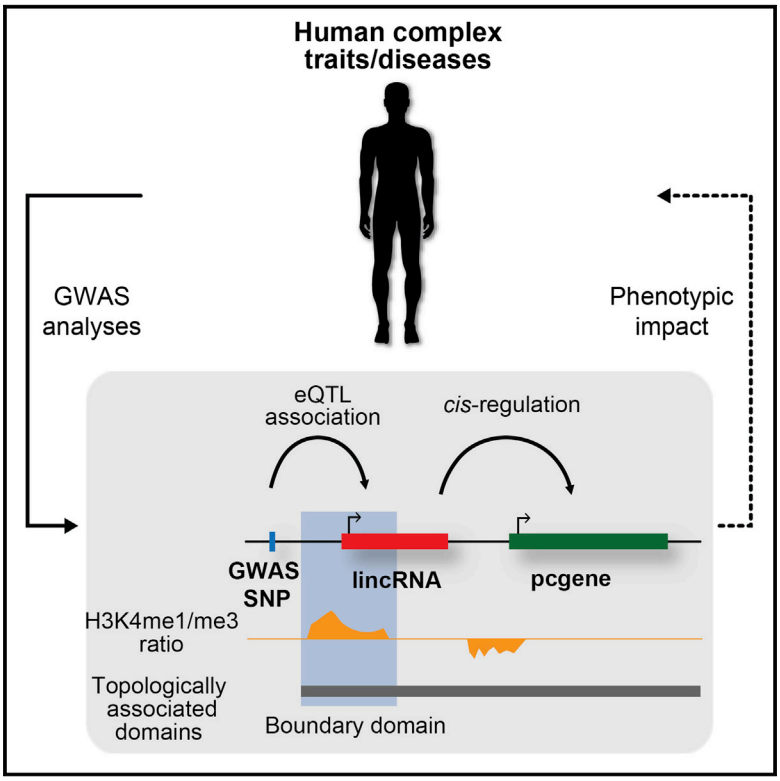
Cell Reports

Resource

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

Graphical Abstract



Authors

Jennifer Yihong Tan, Adam Alexander Thil Smith, Maria Ferreira da Silva, ..., Zoltán Kutalik, Sven Bergmann, Ana Claudia Marques

Correspondence

jennifer.tan@unil.ch (J.Y.T.), anaclaudia.marques@unil.ch (A.C.M.)

In Brief

Tan et al. identify and characterize 69 human complex trait/disease-associated lincRNAs in LCLs. They show that these loci are often associated with *cis*-regulation of gene expression and tend to be localized at TAD boundaries, suggesting that these lincRNAs may influence chromosomal architecture.

Highlights

- We identify 69 lincRNAs associated with human complex traits (TR-lincRNAs)
- TR-lincRNAs are conserved in humans and interact with other disease-relevant loci
- TR-lincRNAs often associate with *cis*-regulation of proximal protein-coding gene expression
- TR-lincRNAs are enriched at TAD boundaries and may modulate chromatin architecture

Tan et al., 2017, Cell Reports *18*, 2280–2288
February 28, 2017
http://dx.doi.org/10.1016/j.celrep.2017.02.009

CellPress

OPEN ACCESS
CellPress

Cell Reports
Resource

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

Jennifer Yihong Tan,[1,2,*] Adam Alexander Thil Smith,[1,2] Maria Ferreira da Silva,[1,2] Cyril Matthey-Doret,[1,2] Rico Rueedi,[2,3] Reyhan Sönmez,[2,3] David Ding,[4] Zoltán Kutalik,[3,5] Sven Bergmann,[2,3] and Ana Claudia Marques[1,2,6,*]
[1]Department of Physiology, University of Lausanne, 1015 Lausanne, Switzerland
[2]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland
[3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[4]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[5]Institute of Social and Preventive Medicine, University Hospital Lausanne (CHUV), 1011 Lausanne, Switzerland
[6]Lead Contact
*Correspondence: jennifer.tan@unil.ch (J.Y.T.), anaclaudia.marques@unil.ch (A.C.M.)
http://dx.doi.org/10.1016/j.celrep.2017.02.009

