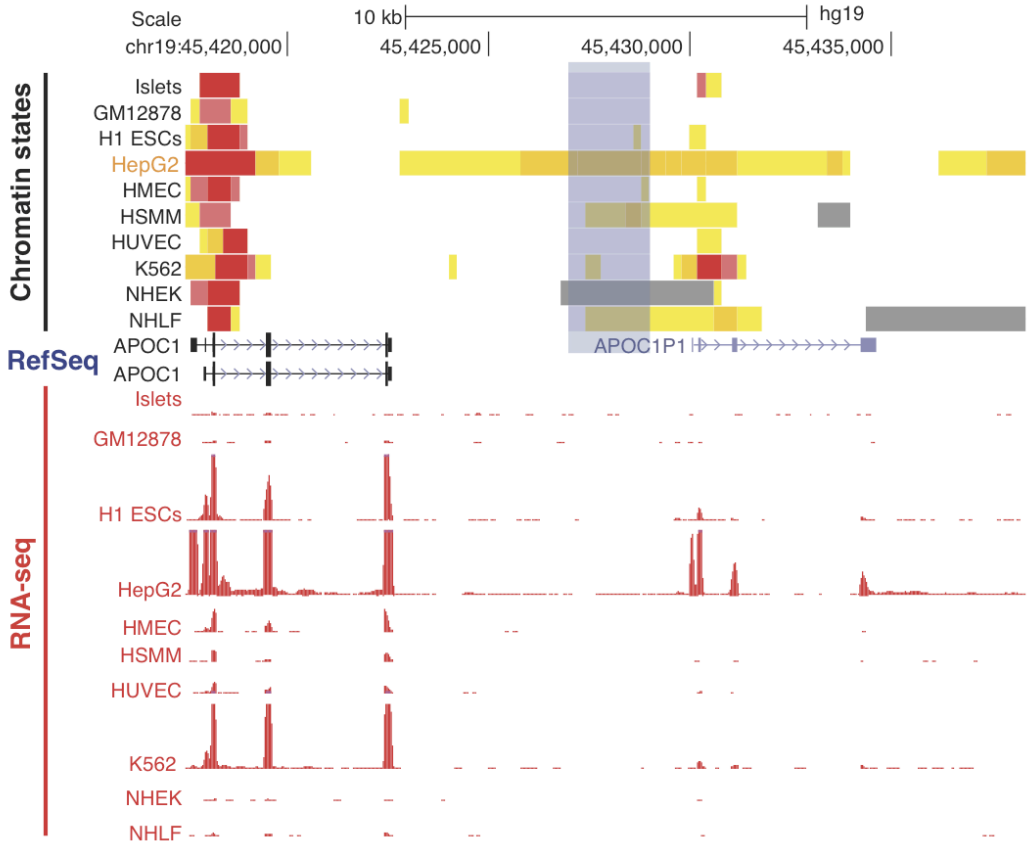

A. Beta globin LCR
10 kb
hg19
chr11
5,280,000
5,290,000
5,300,000
5,310,000
Chromatin states
Islets
GM12872
H1 ESC
HepG2
HMEC
HSMM
HUVEC
K562
NHEK
NHLF
HS1
HS2
HS3
HS4
HS5
RefSeq
HBG1
HBG2
HBE1
7 kb
4.5 kb
2 kb
B. Alpha globin LCR
Scale
20 kb
hg19
chr16:
150,000
160,000
170,000
180,000
190,000
200,000
Chromatin states
Islets
GM12878
H1 ESCs
HepG2
HMEC
HSMM
HUVEC
K562
NHEK
NHLF
-55
-48
-45
-40
-33
-14
-10
-8
RefSeq
NPRL3
NPRL3
HBZ
RNA-seq
Islet
GM12878
H1 ESCs
HepG2
HMEC
HSMM
HUVEC
K562
NHEK
NHLF

## C. Thymic regulatory region (*ADA*)

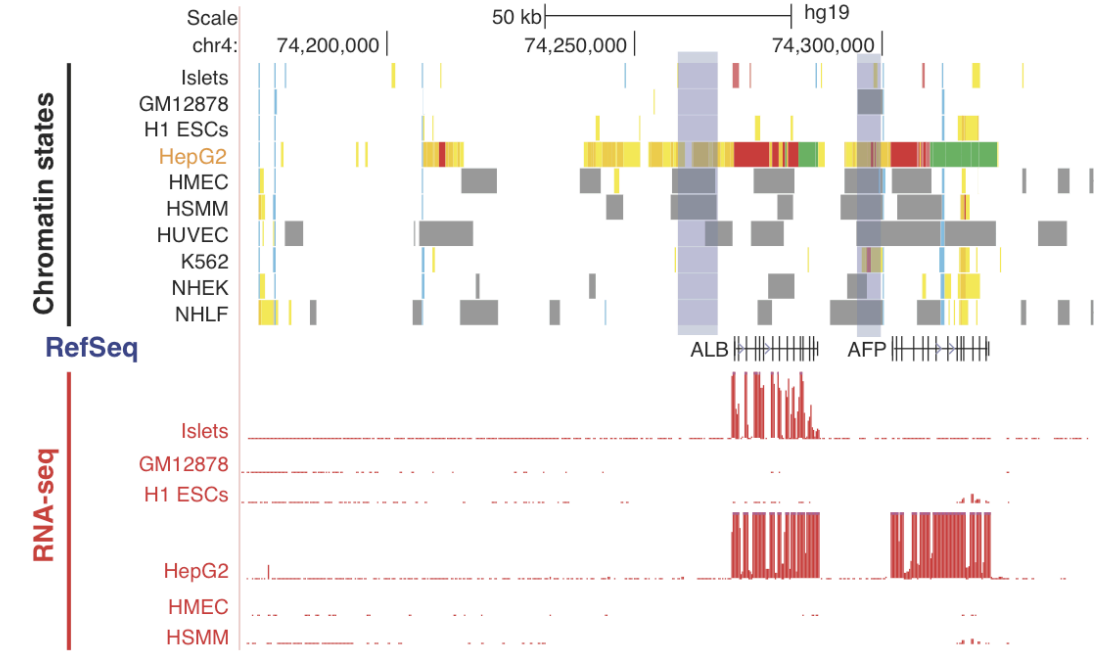

## D. *APOB*



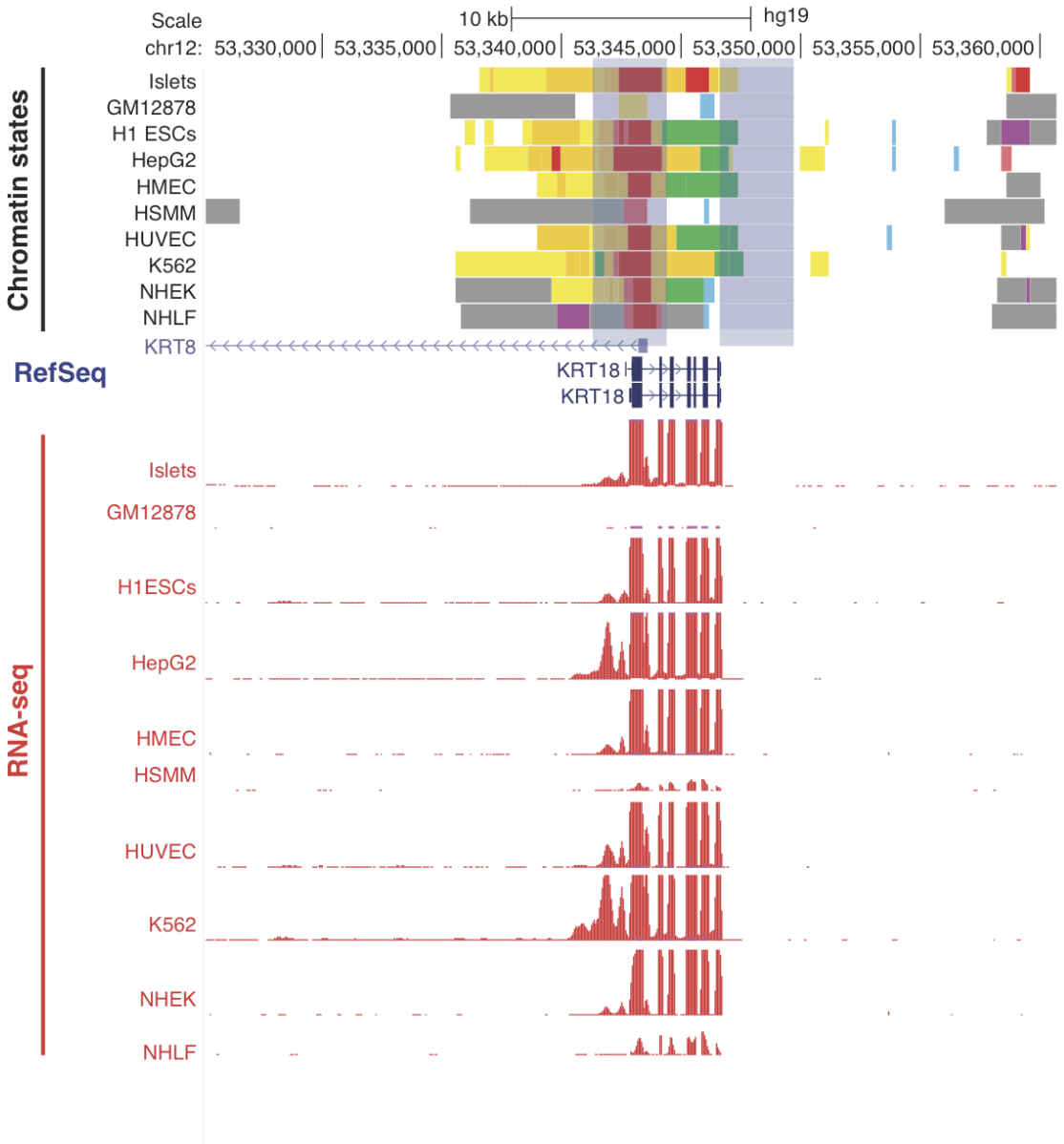
## E. Hepatic control region (*APOE/C1*)



