Cell Reports, Volume *18*

Supplemental Information

cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture

Jennifer Yihong Tan, Adam Alexander Thil Smith, Maria Ferreira da Silva, Cyril Matthey-Doret, Rico Rueedi, Reyhan Sönmez, David Ding, Zoltán Kutalik, Sven Bergmann, and Ana Claudia Marques

Supplemental Data.

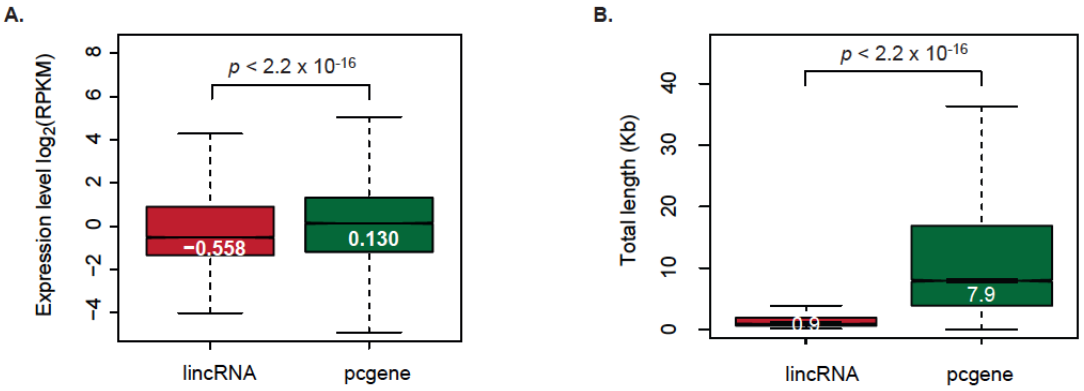

Figure S1. lincRNAs are generally shorter and more lowly expressed than protein-coding genes. Related to Figure 1. (A) Distribution of the expression levels [$\log_2$(RPKM)] and B) transcript length (Kb) of LCL-expressed lincRNAs (red) and protein-coding genes (green). Differences between groups were tested using a two-tailed Mann-Whitney *U* test and *p*-values are indicated.