SUMMARY

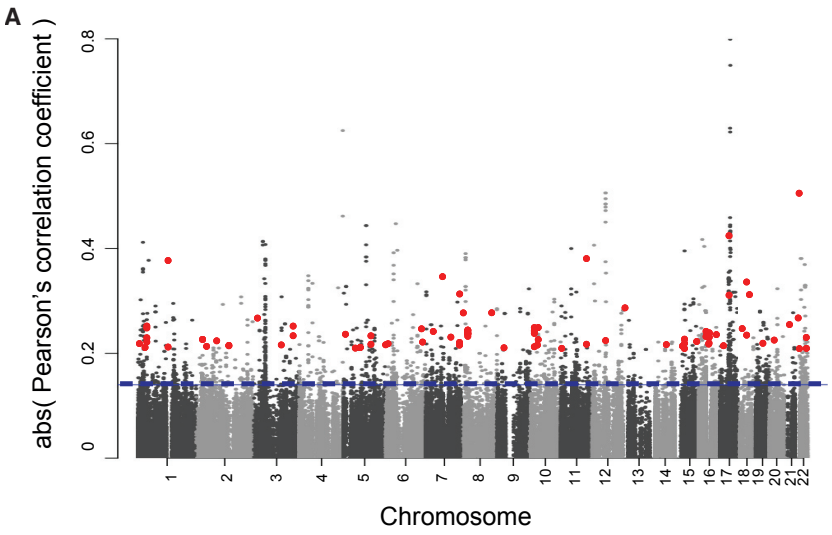
Intergenic long noncoding RNAs (lincRNAs) are the largest class of transcripts in the human genome. Although many have recently been linked to complex human traits, the underlying mechanisms for most of these transcripts remain undetermined. We investigated the regulatory roles of a high-confidence and reproducible set of 69 trait-relevant lincRNAs (TR-lincRNAs) in human lymphoblastoid cells whose biological relevance is supported by their evolutionary conservation during recent human history and genetic interactions with other trait-associated loci. Their enrichment in enhancer-like chromatin signatures, interactions with nearby trait-relevant protein-coding loci, and preferential location at topologically associated domain (TAD) boundaries provide evidence that TR-lincRNAs likely regulate proximal trait-relevant gene expression in *cis* by modulating local chromosomal architecture. This is consistent with the positive and significant correlation found between TR-lincRNA abundance and intra-TAD DNA-DNA contacts. Our results provide insights into the molecular mode of action by which TR-lincRNAs contribute to complex human traits.

INTRODUCTION

An increasing number of reports suggest that long intergenic noncoding RNAs (lincRNAs), which were previously regarded as "junk RNA" (Hüttenhofer et al., 2005), can contribute to normal and disease phenotypes in humans (Esteller, 2011). For example, candidate screens followed by detailed functional characterization of a few individual trait-associated lincRNAs illustrate how genetic variants affecting the lincRNA sequence can underlie human complex traits (Ishii et al., 2006; Zheng et al., 2016). Recently, RNA capture followed by sequencing in multiple disease-associated protein-coding gene deserts led to the identification of lowly and tissue-specifically expressed lincRNA loci (Mercer et al., 2014). Detailed experimental analysis of these lincRNA candidates is now required to establish whether and how these loci contribute to disease.

Although thousands of common genetic variants have been associated with complex human traits through genome-wide association studies (GWASs), only a small proportion fall within exonic coding sequences (Hindorff et al., 2009; Maurano et al., 2012). Instead, most GWAS variants map within noncoding regulatory regions that are enriched in population and tissue-specific expression quantitative trait loci (eQTLs) (Edwards et al., 2013). eQTL analysis has previously led to the identification of protein-coding genes and pathways that are disrupted in human complex traits (for example, Emilsson et al., 2008; Fairfax et al., 2012; Gilad et al., 2008). Recently, lincRNAs whose expression correlate with GWAS variants were also identified using this approach (Kumar et al., 2013; Lappalainen et al., 2013; McDowell et al., 2016; Popadin et al., 2013), suggesting that the transcription or the transcripts arising from lincRNA loci in eQTLs with GWAS variants may similarly contribute to phenotypes. Although a handful of studies have investigated the relationship between individual lincRNAs with risk-variant-associated expression and their linked traits (for example, Ishii et al., 2006; Jendrzejewski et al., 2012), the underlying mechanism of action for most remains undetermined.