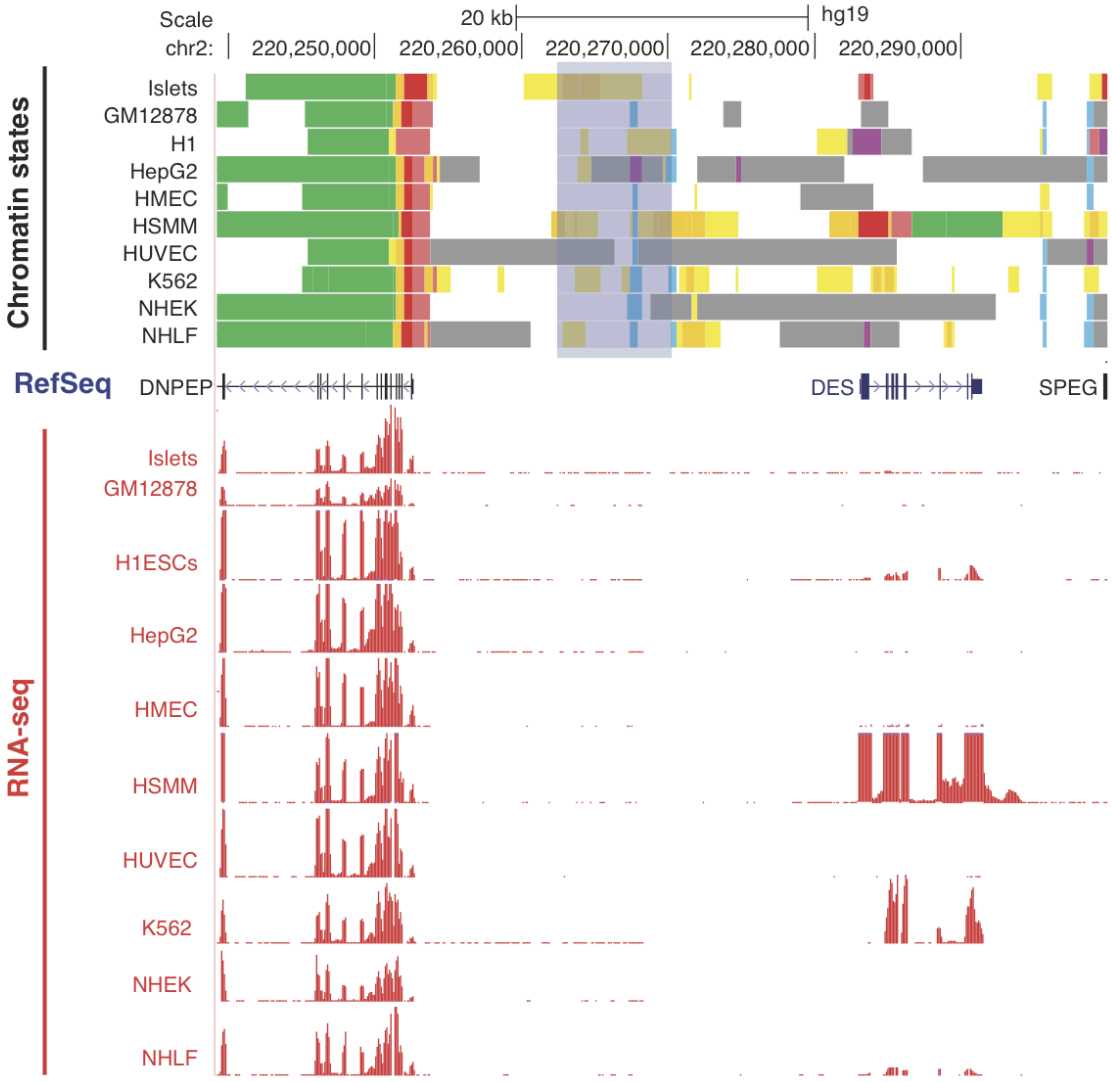

## F. *ALB/AFP* LCR



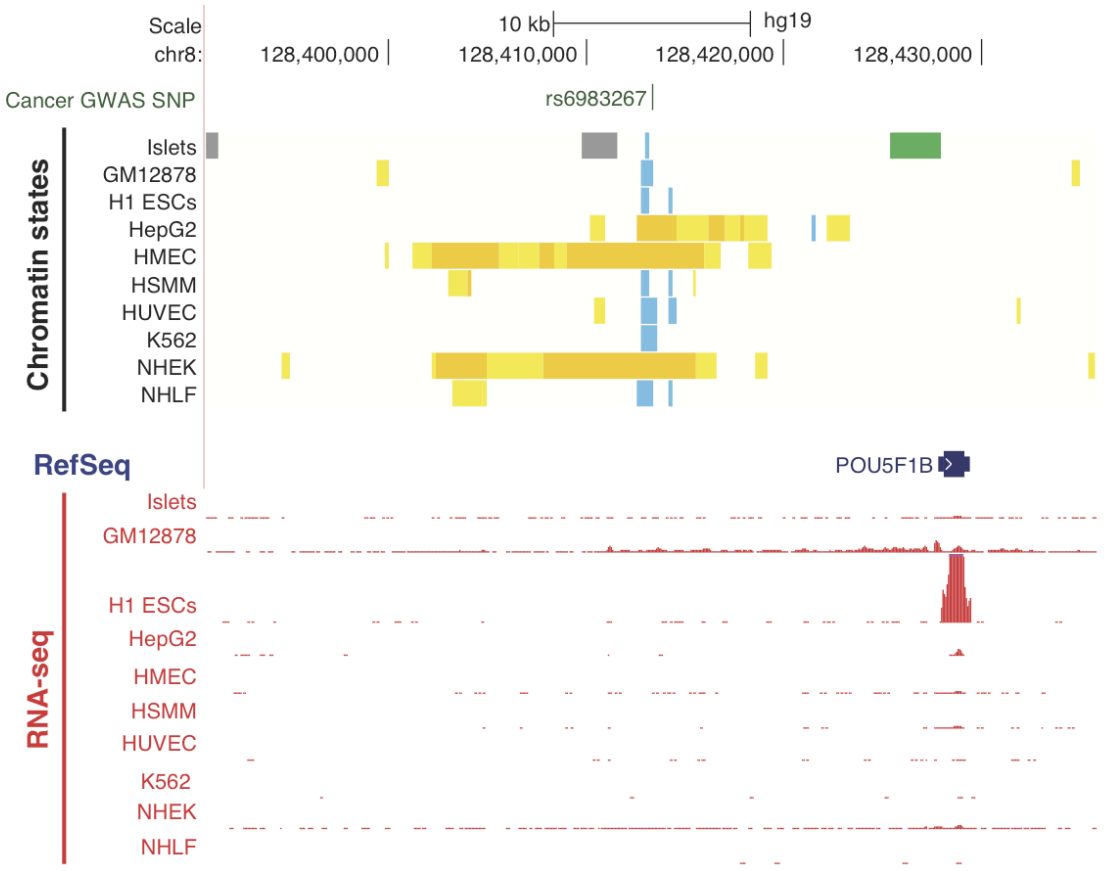
## G. Keratin LCR (*KRT18*)