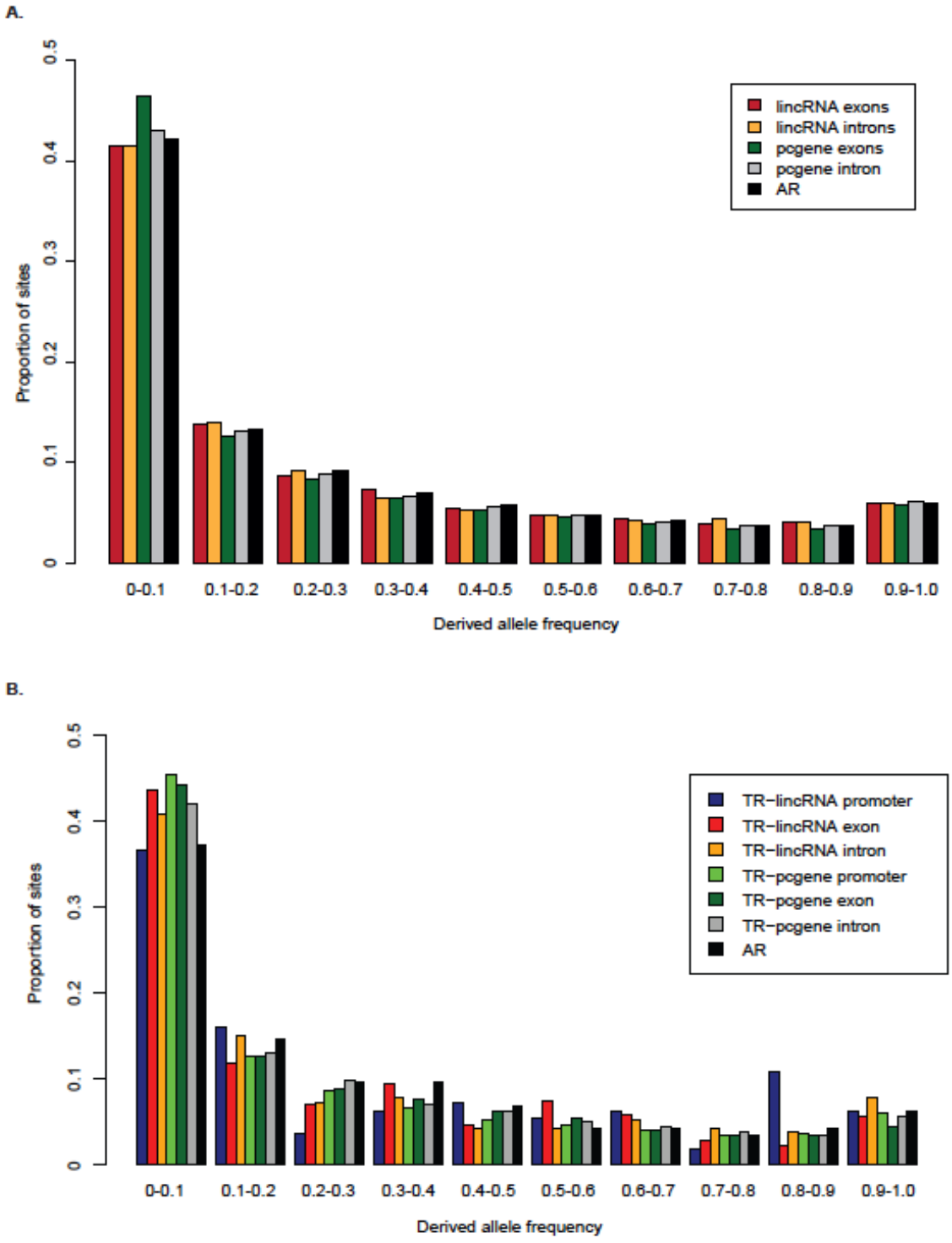

Figure S2. TR-lincRNAs evolved under purifying selection during recent human history. Related to Figure 2. Distribution of derived allele frequency (DAF) for (A) variants within exons and introns of LCL-expressed lincRNAs (exon-red, intron-orange) and protein-coding genes (exon-green, intron-grey) and for (B) variants within putative promoter regions (+/-1 Kb from TSS), exons and introns of TR-lincRNAs (promoter-dark blue, exons-red, introns-orange) and TR-pcgenes (promoter-light green, exon-dark green, intron-grey), and local ancestral repeats (ARs, black).