So far, functionally characterized lincRNAs have been implicated in both transcriptional and post-transcriptional regulation of local or distal genes (Vance and Ponting, 2014). We have previously shown that chromatin signatures at lincRNA transcriptional start sites allow the distinction between these two regulatory classes (Marques et al., 2013). Specifically, the expression of lincRNAs arising from regulatory elements that carry enhancer-like chromatin signatures correlates with neighboring protein-coding gene abundance, suggesting that transcription at these loci contributes to local regulation of expression (Marques et al., 2013). Interestingly, eQTL GWAS variants are enriched within enhancer regions (Ernst et al., 2011; Schaub



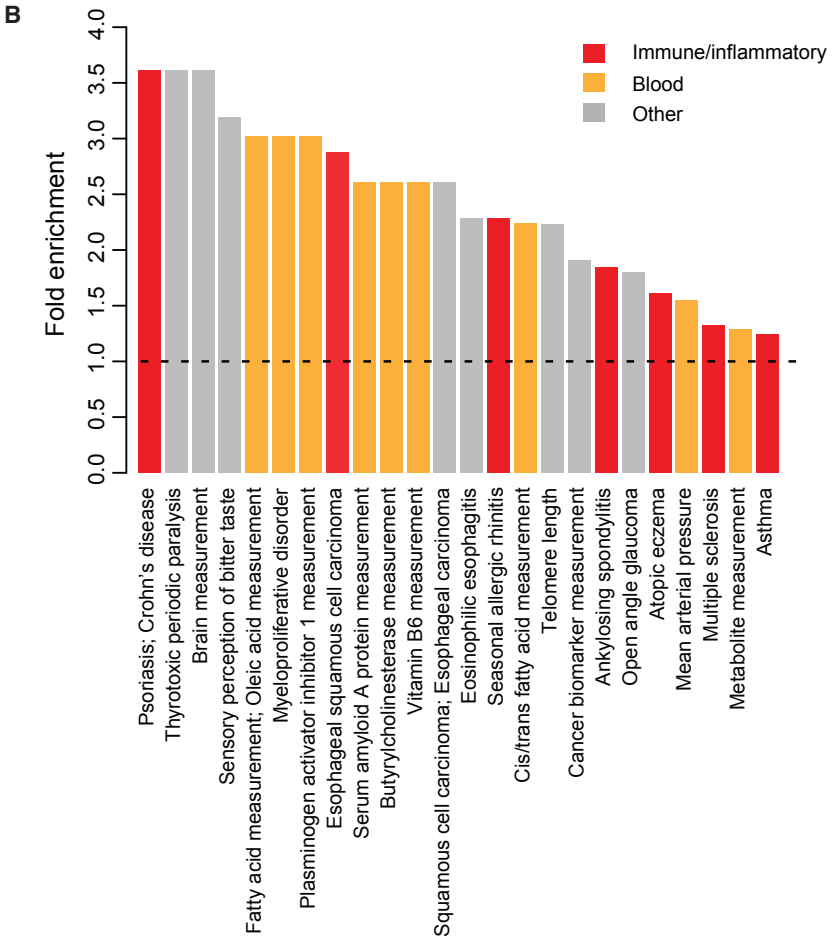


Figure 1. Identification of GWAS *cis*-eQTLs for lincRNAs and Protein-Coding Genes

(A) Manhattan plot showing absolute Pearson's correlation coefficient (*r*) calculated for all possible GWAS *cis*-eQTL associations with LCL-expressed lincRNAs (TR-lincRNAs) and protein-coding genes (TR-pcgenes) across human autosomes. Significance cutoff is represented by a horizontal dashed line (absolute *r* of 0.145). Significant TR-lincRNA *cis*-eQTLs are highlighted in red.