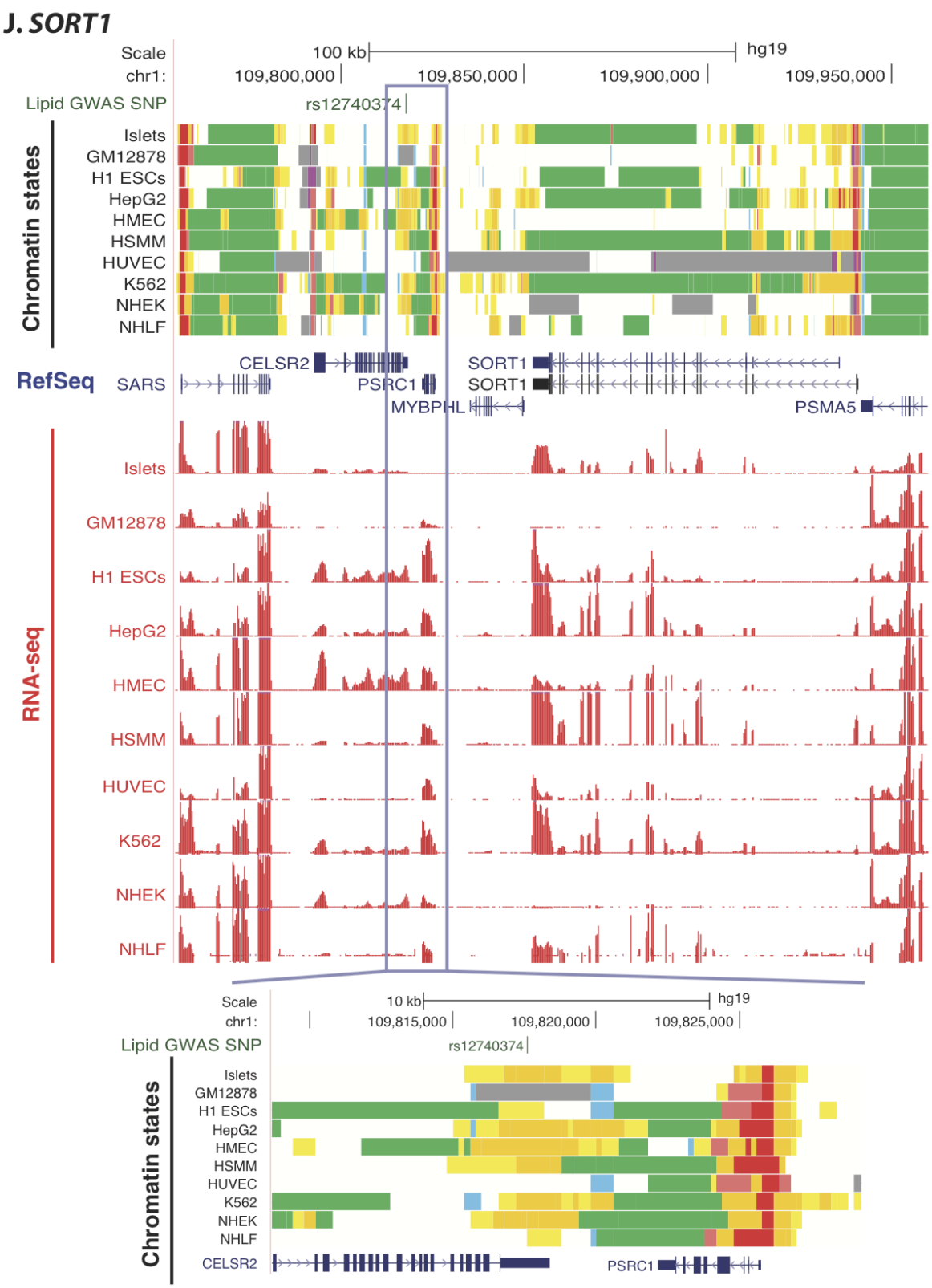


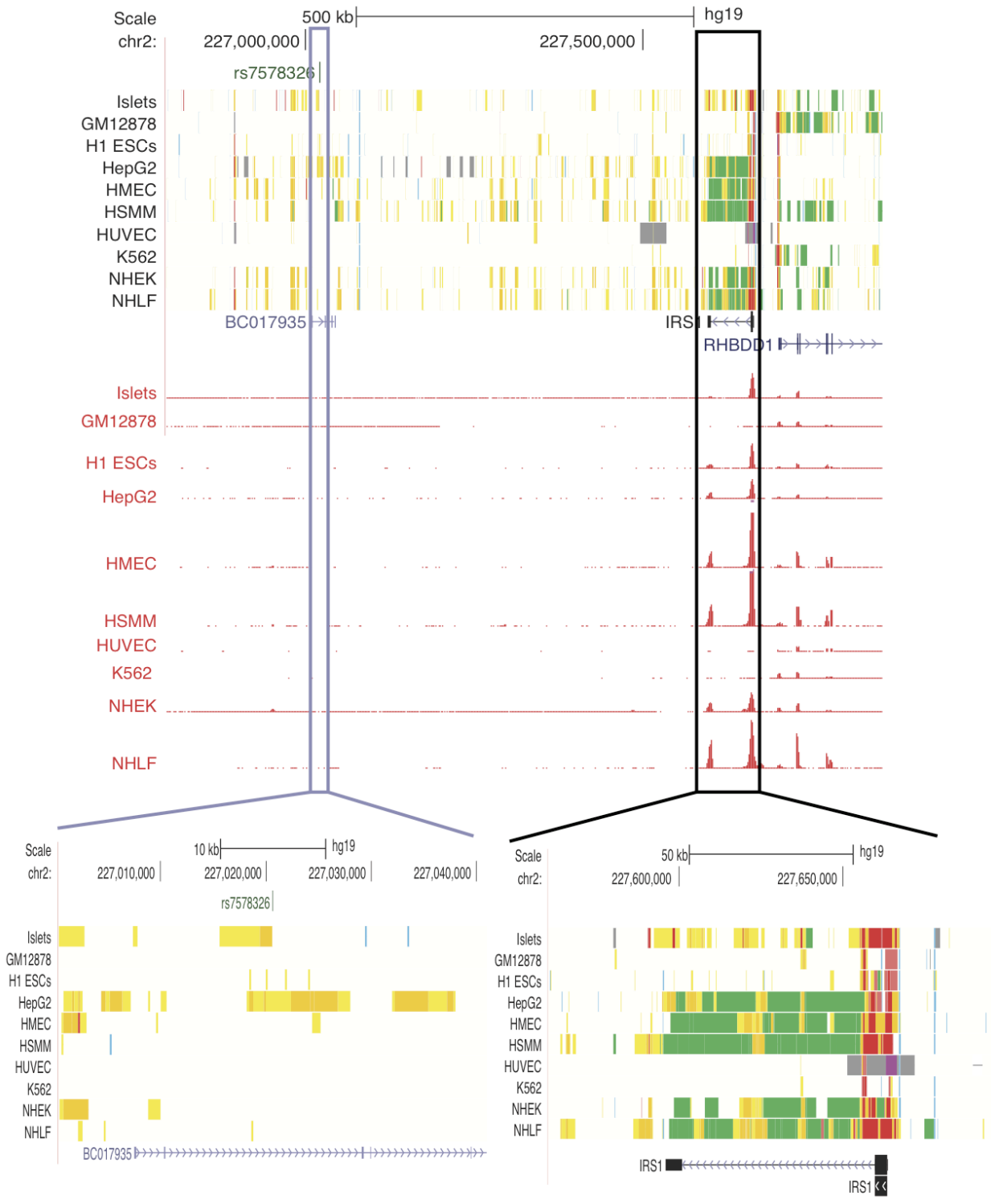
## H. Desmin LCR (*DES*)



## I. *MYC* enhancer SNP

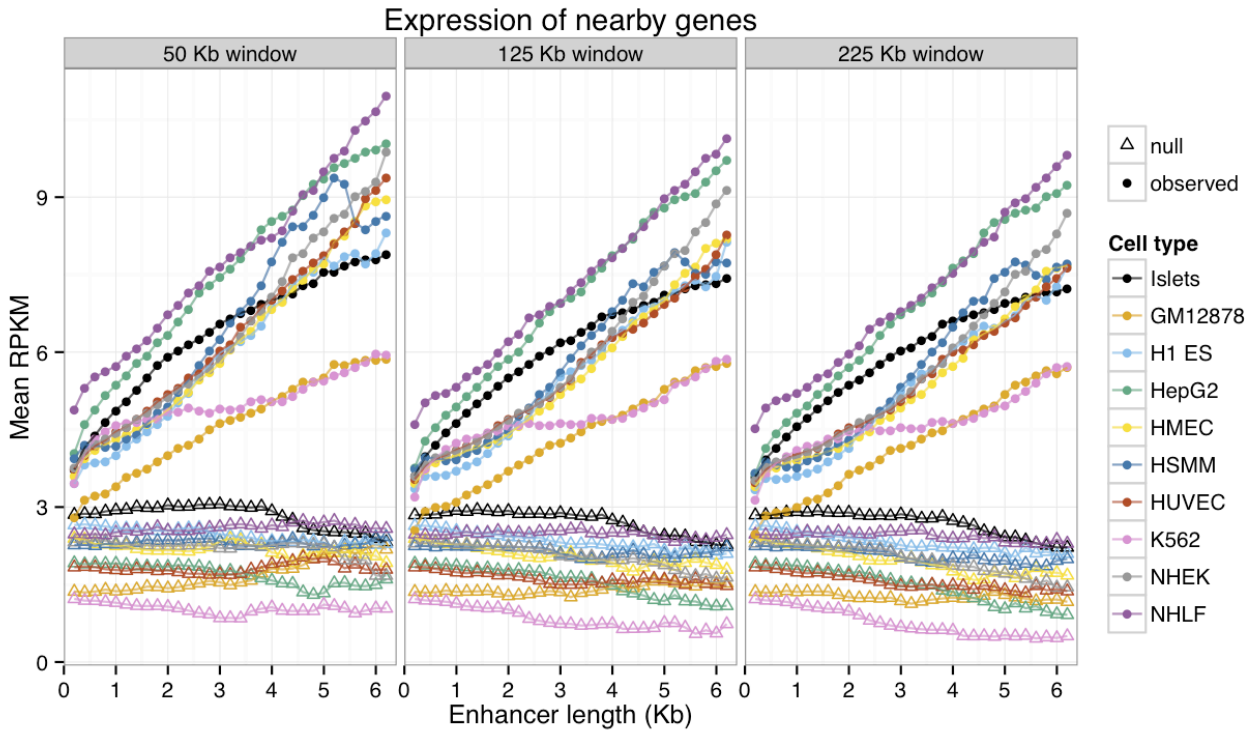



J. SORT1
Scale
100 kb
hg19
chr1:
109,800,000
109,850,000
109,900,000
109,950,000
Lipid GWAS SNP
rs12740374
Chromatin states
Islets
GM12878
H1 ESCs
HepG2
HMEC
HSMM
HUVEC
K562
NHEK
NHLF
RefSeq
SARS
CELSR2
PSRC1
SORT1
SORT1
MYBPHL
PSMA5
RNA-seq
Islets
GM12878
H1 ESCs
HepG2
HMEC
HSMM
HUVEC
K562
NHEK
NHLF
Scale
10 kb
hg19
chr1:
109,815,000
109,820,000
109,825,000
Lipid GWAS SNP
rs12740374
Chromatin states
Islets
GM12878
H1 ESCs
HepG2
HMEC
HSMM
HUVEC
K562
NHEK
NHLF
CELSR2
PSRC1

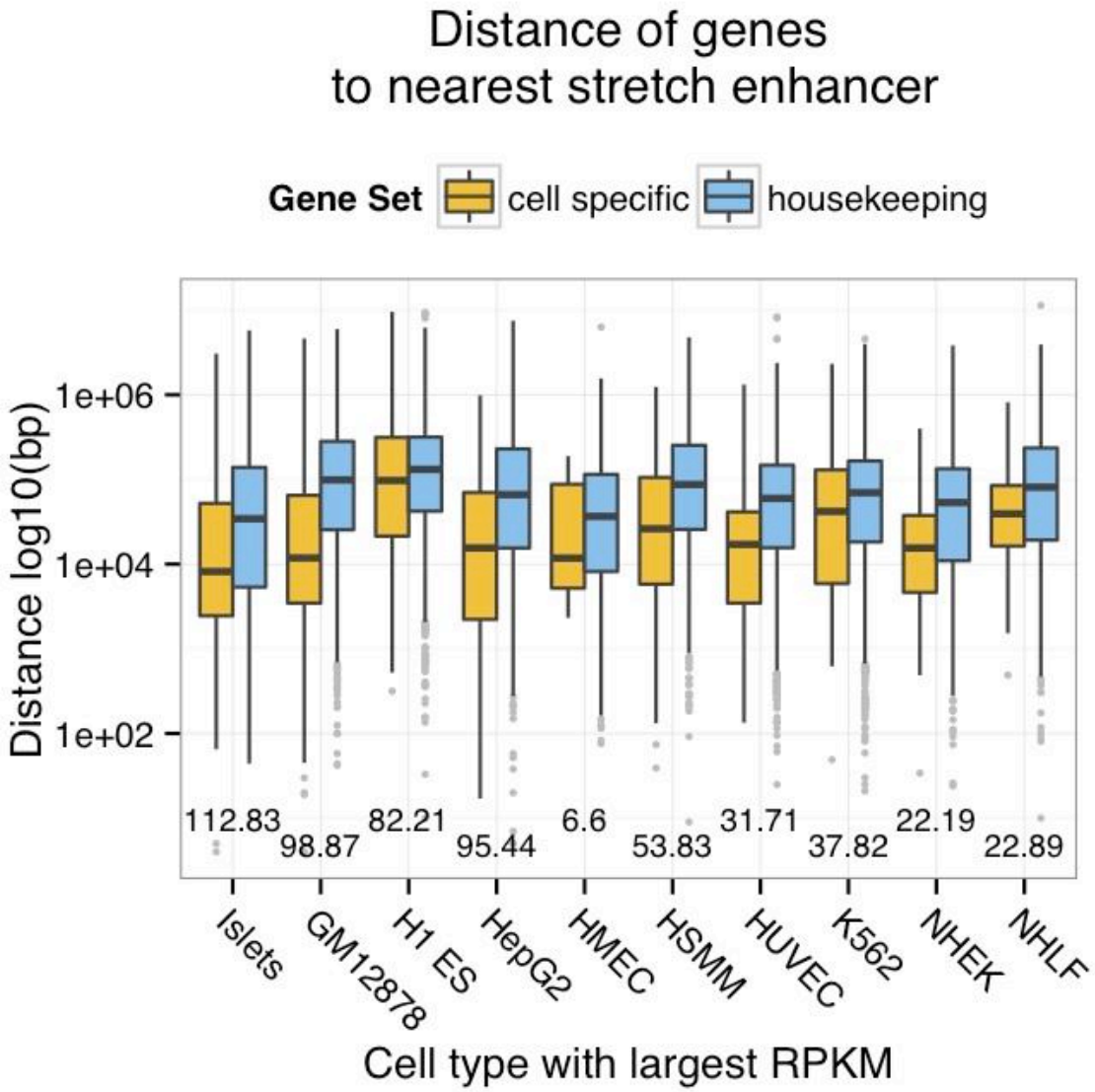



*Supplementary Figure 7*. **Locus control regions (LCRs) and GWAS enhancer SNPs overlap stretch enhancers.** UCSC Genome Browser views of LCRs and regions with GWAS