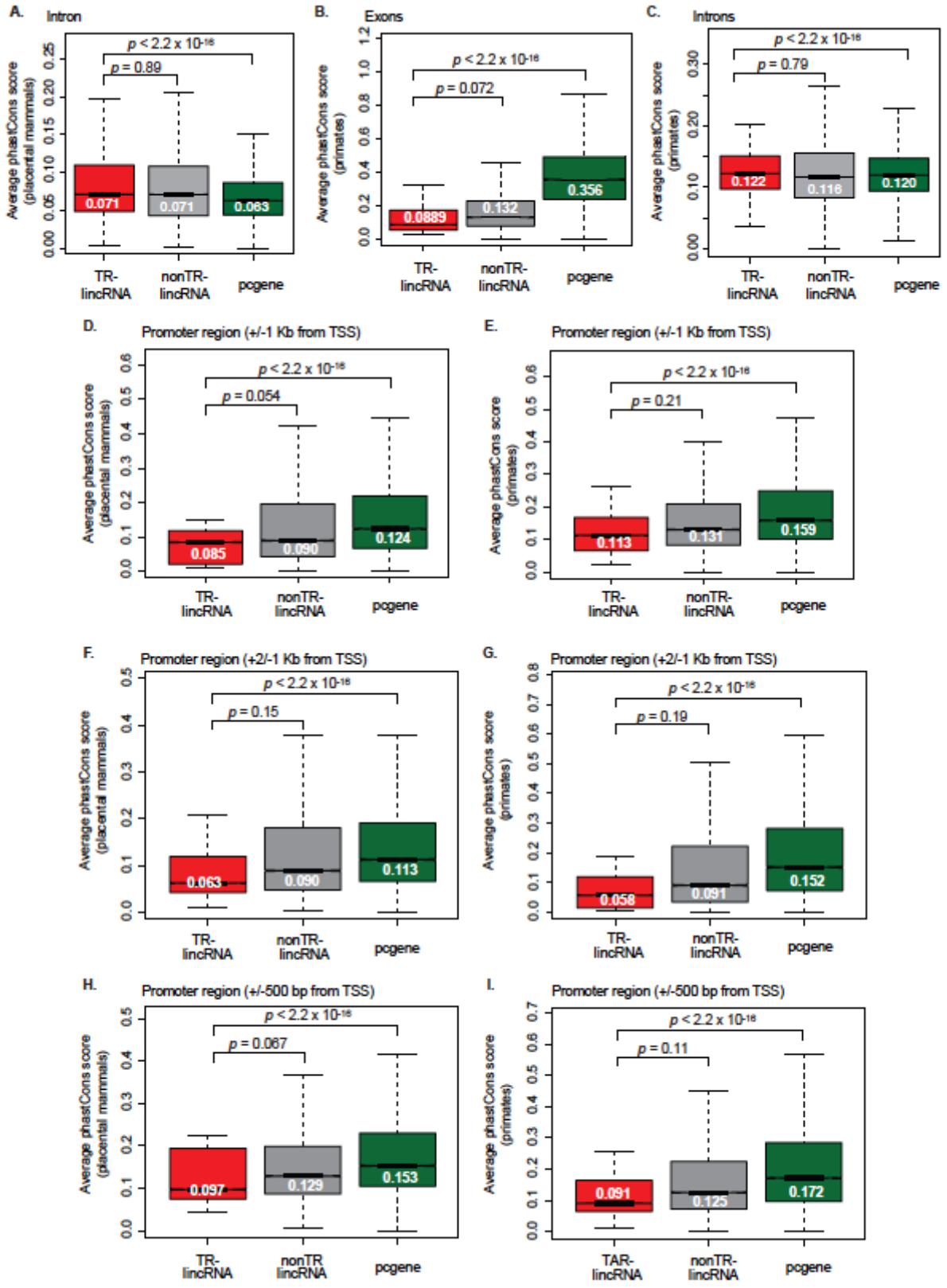

Figure S3. No evidence of constraint during TR-lincRNA evolution across mammals and primates. Related to Figure 2. Distribution of the average phastCons score of TR-lincRNA (red), other LCL-expressed lincRNA (grey), and protein-coding gene (green) for A) introns in placental mammals, (B) exons and (C) introns in primates, and putative promoter regions (+/-1 Kb, +2/-1 Kb, and +/- 500 bp from TSS) in D,F,H) placental mammals and E,G,I) primates. Differences between groups were tested using a two-tailed Mann-Whitney *U* test and *p*-values are indicated.