(B) The GWAS human complex traits that are significantly enriched (fold-enrichment, $p < 0.05$, hypergeometric test) within genome-wide significant *cis*-eQTLs (TR-lincRNAs + TR-pcgenes), relative to all possible GWAS *cis*-eQTL associations. Traits are grouped into immune/inflammatory responses (red), blood-related traits (orange), and others (gray).

See also Figure S1 and Tables S1 and S2.

et al., 2012), suggesting a link between enhancer-associated lincRNAs and complex human traits.

Here, we used functional, evolutionary, and population genomics to extensively characterize the regulatory interactions between a high-confidence set of trait-associated lincRNAs and protein-coding genes identified through GWAS *cis*-eQTL analysis. Our results demonstrate that most human complex-trait-associated lincRNAs arise from enhancer-like regions and are frequently located at the boundaries of topologically associated domains (TADs), which have been previously shown to contribute to chromosomal architecture and gene transcription regulation (Rao et al., 2014). Together, these findings support that the transcription of trait-relevant lincRNAs contributes to chromosomal architecture and thereby the regulation of nearby trait-associated protein-coding gene expression levels.

RESULTS

Identification of Trait-Relevant lincRNAs and Protein-Coding Genes

We considered all lymphoblastoid cell line (LCL)-expressed de novo (Experimental Procedures) and GENCODE-annotated loci with at least one genome-wide significant ($p < 5 \times 10^{-8}$) GWAS SNP (7,451 GWAS SNPs) (Welter et al., 2014) in their vicinity (Experimental Procedures). We calculated the Pearson's correlation between the expression of these coding and noncoding loci and the corresponding genotype of their neighboring GWAS SNPs in a panel of 373 LCLs derived from individuals of European descent (Lappalainen et al., 2013). This led to the identification of 111 and 1,479 GWAS *cis*-eQTLs significantly correlated (false discovery rate [FDR] < 5%; Experimental Procedures) with the expression levels of 73 lincRNAs and 756 protein-coding genes, respectively (Figure 1A). We asked whether differences in length and expression level (Figure S1) between lincRNAs and mRNAs would account for

the relatively lower number of eQTL-lincRNAs. After restricting our analysis to length- and expression-matched mRNAs, we found that the proportion of eQTL-lincRNAs (2.9%) is statistically indistinguishable from that of eQTL-mRNAs (3.2% of size- and expression level-matched mRNAs; $p = 0.68$, two-tailed χ^2 test), suggesting that lincRNA properties indeed limit the power to identify lincRNA-eQTLs. Despite the restricted power in lincRNA *cis*-eQTL detection, most of the identified GWAS lincRNA *cis*-eQTLs (68%; Table S1) could be replicated using data from an independent set of LCLs, derived from 555 individuals of European descent from the Lausanne population (Cohorte Lausannoise [CoLaus]; Firmann et al., 2008). The proportion of replicated lincRNA associations is similar to what was found for mRNA *cis*-eQTLs (71%, $p = 0.69$, two-tailed Fisher's exact test), corroborating the robustness of our *cis*-eQTL findings.

Evidence that these GWAS *cis*-eQTLs are enriched in immune/inflammatory response and blood-related traits, including metabolite levels (Figure 1B), suggests that despite known limitations (Choy et al., 2008), lymphoblastoid cells are suitable to investigate the contributions of lincRNA loci to human complex traits.