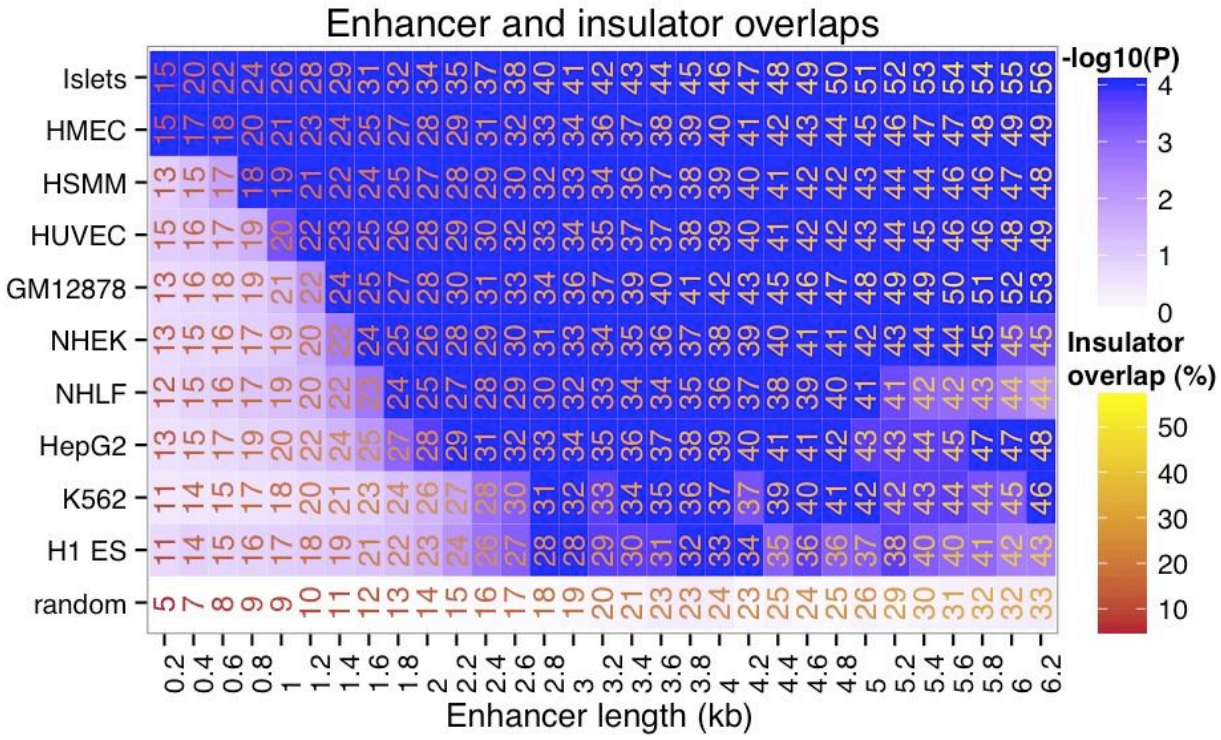
SNPs in enhancers or sites of long range interaction (described in Table S5). Overlaps with stretch enhancer signatures are displayed for beta globin (A) and alpha globin (B) in erythryoleukemia cells (K562), thymic regulatory region (*ADA*) in lymphoblastoid cells (GM12878) (C), *APOB* LCR (D), hepatic control region (*APOE/C1*) (E), and *AFP*/*ALB* (F) in hepatocellular carcinoma cells (HepG2), keratin (*KRT18*) in keratinocytes (NHEK) (G), and desmin (*DES*) in smooth muscle myoblasts (HSMM) (H). GWAS SNPs altering enhancer activity and affecting *MYC* (I) or *SORT1* (J) expression or participating in a long-range interaction with *IRS1* (K) reside in stretch enhancers. Chromatin state assignments are as shown in Figure 1A. The shaded blue box overlapping the chromatin state assignments denotes each LCR location. **Table S5** contains hg19 coordinate information and references used to determine the location of each LCR (16–28). Zoomed views of select regions are represented by purple or black rectangles, and the magnified view is shown below the rectangles.



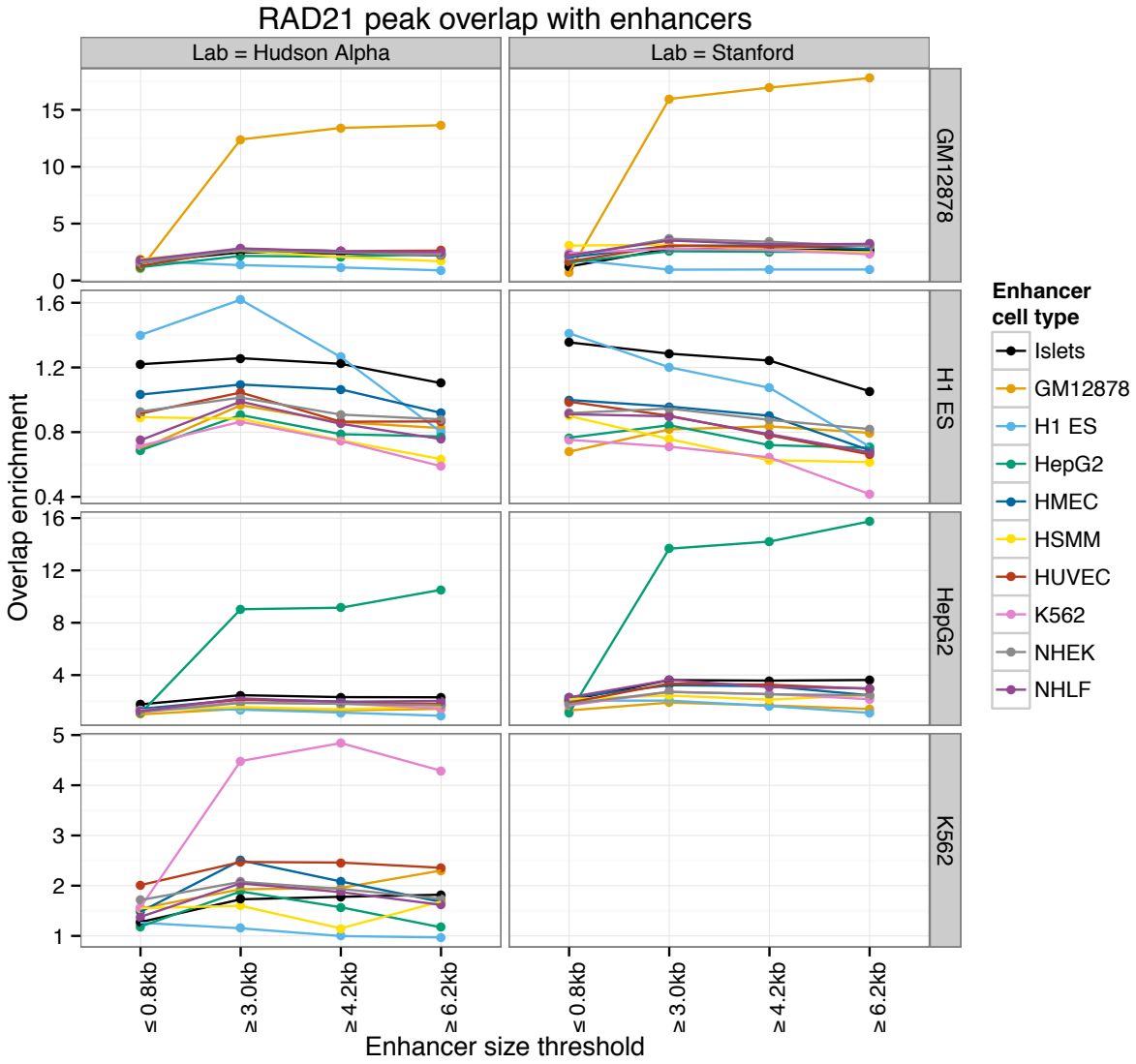

*Supplementary Figure 8*. Expression levels of genes nearby stretch enhancers at different gene-enhancer linking window thresholds (50 kb, 125 kb, 225 kb). Null dataset (empty triangles) generated by randomly re-assigning gene expression patterns.



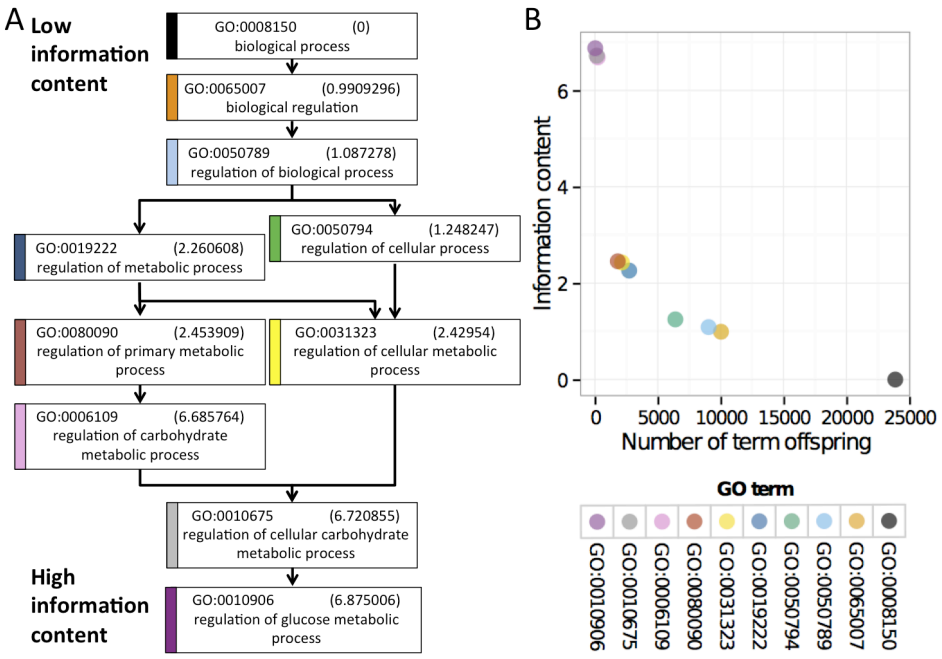

*Supplementary Figure 9*. Cell specific genes (those with normalized information content > 0.75) are significantly closer to stretch enhancers compared to housekeeping genes (those with normalized information content < 0.25). Genes were filtered for those that are expressed at a level of at least 3 RPKM in any cell type. Numbers below each pair of box plots represent -log10(P) from a Wilcoxon rank sum test; note that all comparisons are statistically significant.



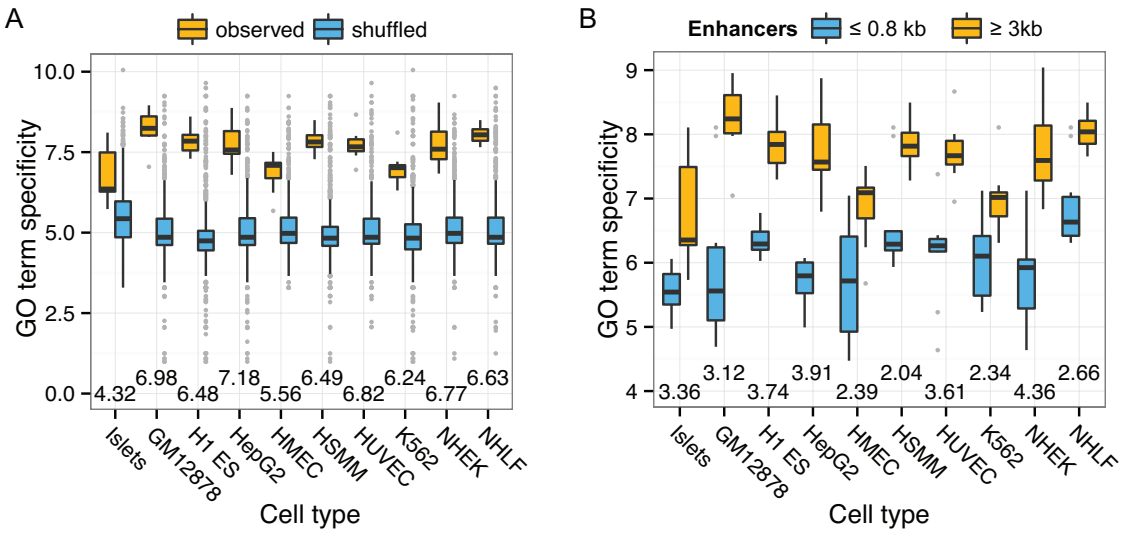

*Supplementary Figure 10*. Enhancer states overlap insulator (CTCF) sites used by other cell types with increasing length. The fraction of enhancers that overlap insulator states in other cell types is indicated with text and colored proportional to the level of overlap (brown = low, yellow = high). Blue shading indicates the significance of the observed overlap based on a permutation test (Methods). Note that random enhancers generated by shuffling islet enhancers show no significant overlap (bottom row).