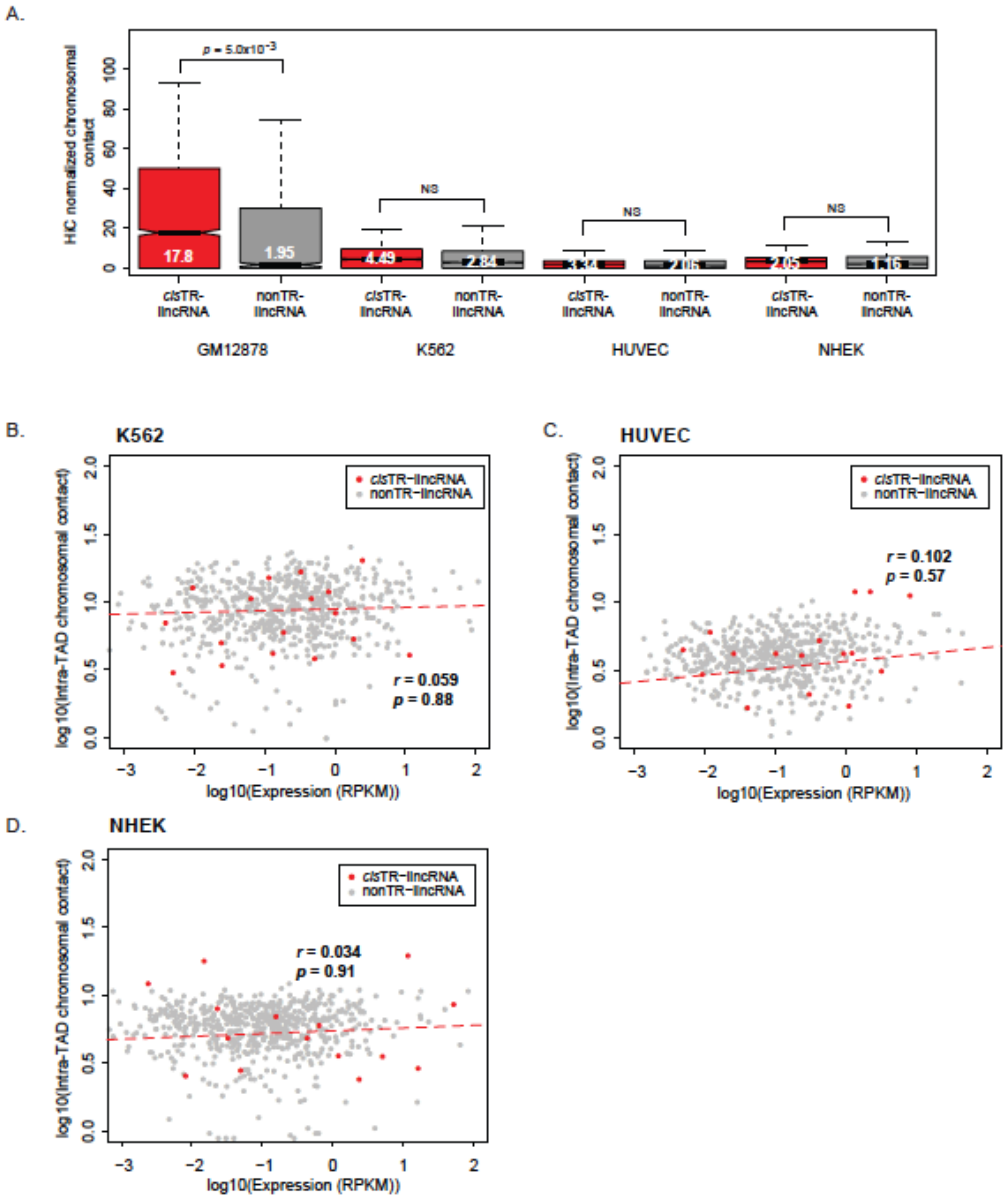

Figure S4. TR-lincRNAs regulate proximal TR-pcgenes in *cis* likely by modulating chromatin architecture. Related to Figure 4. (A) Average chromosomal contacts within TAD containing *cis*TR-lincRNAs (red) and other LCL-expressed lincRNAs (grey) in GM12878, K562, HUVEC and NHEK cell lines. Differences between groups were tested using a two-tailed Mann-Whitney *U* test and *p*-values are indicated. (B-D) Correlations (Spearman's) between expression levels of *cis*TR-lincRNAs ($p>0.05$, red) and other LCL-expressed lincRNAs ($p>0.05$, grey) with the average chromosomal contacts within their containing TADs in K562, HUVEC, and NHEK cell lines.