Genetic variants do not segregate randomly in the human population and SNPs found within the same linkage disequilibrium (LD) block are likely to correlate, to some extent, with the expression levels of all gene loci within the same LD block, leading to false-positive *cis*-eQTL associations between GWAS SNPs and gene expression (Stranger et al., 2007). To address this issue, we used regulatory trait concordance (RTC), an empirical method that accounts for local LD structure (Nica et al., 2010). We estimated the rank of the identified GWAS *cis*-eQTL among all nearby common SNPs based on decreasing absolute correlation with gene expression, thus assessing the likelihood that the identified *cis*-eQTL is most likely driven by the complex-trait-associated genetic variant and not due to local LD with another SNP. This approach does not exclude, however, that the expression of the coding or noncoding loci could be under the influence of an unknown variant in linkage with the GWAS *cis*-eQTL. After applying a previously tested RTC threshold (0.9) to identify high-confidence eQTL associations (Nica et al., 2010), we obtained 69 lincRNAs that are likely true trait-relevant gene candidates (trait-relevant lincRNAs [TR-lincRNAs]), as well as 723 protein-coding genes (TR-pcgenes; Table S1). Importantly, 73% of the GWAS *cis*-eQTLs associated with TR-lincRNAs and TR-pcgenes were validated in CoLaus, a significant 11% increase in replication rate from all identified *cis*-eQTLs ($p < 0.05$, two-tailed Fisher's exact test), reinforcing the reliability of this set.

TR-lincRNAs are likely involved in pathways relevant to their associated traits. Specifically, we asked whether the expression levels of trait-relevant loci are correlated with those of other genes associated with the same trait, as would be expected if they contribute to the same phenotype. For each trait-relevant loci, we used the pathway scoring algorithm "Pascal" (Lamparter et al., 2016) to identify all loci located within LD blocks containing other significant GWAS ($p < 5 \times 10^{-8}$) variants for that trait, and we tested for their co-expression with the *cis*-eQTL loci candidates, a surrogate for genetic interaction. We found that 83% of TR-lincRNAs (57/69) are significantly co-expressed ($p < 0.05$, permutation test; Experimental Procedures) with genes associated with the same trait, a proportion similar to that found for TR-pcgenes (89% [642/723], $p = 0.17$, two-tailed Fisher's exact test; Table S2).

Trait-Relevant lincRNAs Are Conserved in Humans

The biological relevance of lincRNA transcription is generally unclear, and there is ongoing debate as to whether it is the transcript or the act of transcription that underlies the function of most noncoding loci (Wilusz et al., 2009). Evolutionary analyses can provide initial insights into this question, as selective constraint at exons would not be required if it is the act of transcription and not the transcript sequence that underlies function.

We investigated the evolution of TR-lincRNAs' exons in humans and found that they exhibit a significantly higher proportion of low-frequency alleles (derived allele frequency [DAF] < 0.1) compared to local neutrally evolving sequences (ancestral repeats [ARs]), TR-lincRNA intronic regions, and other LCL-expressed lincRNA exons ($p < 0.05$, two-tailed Fisher's exact test; Figure 2A). The proportion of SNPs with DAF < 0.1 found within TR-lincRNA and protein-coding gene exons is statistically indistinguishable ($p = 0.56$, two-tailed Fisher's exact test; Figure 2A). This is in contrast to exons of all LCL-expressed lincRNAs, which have a similar proportion of low derived allele frequency polymorphic sites as local ARs ($p = 0.15$, two-tailed Fisher's exact test; Figure S2A), consistent with previous analyses (Haerty and Ponting, 2013). No statistically significant difference in derived allele frequency was observed between introns and exons of all LCL-expressed lincRNAs ($p = 0.89$, two-tailed Fisher's exact test; Figure S2A). Our results indicate that purifying selection has acted to remove deleterious mutations within TR-lincRNA exons during recent human evolution, which reinforces the functional relevance of these noncoding transcripts in humans. Surprisingly, analysis of putative promoters of TR-lincRNAs suggests that these regions evolved neutrally or nearly neutrally (Figure S2B). The difference in evolutionary constraint between the promoter and exon sequences can likely be explained by inaccurate prediction of proximal promoter regions, which would result in reduced power to infer their constraint. Despite limitations, our analysis of exonic sequence evolution supports that TR-lincRNA transcripts were preserved during recent human evolution.