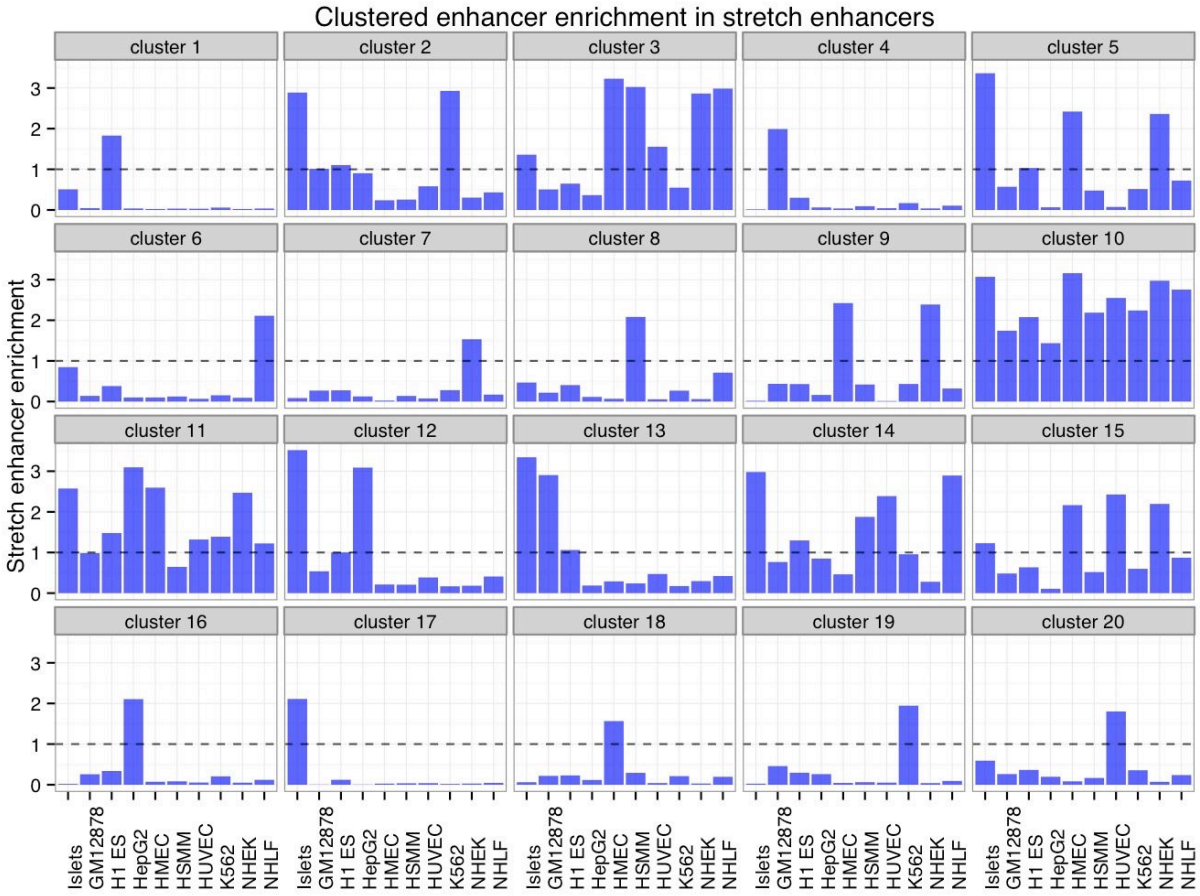

*Supplementary Figure 11*. Stretch enhancers are highly enriched to overlap the cohesin complex component RAD21 in relevant cell types. We used ENCODE RAD21 ChIP-seq peaks that do not overlap CTCF peaks for GM12878 (top row), H1 ES (2nd row from top), HepG2 (3rd row from top), and K562 (bottom row) cells. Enhancers that represent the median size (0.8 kb) or smaller are not enriched, whereas stretch enhancers greater than or equal to different thresholds (3.0 kb = 90%, 4.2 Kb = 95%, 6.2 kb = 99%) are highly enriched in all differentiated cell types. Notably, the enrichment is specific to the relevant cell type—for example, GM12878 stretch enhancers are enriched to mark RAD21 ChIP-seq peaks in GM12878 cells and not any other cell types (top row). H1 ES cells are not enriched, which supports the concept that stretch enhancers are a mark of differentiated cell types.



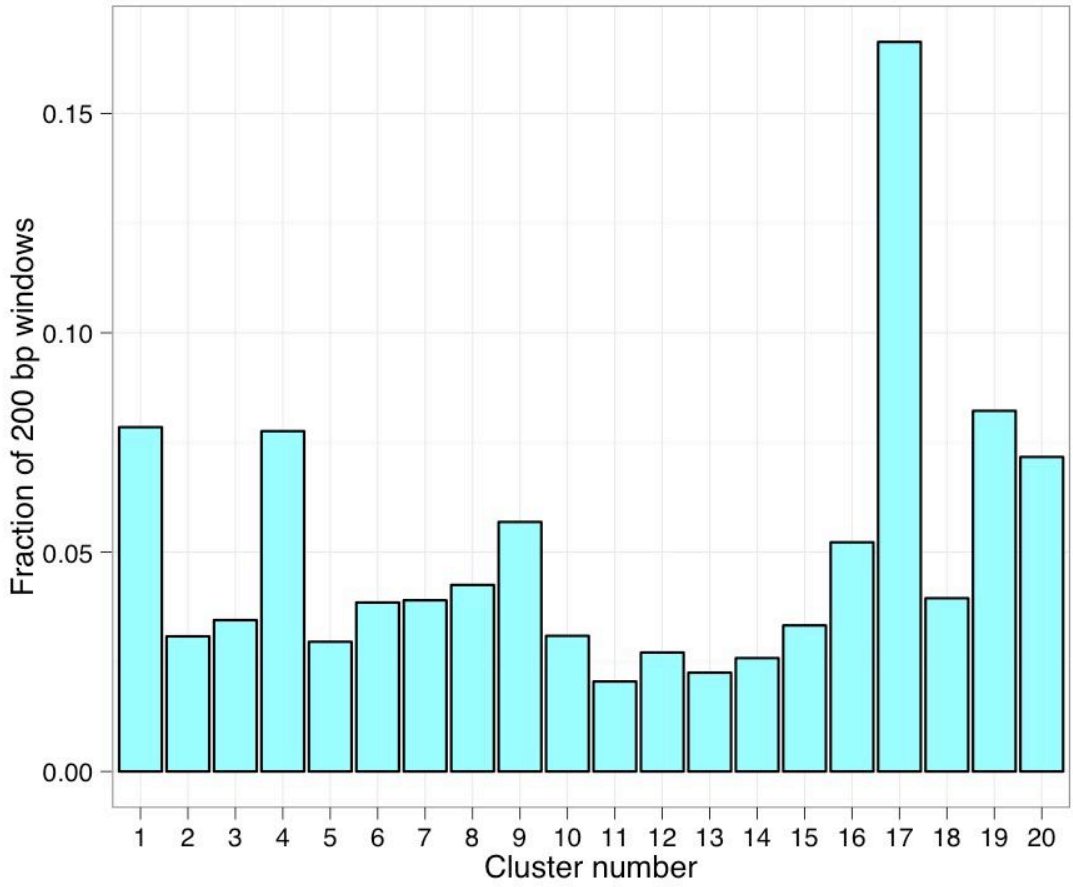

*Supplementary Figure 12*. **A**, GO tree for GO:0010906 (regulation of glucose metabolic process), the 10th most enriched term in Islet enhancers ≥ 6.2 kb. Each box represents a GO term. Within each box is the GO ID, information content (in parenthesis), and term description. For simplicity, "*is a*" relationships are depicted. Note that GO:0008150 is a root ontology term and has an information content of 0. **B**, GO terms from **(A)** plotted by their information content and number of offspring (*i.e.* the children of a term, their children, and so on), calculated using the GO.db R library. We use information content as a measure of GO term specificity.



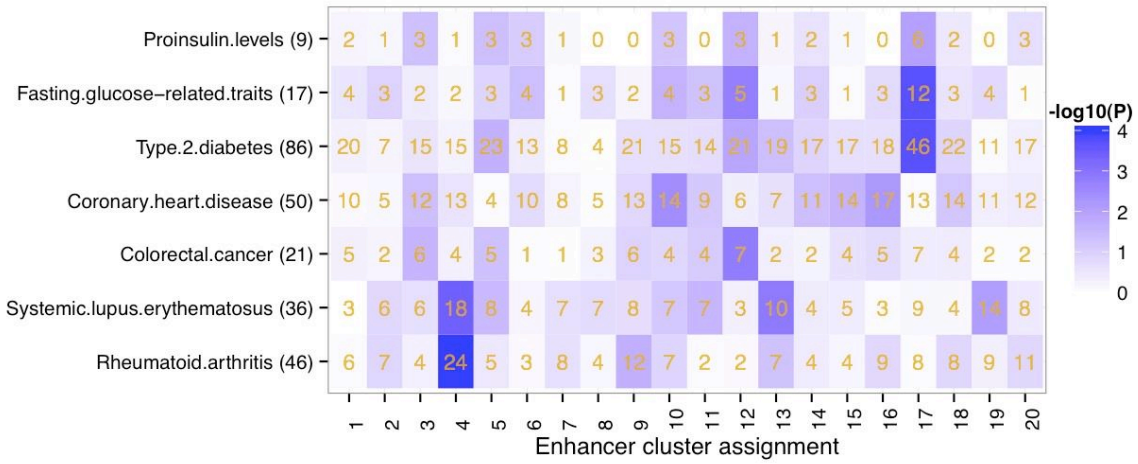

*Supplementary Figure 13*. GO term specificity in stretch enhancers compared to control enhancer regions. **A,** Specificity scores for GO terms associated with stretch enhancers (≥ 3 kb) in observed (yellow) and shuffled (blue) datasets. **B,** Specificity scores for GO terms associated with enhancers less than or equal to the median size (≤ 0.8 kb) (blue) or stretch enhancers (≥ 3 kb) (yellow). Numbers at the bottom of each pair of box plots represent -log10(*P*) for a Wilcoxon rank sum test.