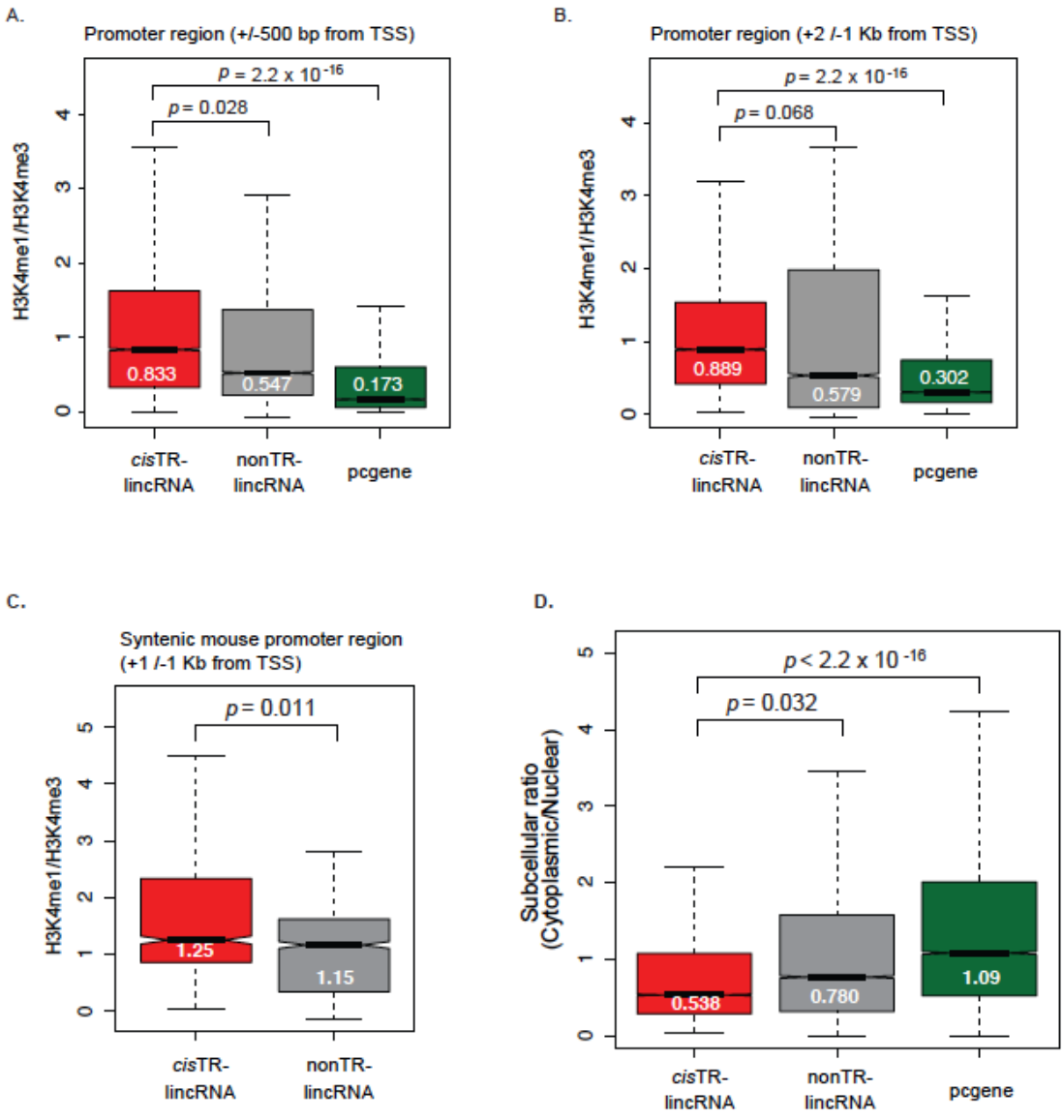

Figure S5. TR-lincRNAs promoter regions are enriched in enhancer-associated chromatin marks. Related to Figure 5. (A-C) Ratio of the number of H3K4me1 to H3K4me3 sequencing reads mapped to the putative promoter regions [A] 500 bp upstream and downstream of TSS and B) 2 Kb upstream and 1 Kb downstream in human GM12878 LCLs and C) 1 Kb upstream and downstream of TSS in mouse CH12 LCLs] and D) subcellular localization ratio (cytoplasmic/nuclear) in LCLs (GM12878) for TR-lincRNAs (red), other LCL-expressed lincRNAs (grey), and protein-coding genes (green). Differences between groups were tested using a two-tailed Mann-Whitney *U* test and *p*-values are indicated.

Table S1. Related to Figure 1. TR-lincRNAs and TR-pcgenes with their associated GWAS *cis*-eQTLs and regulatory trait concordance (RTC) scores.

Table S2. Related to Figure 1. Co-expression of TR-lincRNAs and TR-pcgenes with trait-relevant genes, as predicted by Pascal.

Table S3. Related to Figure 2. Fold enrichment in repeat elements within TR-lincRNA exons and putative promoter regions relative to other LCL-expressed lincRNAs.

Table S4. Related to Figure 5. Coordinates of the genomic loci of TR-lincRNAs and their proximal TR-pcgenes linked to the same traits.

Table S5. Related to Figure 5. Linear regression analysis results of TR-lincRNAs with proximal TR-pcgenes that are associated with the same complex trait or disease through *cis*-eQTLs.

Table S6. Related to Figure 3. Regulatory trait concordance (RTC) score and correlation in expression levels between pairs of TR-lincRNAs and their proximal TR-pcgene(s) that are associated with the same complex trait variant, relative to randomly shuffled lincRNA expression.

Table S7. Related to Figure 3. Copy number variations (CNVs) that uniquely encompass *cis*TR-lincRNAs and TR-pcgenes.