Unexpectedly, the higher selective constraint observed for TR-lincRNAs relative to other LCL-expressed lincRNAs appears to be an evolutionary signature specific to recent human evolution, as we found no significant differences in their sequence conservation during either mammalian or primate evolution, estimated using phastCons scores, a measure of nucleotide conservation (Siepel et al., 2005) (Figures 2B and S3). Specifically, relative to other LCL-expressed lincRNAs, TR-lincRNA exons, introns, and promoters exhibit statistically indistinguishable median phastCons scores (Figure S3). This observation could be the result of rapidly evolving repetitive elements within TR-lincRNAs (Kapusta et al., 2013; Kelley and Rinn, 2012). Indeed, we found that TR-lincRNA exons and promoters are enriched in long terminal repeat (LTR)-derived transposable elements relative to other LCL-expressed lincRNAs (3.8- to 7.9-fold enrichment, $p < 0.05$). In particular, TR-lincRNAs exons and promoters are enriched in human endogenous retrovirus K (ERVK) LTRs (1.6- to 2.2-fold enrichment,

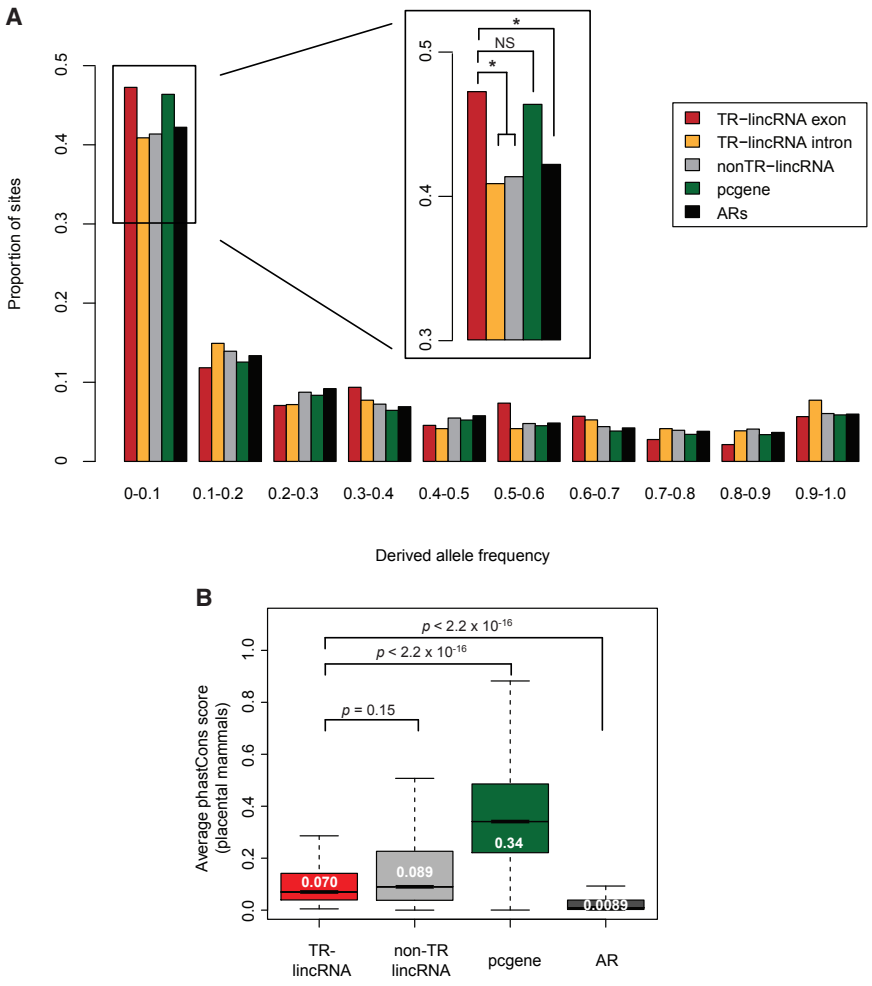


Figure 2. TR-lincRNAs Evolved under Purifying Selection during Recent Human History

(A) Distribution of derived allele frequency (DAF) for variants within exons (red) and introns (yellow) of TR-lincRNA, LCL-expressed lincRNA exons (gray), protein-coding gene exons (green), and ancestral repeats (ARs; black). Low-frequency polymorphic sites (DAF < 0.1) for all classes of genes are depicted in the insert. Asterisks indicate levels of significance in the comparison (*$p < 0.05$; NS, not significant [$p > 0.05$]; two-tailed Fisher's exact test).