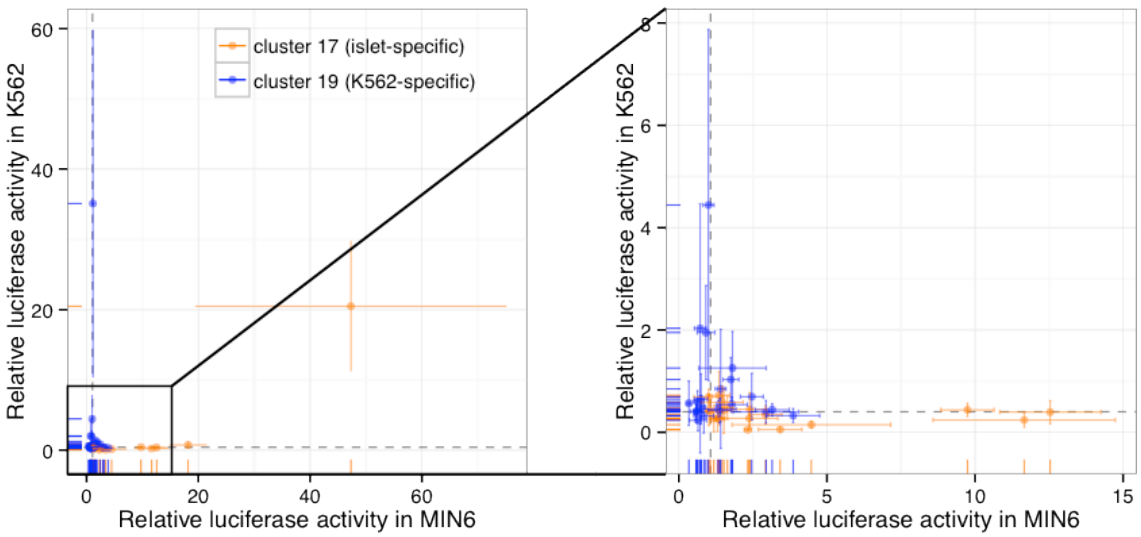

*Supplementary Figure 14*. Cell-type specific enhancer clusters are enriched to occur in stretch enhancers in relevant cell types. For example, islet-specific cluster 17 is specifically enriched in islet stretch enhancers. Enrichment was calculated as the fraction of enhancers in a cluster that overlap stretch enhancers in a cell type divided by the fraction of stretch enhancers in that cell type. The horizontal dashed line at 1 indicates no enrichment.



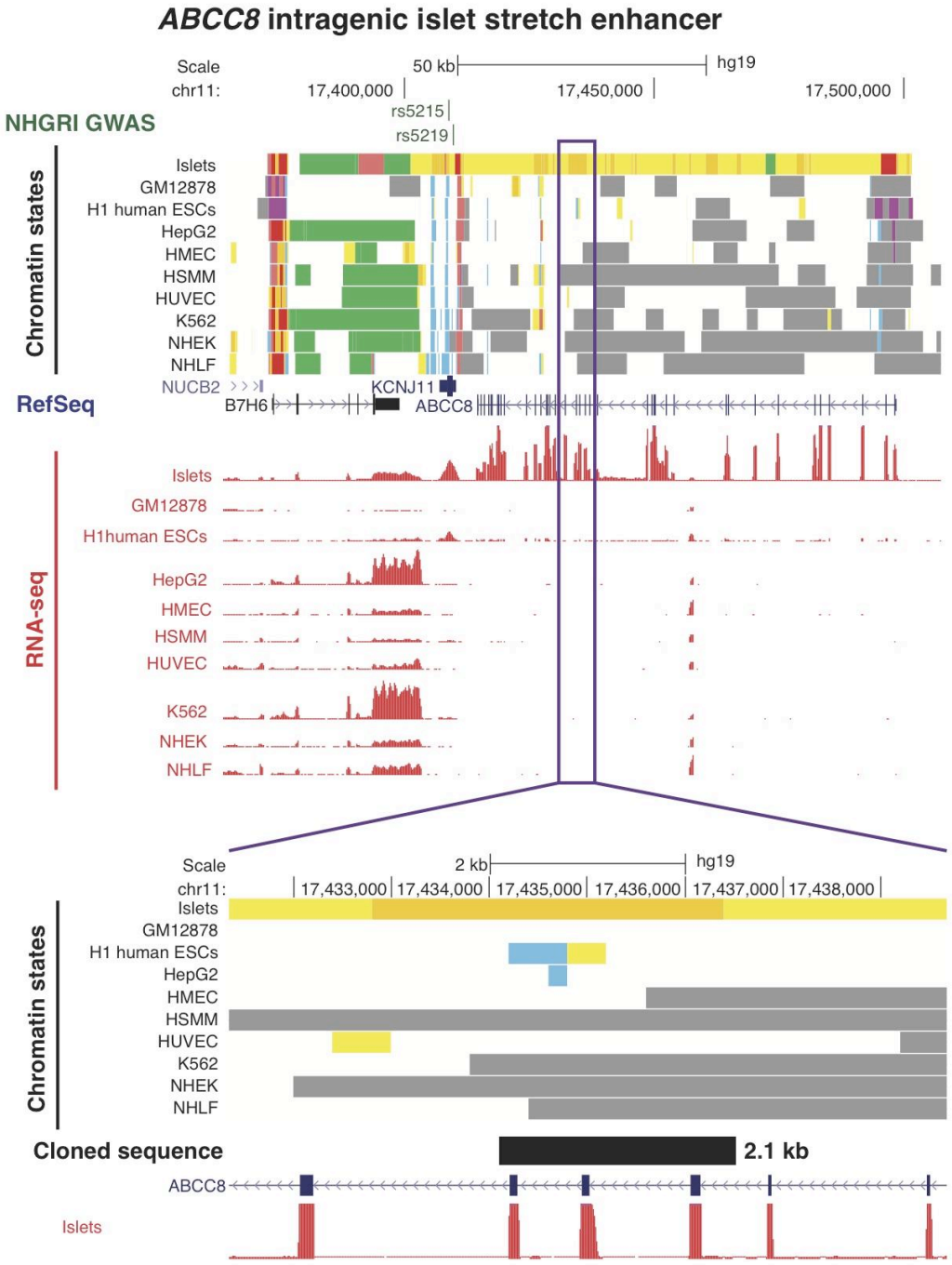

*Supplementary Figure 15*. Enhancer cluster representation for all 200 bp enhancer windows.



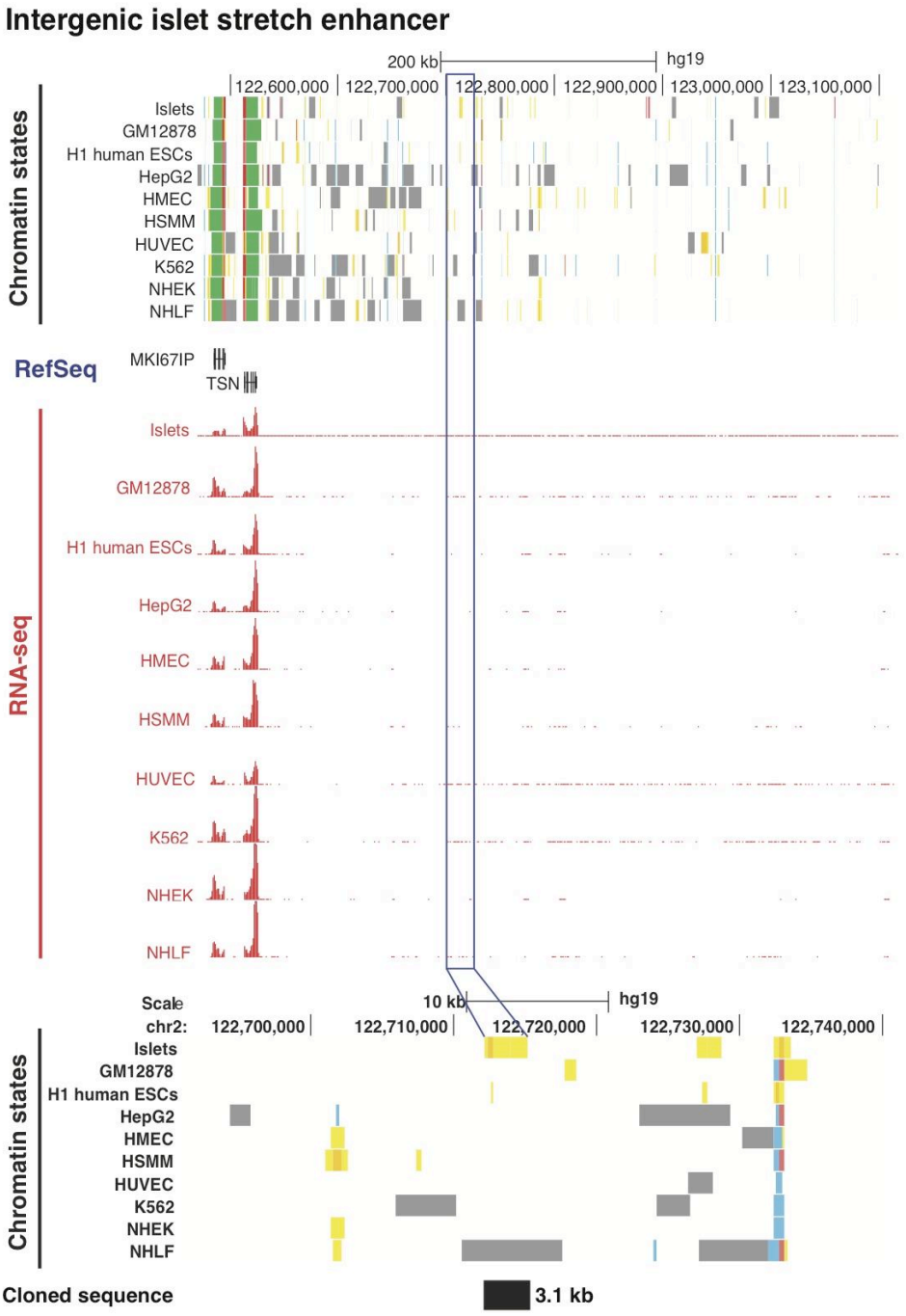

*Supplementary Figure 16*. Positions of index and tightly linked ($r^2 \geq 0.8$) SNPs for different diseases or traits (y-axis) were overlapped with those of enhancer states for each cell type (x-axis). The number of SNP loci overlapping enhancer states in each cell type is indicated in orange. Blue shading indicates the significance of SNP locus enrichment relative to a null distribution (Methods). The total number of GWAS loci for each trait is indicated in parentheses on the y-axis. Notably, T2D GWAS SNPs are significantly enriched to overlap islet-specific cluster 17 enhancers.




*Supplementary Figure 17*. Luciferase reporter construct activity for cluster 17 (islet-specific) and 19 (K562-specific) stretch enhancers tested in MIN6 and K562 cells show cell-type specific activity. Points represent the mean of at least three experiments and error bars represent one standard deviation. Dashed gray lines represent the median activity of cluster 17 enhancers in K562 cells (horizontal line) and cluster 19 enhancers in MIN6 cells (vertical line), and are used as a reference for a random expectation. Short rug lines at the base of each axis are used as a reference for each point.




*Supplementary Figure 18*. Browser shot showing *ABCC8* intragenic region used to generate transgenic mice. Annotated genes are indicated in the RefSeq portion of the view. NHGRI GWAS catalog SNPs associated with type 2 diabetes are indicated in green at the top. The lower panel with chromatin states and islet-only RNA-seq data is a zoomed-in view of the purple box above. The black rectangle indicates the DNA sequence in this stretch enhancer region cloned and tested for enhancer activity in luciferase and *lacZ* assays.




*Supplementary Figure 19*. Browser shot showing intergenic stretch enhancer region used to generate transgenic mice.