Supplemental Experimental Procedures

RNA sequencing and genotype data

Mapped RNA sequencing reads of EBV-transformed lymphoblastoid cell lines (LCLs) derived from 373 individuals of European descent (CEU, GBR, FIN and TSI) and the corresponding processed genotypes were downloaded from EBI ArrayExpress (accession E-GEUV-1) (Lappalainen et al., 2013). Only single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) greater than 5% were considered in the eQTL analysis.

lincRNA and protein-coding gene expression quantification

Mapped RNA sequencing reads from the ENCODE GM12878 cell line were assembled *de novo* using Cufflinks v2.1.1. Transcripts with a) no overlap with ENSEMBL build 70 protein-coding genes, b) longer than 200 nucleotides and c) with no coding potential as predicted by CPC (Kong et al., 2007) were annotated as *de novo* LCL-expressed lincRNAs.

The number of RNA sequencing reads overlapping lincRNAs (GENCODE version 19 and *de novo* LCL-expressed lincRNAs) and protein-coding genes (GENCODE version 19) was estimated using HTSeq (version 0.6.1, default parameters) (Anders et al., 2015). We estimated the expression level of each gene in each sample as the total number of reads per kilobase per million mapped reads (RPKM) mapping to the total number of exonic nucleotides of the gene. Only genes quantified (RPKM>0) in more than half of LCL samples were considered in the remainder of the analysis (14,846 protein-coding genes and 1,510 lincRNAs).

Potential technical variation across samples were regressed out using PEER (Stegle et al., 2012) as described previously (Lappalainen et al., 2013). The PEER-corrected expression values were transformed to follow a centered and standardized normal distribution using the *rntransform* function from the GenABEL R package (Aulchenko et al., 2007) as described previously (Lappalainen et al., 2013).

To predict subcellular localization, we estimated the number of poly(A)-selected RNA sequencing reads derived from nuclear and cytoplasmic fractions of human LCLs (GM12878) that mapped to exons of lincRNAs and protein-coding genes, as described above (Encode Project Consortium, 2012). Only genes expressed > 0.3 RPKM in both the cytoplasmic and nuclear fractions of the cells were considered in the analysis (Ramskold et al., 2009).

Cis-eQTL analysis

Expression quantitative trait locus (eQTL) analysis was performed for genome-wide significant ($p<5x10^{-8}$ (Welter et al., 2014)) trait-associated autosomal SNPs located within a 2 Mb window centered on the predicted transcription start site (TSS) of each expressed lincRNA and protein-coding gene.

We estimated Pearson's correlation (r_{obs}) between gene expression levels (PEER-corrected and standard normal distribution-transformed) and trait-associated SNP genotypes. To assess the significance of the correlations globally, we permuted the expression levels of each gene 1000 times and recorded the maximum absolute Pearson correlation (r_{exp}). We considered *cis*-eQTLs with an absolute r_{obs} higher than 95% of r_{exp} values for all possible SNP-gene pairs (FDR 5%) to be significant (Lappalainen et al., 2013). *Cis*-eQTLs mapped to the human leukocyte antigen locus (chr6: 29,523,406-33,3777,701 (Shiina et al., 2009)) were excluded from the study due to the complex genomic architecture at this locus (173 protein-coding and 17 lincRNA *cis*-eQTLs).

Generation of gene expression/length matched data set

To account for differences in expression levels and length between lincRNAs and protein-coding genes, we identified a random subset of protein-coding genes with matched expression levels and length to lincRNAs. We divided all human genes into two sets of 10 equally sized bins based on their expression levels and length, independently. For each lincRNA, protein-coding genes were randomly drawn without replacement from the intersection of their expression-level-matched and a length-matched gene bins.

Regulatory Trait Concordance

We calculated the regulatory trait concordance (RTC) score for each GWAS *cis*-eQTL. As described previously (Nica et al., 2010), RTC is measured by identifying the ranking of the GWAS *cis*-eQTL correlations (absolute Pearson's correlation) relative to those calculated for all other SNP eQTLs (dbSNP build 142 (Rosenbloom et al., 2015)) in the 2 Mb window centered on the gene's TSS.

Enrichment analysis of GWAS traits

Human complex traits-associated with *cis*-eQTLs were obtained from the NHGRI GWAS catalogue (Welter et al., 2014) as described above. Hypergeometric tests were performed for all individual traits, followed by Benjamini and Hochberg multiple testing correction.

Co-expression analysis

Genes implicated in human complex traits were predicted using the *Pathway scoring algorithm* "Pascal" (Lamparter et al., 2016) with default parameters. As input, we used all other trait-associated SNPs that reach a genome-wide significance threshold ($p<5x10^{-8}$), excluding that of the GWAS *cis*-eQTL, deposited in the NHGRI GWAS catalogue (Welter et al., 2014) and the LD structure of the European population (1000 Genomes Project Consortium, 2012).

We determined the absolute pairwise Pearson's correlation in expression in LCLs between TR-lincRNAs and all genes associated with the same trait (according to Pascal). We compared these values to what would be expected based on the absolute median pairwise correlation between the trait-relevant genes and 1,000 randomly permutated lincRNA expression levels.

Median absolute Pearson's correlation in expression was calculated between all LCL-expressed lincRNAs and protein-coding genes, as well as pairs of protein-coding genes, located within < 20 Kb, 20 Kb-100 Kb, 100 Kb-500 Kb, and 500 Kb-2 Mb of each other.

Impact of copy number variation on gene expression

Copy number variants (CNVs) (1000 Genomes Project Consortium, 2012) that encompass pairs of *cis*TR-lincRNAs or TR-pgenes that share the same *cis*-eQTL were obtained to assess the impact of the changes in TR-lincRNA or TR-pcgene copies on the expression levels of their nearby associated TR-pcgene and TR-lincRNA, respectively. After excluding CNVs that overlap the shared GWAS *cis*-eQTL or those that contain both the linked *cis*TR-lincRNA and TR-pcgene, we obtained 103 samples with a total of 4 CNVs that encompassed *cis*TR-lincRNAs and 88 samples with 3 CNVs that overlapped 5 TR-pcgenes (Table S7).

Enhancer-associated TR-lincRNAs

Coordinates of ENCODE-predicted enhancer elements, H3K4me1 and H3K4me3 ChIP sequencing reads in human GM12878 and mouse CH12 LCLs (Encode Project Consortium, 2012; Mouse Encode Consortium, 2012) were downloaded from the UCSC database (Rosenbloom et al., 2015). We estimated the ratio of H3K4me1 to H3K4me3 reads mapping to putative promoter regions of lincRNAs (using HTseq version 0.6.1 (Anders et al., 2015)). All analyses were performed using three definitions of putative promoter regions (2 Kb upstream to 1 Kb downstream of TSS, 1 Kb upstream and downstream of TSS, and 500 bp upstream and downstream of TSS). Syntenic regions of lincRNA putative promoter regions (1 Kb upstream and downstream of TSS) in mouse was obtained using liftOver (Meyer et al., 2013) with parameters: -minMatch = 0.2 -minBlocks = 0.01. Regions within the ENCODE Data Analysis Consortium Blacklisted Regions (Hoffman et al., 2013) were excluded from this analysis.

Linear regression analysis of enhancer-associated TR-lincRNAs

We used hierarchical regression to test whether adding the expression of the TR-lincRNA ($\text{TR-pcgene}_{expr} \sim \text{SNP}_{GWAS} + \text{TR-lincRNA}_{expr}$) improves on the *cis*-eQTL regression model ($\text{TR-pcgene}_{expr} \sim \text{SNP}_{GWAS}$) by assessing the difference in the proportion of variance explained by the models.

Genome-wide enrichment of genetic elements

Enrichment or depletion of TR-lincRNA overlaps with genetic elements, relative to the expectation, were estimated using the Genome Association Tester (GAT) (Heger et al., 2013). GAT tests for enrichment by comparing the observed overlap against a null distribution obtained by randomly sampling 10,000 times (with replacement) segments of the same length and matching GC content as the tested loci within mappable intergenic regions of the hg19 genome (Encode Project Consortium, 2012) or the location of all LCL-expressed lincRNA exons. To control for potential confounding variables that correlate with GC content, such as gene density, the genome was divided into segments of 10 Kb and assigned to eight isochore bins in the enrichment analysis. The GC content of each bin was as follows: 0-36, 36-38, 38-40, 40-42, 42-44, 44-46, 46-48, 48-100 (Heger et al., 2013).