## Supplementary tables

**Table S1.** Stretch enhancer counts and fraction in the 90$^{th}$ (3 kb), 95$^{th}$ (4.2 kb), and 99th (6.2 kb) percentile of length by cell type

| | All | >=3kb | | >=4.2kb | | >=6.2kb | |
|---|---|---|---|---|---|---|---|
| | N | N | Fraction of All | N | Fraction of All | N | Fraction of All |
| All cell types | 1,108,378 | 111,592 | 0.101 | 59,075 | 0.053 | 24,975 | 0.023 |
| GM12878 | 94,115 | 10,355 | 0.110 | 5,547 | 0.059 | 2,318 | 0.025 |
| H1 ES | 132,250 | 6,426 | 0.049 | 2,699 | 0.020 | 799 | 0.006 |
| HMEC | 125,476 | 12,997 | 0.104 | 6,459 | 0.051 | 2,403 | 0.019 |
| HSMM | 91,956 | 7,284 | 0.079 | 3,157 | 0.034 | 989 | 0.011 |
| HepG2 | 80,433 | 7,969 | 0.099 | 4,445 | 0.055 | 1,884 | 0.023 |
| Islets | 173,055 | 23,013 | 0.133 | 14,720 | 0.085 | 8,173 | 0.047 |
| HUVEC | 89,930 | 10,890 | 0.121 | 5,642 | 0.063 | 2,137 | 0.024 |
| K562 | 102,990 | 10,142 | 0.098 | 5,301 | 0.051 | 2,115 | 0.021 |
| NHEK | 119,265 | 12,658 | 0.106 | 6,168 | 0.052 | 2,317 | 0.019 |
| NHLF | 98,908 | 9,858 | 0.100 | 4,937 | 0.050 | 1,840 | 0.019 |

**Table S2.** Islet sample information

| Sample ID | Sex | Purity | Viability | BMI | Age | Cause of Death | Ethnicity | Isolation site | Approximate amount of crosslinked material (1 islet equivalent (IEQ)= ~1000 cells) | ChIP modifications | RNA integrity | Corresponding designation from Stitzel et al., 2010 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| UG3360 | F | 80 | 97 | 24 | 37 | Not reported | Caucasian | University of Illinois | 18,000 IEQ | Input, K4me1, K4me3, **K36me3** | 7.4 | Islet 4 |
| ULI102 | M | 90 | 96 | 30 | 54 | Intracranial hemorrage | Caucasian | University of Washington | 18,000 IEQ | **Input, K27ac** | 8 | None-new sample |
| VAS050 | F | 85 | >90 | 28 | 47 | Cerebrovascular accident (stroke) | Caucasian | University of Miami | 16,000 IEQ | **Input, K4me3, K27ac** | N/A | None-new sample |
| UGA076 | M | 90 | 95 | 27.9 | 60 | Cerebrovascular hemorrhage | Caucasian | University of Miami | 16,000 IEQ | Input, CTCF | 8.4 | Islet 6 |
| UFK467 | M | 70 | 93 | 26.5 | 16 | Blunt head trauma | Caucasian | University of Alabama Birmingham | 18,000 IEQ | Input, K4me1, **K36me3** | 8.5 | Islet 3 |
| WBB020 | M | 85 | 95 | 24.7 | 36 | Self-inflicted gunshot wound to the head | Caucasian | University of Illinois | 16,000 IEQ | Input, **K36me3** | N/A | Islet 5 |

**Table S3.** Uniquely mappable reads for each data set analyzed in this study

| Cell Type | Data Type | Uniquely Mappable Reads | Source |
|---|---|---|---|
| Islets | CTCF | 17,625,608 | (1) |
| Islets | H3K27ac | 43,995,919 | this study |
| Islets | H3K27me3 | 10,773,905 | Roadmap Epigenomics |
| Islets | H3K36me3 | 23,988,227 | this study |
| Islets | H3K4me1 | 54,894,639 | (1) |
| Islets | H3K4me3 | 63,648,542 | (1) |
| Islets | Input | 147,768,337 | (1) |
| Islets | RNA-seq | 264,307,429 | this study |
| GM12878 | CTCF | 24,534,133 | (2) |
| GM12878 | H3K27ac | 14,488,303 | (2) |
| GM12878 | H3K27me3 | 18,556,632 | (2) |
| GM12878 | H3K36me3 | 20,205,900 | (2) |
| GM12878 | H3K4me1 | 20,377,456 | (2) |
| GM12878 | H3K4me3 | 24,834,648 | (2) |
| GM12878 | Input | 12,063,059 | (2) |
| GM12878 | RNA-seq | 110,323,019 | (3) |
| H1 ES | CTCF | 22,856,737 | (2) |
| H1 ES | H3K27ac | 13,443,751 | (2) |
| H1 ES | H3K27me3 | 15,469,166 | (2) |
| H1 ES | H3K36me3 | 22,627,980 | (2) |
| H1 ES | H3K4me1 | 24,126,999 | (2) |
| H1 ES | H3K4me3 | 17,237,666 | (2) |
| H1 ES | Input | 13,085,773 | (2) |
| H1 ES | RNA-seq | 111,621,629 | (3) |
| HepG2 | CTCF | 13,012,438 | (2) |
| HepG2 | H3K27ac | 12,576,850 | (2) |
| HepG2 | H3K27me3 | 10,345,256 | (2) |
| HepG2 | H3K36me3 | 11,127,564 | (2) |
| HepG2 | H3K4me1 | 12,589,777 | (2) |
| HepG2 | H3K4me3 | 16,527,917 | (2) |
| HepG2 | Input | 10,585,987 | (2) |
| HepG2 | RNA-seq | 90,007,371 | (3) |
| HMEC | CTCF | 15,564,014 | (2) |
| HMEC | H3K27ac | 19,238,712 | (2) |
| HMEC | H3K27me3 | 15,320,886 | (2) |
| HMEC | H3K36me3 | 17,833,196 | (2) |
| HMEC | H3K4me1 | 30,541,298 | (2) |

| | | | |
|---|---|---|---|
| HMEC | H3K4me3 | 23,921,373 | (2) |
| HMEC | Input | 15,574,355 | (2) |
| HMEC | RNA-seq | 146,247,741 | (3) |
| HSMM | CTCF | 17,003,672 | (2) |
| HSMM | H3K27ac | 18,965,651 | (2) |
| HSMM | H3K27me3 | 18,396,680 | (2) |
| HSMM | H3K36me3 | 33,459,808 | (2) |
| HSMM | H3K4me1 | 16,712,016 | (2) |
| HSMM | H3K4me3 | 17,027,960 | (2) |
| HSMM | Input | 16,481,856 | (2) |
| HSMM | RNA-seq | 116,648,081 | (3) |
| HUVEC | CTCF | 11,400,168 | (2) |
| HUVEC | H3K27ac | 21,122,677 | (2) |
| HUVEC | H3K27me3 | 19,847,745 | (2) |
| HUVEC | H3K36me3 | 17,264,410 | (2) |
| HUVEC | H3K4me1 | 23,587,423 | (2) |
| HUVEC | H3K4me3 | 18,583,347 | (2) |
| HUVEC | Input | 17,677,974 | (2) |
| HUVEC | RNA-seq | 90,086,549 | (3) |
| K562 | CTCF | 17,723,407 | (2) |
| K562 | H3K27ac | 16,665,144 | (2) |
| K562 | H3K27me3 | 17,432,398 | (2) |
| K562 | H3K36me3 | 17,262,071 | (2) |
| K562 | H3K4me1 | 19,662,679 | (2) |
| K562 | H3K4me3 | 19,463,671 | (2) |
| K562 | Input | 17,310,513 | (2) |
| K562 | RNA-seq | 110,980,921 | (3) |
| NHEK | CTCF | 11,927,593 | (2) |
| NHEK | H3K27ac | 20,386,596 | (2) |
| NHEK | H3K27me3 | 15,806,100 | (2) |
| NHEK | H3K36me3 | 16,243,945 | (2) |
| NHEK | H3K4me1 | 21,064,582 | (2) |
| NHEK | H3K4me3 | 16,927,909 | (2) |
| NHEK | Input | 17,093,736 | (2) |
| NHEK | RNA-seq | 125,098,881 | (3) |
| NHLF | CTCF | 15,551,676 | (2) |
| NHLF | H3K27ac | 15,358,166 | (2) |
| NHLF | H3K27me3 | 15,148,729 | (2) |
| NHLF | H3K36me3 | 21,736,486 | (2) |
| NHLF | H3K4me1 | 18,174,021 | (2) |
| NHLF | H3K4me3 | 26,943,869 | (2) |

| | | | |
|---|---|---|---|
| NHLF | Input | 11,747,259 | (2) |
| NHLF | RNA-seq | 164,643,571 | (3) |