Enrichment of repeat elements (Jurka et al., 2005) were tested across TR-lincRNA promoters and exons. To test for preferential TR-lincRNA location relative to TR-pcgenes, we defined TR-pcgene territories (genomic regions containing all nucleotides that are closer to the TR-pcgene of interest than they are to its most proximal up- and downstream protein-coding genes) and tested for TR-lincRNA nucleotide enrichment in these regions. Binding sites of cohesin (union of RAD21 and SMC3 sites) and CTCF binding sites were obtained from ENCODE (Encode Project Consortium, 2012). Topologically associating domains (TADs) that do not completely overlap other smaller TADs in GM12878 (Rao et al., 2014) were divided into 10 equal sized segments and tested for enrichment of TR-lincRNAs. A TAD boundary is defined as 20% of the TAD's length inwards from the TAD's border.

Spatial chromosomal architecture analysis

Intra-chromosome interactions were calculated using Hi-C contact matrices for four ENCODE cell lines, GM12878, K562, HUVEC, and NHEK (Rao et al., 2014). All computations were performed on 5 Kb resolution matrices with a MAPQ score above 30. The raw data were normalized using the KR normalization vectors except for chr 9, where SQRTVC normalization was used as the KR normalization algorithm failed to converge (Rao et al., 2014). Average intra-chromosomal contact was estimated for each TAD that encompasses the gene loci of interest. Spearman's correlation was estimated between gene expression levels and the average density of contact within the TAD where the gene resides. Comparisons between Spearman's correlations was performed using the two-sided Fisher's *z* test (1925) based on independent groups implemented in the "cocor" R package (Diedenhofen and Musch, 2015).

Evolutionary rate and sequence conservation analysis

The ancestral allele and the frequency of each polymorphic site in the European population were obtained from the 1000 Genomes Project (1000 Genomes Project Consortium, 2012) and were used to determine derived allele frequency (DAF), as previously described (Haerty and Ponting, 2013). The DAF spectrum was determined using all human common SNPs (dbSNP build 142) mapped within lincRNA and protein-coding gene exons, introns and putative promoter regions. As a control, we compared the estimated DAF spectrum to that of local ancestral repeats (ARs – transposable elements shared between human and mouse) within a 2 Mb window centered on each lincRNA's TSS, which is used as a proxy for neutrally-evolving sequences. Human and mouse transposable elements were downloaded from RepBase build 21 (Jurka et al., 2005). Human repeat elements whose syntenic regions in mouse (obtained using liftOver (Meyer et al., 2013) with parameters: -minMatch = 0.2 -minBlocks = 0.01) also overlapped a murine transposable element by at least 1 bp were used as ARs.

PhastCons scores computed using the multiple alignments of 45 vertebrate genomes to the human genome (hg19) were downloaded from the UCSC database (Rosenbloom et al., 2015) for placental mammals and primates (Siepel et al., 2005). We calculated the average phastCons score across lincRNA and protein-coding gene exons, putative promoter regions, and local ARs.

Statistical tests

All statistical analyses were performed using the R software environment for statistical computing and graphics (R Development Core Team, 2008).

Supplemental References

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Aulchenko, Y.S., Ripke, S., Isaacs, A., and van Duijn, C.M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics *23*, 1294-1296.

Diedenhofen, B., and Musch, J. (2015). cocor: a comprehensive solution for the statistical comparison of correlations. PloS one *10*, e0121945.

Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Haerty, W., and Ponting, C.P. (2013). Mutations within lncRNAs are effectively selected against in fruitfly but not in human. Genome biology *14*, R49.

Heger, A., Webber, C., Goodson, M., Ponting, C.P., and Lunter, G. (2013). GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics *29*, 2046-2048.

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E.*, et al.* (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic acids research *41*, 827-841.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research *110*, 462-467.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic acids research *35*, W345-349.

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLoS computational biology *12*, e1004714.

Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G.*, et al.* (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506-511.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B.*, et al.* (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic acids research *41*, D64-69.

Mouse Encode Consortium (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome biology *13*, 418.

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS genetics *6*, e1000895.

R Development Core Team (2008). R: A language and environment for statistical computing. (R Foundation for Statistical Computing).

Ramskold, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS computational biology *5*, e1000598.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S.*, et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M.*, et al.* (2015). The UCSC Genome Browser database: 2015 update. Nucleic acids research *43*, D670-681.

Shiina, T., Hosomichi, K., Inoko, H., and Kulski, J.K. (2009). The HLA genomic loci map: expression, interaction, diversity and disease. Journal of human genetics *54*, 15-39.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S.*, et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research *15*, 1034-1050.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature protocols *7*, 500-507.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L.*, et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research *42*, D1001-1006.