**Table S4.** Stretch enhancer constructs used in luciferase and lacZ assays

| TARGETED REGION | | | | | | | | CLONED AMPLICON | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Construct** | **Chromosome** | **start** | **stop** | **map** | **repeat** | **dhsCount** | **tssDist** | **start** | **stop** | **Amplicon size (bp)** | **Forward amplification primer sequence** | **Reverse amplification primer sequence** | ***lacZ*** |
| Islet-1 | chr22 | 46420000 | 46422600 | 0.905064 | 0.43885 | 0 | 17505 | 46419941 | 46422713 | 2772 | TGACAACACAGCAGTCACCTC | CAGGGACCTCACCAAGAGC | |
| Islet-2 | chr2 | 155544200 | 155546800 | 0.957157 | 0.10692 | 0 | 8293 | 155544121 | 155546884 | 2763 | GCTGCCACCACAACTAACTTT | AATTCCTCCCAGTTTTTACTGAAG | |
| Islet-4 | chr5 | 50274200 | 50277400 | 0.976603 | 0.08 | 0 | 311429 | 50274079 | 50277447 | 3368 | TGGCTTATATGATTGGTCATGC | GGACCCCTGCTCAAATGC | |
| Islet-5 | chr6 | 136644600 | 136647600 | 0.923843 | 0.19567 | 1 | 33612 | 136644543 | 136647656 | 3113 | CAATGAGGCAAAACCCATA | CAGACCGTGAGAGCCATTTA | + |
| Islet-6 | chr19 | 42203000 | 42206000 | 0.927772 | 0.09733 | 1 | 6530 | 42202926 | 42206069 | 3143 | ATTGAGTTCATTCCCCAGGT | CAGAACCCCACTAGCACGAC | |
| Islet-7 | chr5 | 133163000 | 133165800 | 0.91573 | 0.1825 | 1 | 138606 | 133162946 | 133165886 | 2940 | CCTCTGACAGGTAATGAATCG | ACAGTTCTGGGTCAGGGCTA | |
| Islet-10 | chr5 | 63960000 | 63962600 | 0.994508 | 0.20692 | 0 | 23535 | 63959931 | 63962666 | 2735 | CCTGCTGGCCTGTTTTACAAT | TCACAAGGCTGTTAGCTGATG | |
| Islet-11 | chr1 | 177590200 | 177593400 | 0.992848 | 0.36 | 0 | 345650 | 177590150 | 177593483 | 3333 | CATGGCACATCTGGTTTTCA | TCCTCACATGACCTCGACTG | |
| Islet-13 | chr12 | 94927200 | 94930000 | 0.912083 | 0.39714 | 2 | 25565 | 94927149 | 94930061 | 2912 | AGTCCCCACCATACTGTGTTT | ATGTGTCCCTGCATCAACTG | |
| Islet-16 | chr2 | 205122600 | 205125200 | 1 | 0.01038 | 1 | 285316 | 129463351 | 129466958 | 3607 | GACAGGATGTCCATGACAGAA | AATGTCCAGATGTGGCTATGG | |
| Islet-17 | chr10 | 84922400 | 84925200 | 0.931766 | 0.42786 | 0 | 973985 | 84922334 | 84925292 | 2958 | GCCTGGGTGATCCTAGAACT | TTCCAAAAGCAAAATGTAAATATGA | |
| Islet-18 | chr5 | 71302000 | 71305000 | 0.988364 | 0.35933 | 0 | 98118 | 71301929 | 71305062 | 3133 | TGTAATAAAAGGAAATCAATCTATGC | CCTACTGCCTGGACTGTGAA | |
| Islet-19 | chr4 | 181594600 | 181597400 | 1 | 0.10429 | 0 | 482902 | 181594510 | 181597509 | 2999 | TGTGAGTTCTATGGCCTCAAC | TTGGGCAAGGAAGACATTTG | |
| Islet-20 | chr9 | 72600600 | 72603200 | 0.969077 | 0.27192 | 0 | 55297 | 72600507 | 72603303 | 2796 | CCTGTAGAAGATTTGTCATCCACTT | TGCCAGAAGCATAAAACCAG | |
| Islet-21 | chr1 | 208698600 | 208701400 | 0.999881 | 0.1625 | 1 | 280936 | 208698513 | 208701484 | 2971 | AGAACATCTGGCGAACCTG | CACCACATTTATTGAGCAGTGA | |
| Islet-25 | chr6 | 55575000 | 55577800 | 1 | 0.27107 | 0 | 130989 | 55574919 | 55577851 | 2932 | AGTGGGTAGATTCCTTCCTTTTT | ATTCCTATCCCCGCAACAAT | |
| Islet-27 | chr11 | 21928600 | 21931600 | 0.958086 | 0.27633 | 0 | 283122 | 21928488 | 21931656 | 3168 | GCAGCAATCTAAGAAATAGTGTTCC | GATTGTCTACTGTAAGATTTCACAAA | |
| Islet-28 | chr4 | 59639800 | 59643400 | 0.927564 | 0.20194 | 0 | 1663250 | 59639681 | 59643469 | 3788 | CGGATTAAAGCTGGAAATAGATG | TGGATTTTGTTGGATGGTTTC | |
| Islet-31 | chr6 | 95056600 | 95059200 | 0.966395 | 0.23962 | 1 | 639800 | 95056540 | 95059311 | 2771 | TACACCCTGAGAGAGGGAGA | TCTTGGGATAATTTTTATGATCTTC | + |
| Islet-32 | chr2 | 122712200 | 122715200 | 0.947318 | 0.41933 | 1 | 199080 | 122712120 | 122715253 | 3133 | GCAGTGGTTTATAGTTCTCCTTGC | GCTGGTTCCTGTGACTTTCC | + |
| *ABCC8* | chr11 | 17434587 | 17436700 | 1 | 0.12163 | 2 | 23710 | 17434517 | 17436675 | 2158 | GGGACCATCTGACCAGTACAAAC | TGCATCCATTTACTCCCTTCAC | |
| | | | | | | | | | | | | | |
| K562-3 | chrX | 49946600 | 49949600 | 0.988669 | 0.33467 | 0 | 15404 | 49946457 | 49949658 | 3201 | TGGATCAAAACCTCTTGGACA | GGCTTGGAAAGGCTGAGTAA | |
| K562-5 | chr14 | 36753800 | 36756400 | 0.999808 | 0.08269 | 0 | 33482 | 36753745 | 36756491 | 2746 | TTGGTGAACTTTTAAACTCAAAGC | TAGCCATGTGCCATGAACTG | |
| K562-7 | chr1 | 95548600 | 95551800 | 0.923416 | 0.14625 | 2 | 6273 | 95548541 | 95551861 | 3320 | CCTATAAGTAAAGTAAAGGCAATTATGT | TCCAAAAGACTAGGGTTGTTCA | |
| K562-8 | chr16 | 79876600 | 79880400 | 0.943822 | 0.36737 | 1 | 241979 | 79876523 | 79880454 | 3931 | CGTCAACCTACTCTACCTCTCC | GAAACTGTGCCTGCCAGATAG | |
| K592-9 | chr17 | 31241000 | 31243800 | 0.961548 | 0.36214 | 1 | 11128 | 31240940 | 31243848 | 2908 | CAATGATGGGAGGTGCTACTT | TTCCAGCTTCCCTGAAAAA | |
| K582-10 | chr5 | 173261000 | 173263800 | 1 | 0.14857 | 0 | 51531 | 173260941 | 173263897 | 2956 | CGGAAAACAAGGACTCACTG | TCTTCCTCTCAGTAGAACAATGAA | |
| K562-14 | chr3 | 33928400 | 33931600 | 1 | 0.41375 | 1 | 88338 | 33928292 | 33931695 | 3403 | TTTGGTAAAACTGAGAAAAATCACA | CAGAGCAAAGCCCTAATCTTG | |
| K562-15 | chr20 | 37319400 | 37322600 | 0.906724 | 0.33187 | 2 | 30505 | 37319251 | 37322654 | 3403 | CCGTTCTGTAGCTCAGGAATTA | CATAGAAGAGGAGGGGAACG | |
| K562-16 | chr5 | 8011600 | 8015400 | 0.904401 | 0.33658 | 1 | 142384 | 8011386 | 8015485 | 4099 | CAATTTGCCTTTTTCAATTATG | GGCCCCTAGTCATCTGAACA | |
| K562-17 | chr4 | 183930000 | 183933600 | 0.998773 | 0.17556 | 3 | 25218 | 183929948 | 183933687 | 3739 | CAGGCATCAAAAACCCAGAT | CCCTGTGGGAGCCATACATA | |
| K562-18 | chr4 | 10189000 | 10192000 | 0.989553 | 0.41933 | 0 | 70428 | 10188908 | 10192068 | 3160 | TGGCCTTTGACAGCATAAAG | GGACAGATGTGGCATGTTCA | |
| K562-19 | chr4 | 175341200 | 175343800 | 0.901394 | 0.30346 | 0 | 1146 | 175341135 | 175343879 | 2744 | TCTCACAAGCAGTTCACACTCA | CAACATCTTTCATCATTCATTTTTG | |
| K562-20 | chr12 | 90190000 | 90193200 | 0.974057 | 0.28063 | 0 | 87269 | 90189930 | 90193316 | 3386 | GAGAGAGAGAAGAGCATGAGGTG | GTTGGTTTGCACGGCTAGAT | |
| K562-21 | chr7 | 136173800 | 136176400 | 0.985559 | 0.12385 | 0 | 376999 | 136173711 | 136176483 | 2772 | TCACCATCATAAGAAATTGATGT | ACACTCTTCCGAGGGAAAAA | |
| K562-23 | chrX | 109185200 | 109189000 | 0.905052 | 0.40263 | 1 | 56863 | 109185067 | 109189064 | 3997 | TGGAAGGAGCTTCAAAGGAG | CTGTGAACCTAATGCCCAGA | |
| K562-25 | chr3 | 167874600 | 167877600 | 0.995437 | 0.391 | 1 | 61184 | 167874531 | 167877692 | 3161 | TGTTGACGAAGTTTGGACTG | GGGATTGGCAAAGCAGAGA | |
| K562-26 | chr10 | 130373200 | 130376400 | 0.99717 | 0.01031 | 0 | 448733 | 130373102 | 130376480 | 3378 | GCGTTTCAGTTCCTGGGTAG | CGAGATGGCAAAGTCTGGA | |
| K562-27 | chr1 | 39229200 | 39231800 | 0.902869 | 0.28385 | 0 | 93540 | 39229111 | 39231861 | 2750 | TGCTCCTGATGTAGTCTTTGTGA | TCCCTCATTTCATTTGTGTGTC | |
| K562-28 | chr6 | 106253800 | 106256800 | 0.98701 | 0.36767 | 0 | 277395 | 106253725 | 106256853 | 3128 | CAGGAAACCAAGAATAGTGTAAGG | TTTAAGGGAAACATATAGGCACAC | |
| K562-29 | chr3 | 105850600 | 105853600 | 0.998142 | 0.132 | 0 | 262714 | 105850496 | 105853659 | 3163 | TCGTCTAGGGAACAAGAGTCC | GGCCAAGAACAATGCCTACA | |
| K562-30 | chr4 | 10199200 | 10203600 | 0.911398 | 0.30705 | 2 | 80628 | 10198871 | 10203814 | 4943 | TGGCTTCTAGCTTCATGTCC | GGGGCACAGCCTTCTTCTAT | |
| K562-32 | chr4 | 10242200 | 10248200 | 0.911927 | 0.40017 | 1 | 123628 | 10242103 | 10248251 | 6148 | GAGCTGAACAAACCCTTCTATGA | TGACCCTTTGAGGTAAGTCCA | |
| K562-33 | chr13 | 29141200 | 29144600 | 0.904094 | 0.42853 | 1 | 71936 | 29141138 | 29144709 | 3571 | ACGTTTTATCACCTGTTGAGG | CTCAATCCAGGAAGTTGTCA | |

All coordinates are hg19
**Map** = fraction of nucleotide bases in the region that are mapable using 36 bp sizes
**Repeat** = fraction of nucleotide bases in the region that are part of repetitive sequences
**dhsCount** = Number of DNase hypersensitive peaks in the region
**tss Dist** = Distance from the nearest Transcription Start Site (TSS) of a RefSeq gene

**Table S5.** LCR coordinates, ENSEMBL gene designations of "target" genes for RPKM/specificity calculations

LCR stretch enhancer overlaps

| LCR Name | Chrom. | Start | Stop | Target gene | ENSEMBL ID | REFERENCE |
|---|---|---|---|---|---|---|
| Beta globin LCR | 11 | 5312534 | 5296894 | *HBG1* in K562 cells | ENSG00000213934 | (17, 28) |
| Alpha globin LCR | 16 | 147854 | 194854 | *HBA2* | ENSG00000188536 | (18) |
| Thymic regulatory region | 20 | 43270847 | 43272075 | *ADA* | ENSG00000196839 | (19) |
| *INS* open chromatin domain | 11 | 2163424 | 2243424 | *INS/TH* | ENSG00000254647 | (16) |
| Hepatic control region (HCR) | 19 | 45427522 | 45428295 | *APOE/APOC1* | ENSG00000130208 | (20) |
| *ALB/AFP/AFM* | 4 | 74257972 | 74266472 | *ALB* | ENSG00000163631 | (21) |
| *ALB/AFP/AFM* | 4 | 74295433 | 74299433 | *AFP* | ENSG00000081051 | (21) |
| *APOB* | 2 | 21267257 | 21272315 | *APOB* | ENSG00000084674 | (22) |
| *APOB* | 2 | 21222271 | 21223942 | *APOB* | ENSG00000084674 | (22) |
| *Keratin* | 12 | 53341743 | 53344110 | *KRT18* | ENSG00000111057 | (23) |
| Desmin | 2 | 220265099 | 220274099 | *DES* | ENSG00000175084 | (24) |

GWAS enhancer/ChIA-PET SNPs for stretch enhancer overlaps

| SNP ID | Chrom. | Start | Stop | Target gene | ENSEMBL ID | REFERENCE |
|---|---|---|---|---|---|---|
| rs6983267 | 8 | 128413304 | 128413305 | *MYC* | ENSG00000136997 | (25, 26) |
| rs7578326 | 2 | 227020652 | 227020653 | *IRS1* | ENSG00000169047 | (29) |
| rs12740374 | 1 | 109817589 | 109817590 | *SORT1* | ENSG00000134243 | (27) |