(B) Distribution of sequence conservation, as estimated using phastCons scores across placental mammals (y axis), within the exonic sequence of TR-lincRNAs (red), other LCL-expressed lincRNAs (light gray), protein-coding genes (green), and ancestral repeats (dark gray). Differences between groups were tested using a two-tailed Mann-Whitney *U* test, and p values are indicated.

See also Figures S2 and S3 and Table S3.

$p < 0.05$; Table S3; Experimental Procedures), whose transcription was previously shown to be elevated upon immune system stimulation (Manghera and Douville, 2013).

Trait-Relevant lincRNA Transcription Is Associated with *cis* Regulation

lincRNAs can regulate the expression levels of local and distal targets (Vance and Ponting, 2014). To gain insights into the molecular mode of action of TR-lincRNAs, we examined their relationship with TR-pcgenes. For each protein-coding gene, we defined its territory as the genomic region containing all nucleotides that are closer to the gene than they are to its most proximal up- and downstream protein-coding genes. We found that TR-lincRNAs are significantly more likely than expected to reside within TR-protein-coding gene territories (fold enrichment = 2.4, $p < 1 \times 10^{-3}$; Experimental Procedures).

Next, we estimated the median co-expression (Pearson's correlation) in LCLs between pairs of TR-lincRNAs and protein-coding genes in their vicinity (within <20 kb, 20–100 kb, 100–500 kb, and >500 kb of each other). Consistent with their proposed regulatory interactions, we found TR-lincRNAs to be significantly more highly correlated in expression with nearby protein-coding genes than other LCL-expressed lincRNAs (Figure 3A). Furthermore, TR-lincRNAs are over 2.5 times more likely to share an eQTL with at least one nearby protein-coding gene (43/69 [62.3%]) compared to other LCL-expressed lincRNAs (592/2441 [24.3%]), a significantly higher proportion ($p < 1 \times 10^{-3}$, two-tailed Fisher's exact test; Experimental Procedures), suggesting that TR-lincRNAs are more likely than other transcripts to affect the expression of nearby loci.

To dissect the regulatory interaction between TR-lincRNAs and their nearby co-expressed TR-pcgenes, we focused on the 30 trait-relevant lincRNAs with nearby TR-pcgenes that share the same GWAS *cis*-eQTL (Table S4; Experimental Procedures), hereafter referred to as *cis*TR-lincRNAs. We tested, using hierarchical linear regression, whether adding the expression levels of the *cis*TR-lincRNA strengthens the *cis*-eQTL association of its linked TR-pcgene (Experimental Procedures). 87% (26/30) of *cis*TR-lincRNAs significantly improves the association between the expression levels of the nearby TR-pcgenes and their trait-associated variants (Table S5). Furthermore, *cis*TR-lincRNA associations with GWAS *cis*-eQTLs relative to common SNPs in the region (median RTC = 0.97) are significantly higher than those for TR-pcgene associations (median RTC = 0.95, $p < 0.05$, two-tailed Mann-Whitney paired *U*-test; Table S6).

To assess how changes in *cis*TR-lincRNA or TR-pcgene copies impact the expression levels of their nearby associated loci, we identified copy-number variants (CNVs; 1000 Genomes Project Consortium et al., 2012) that uniquely encompass either *cis*TR-lincRNAs or TR-pcgenes (Table S7). CNVs that overlap the shared GWAS *cis*-eQTL or those that contain both the linked *cis*TR-lincRNA and TR-pcgene were excluded. We estimated the absolute fold difference in *cis*TR-lincRNA or TR-pcgene

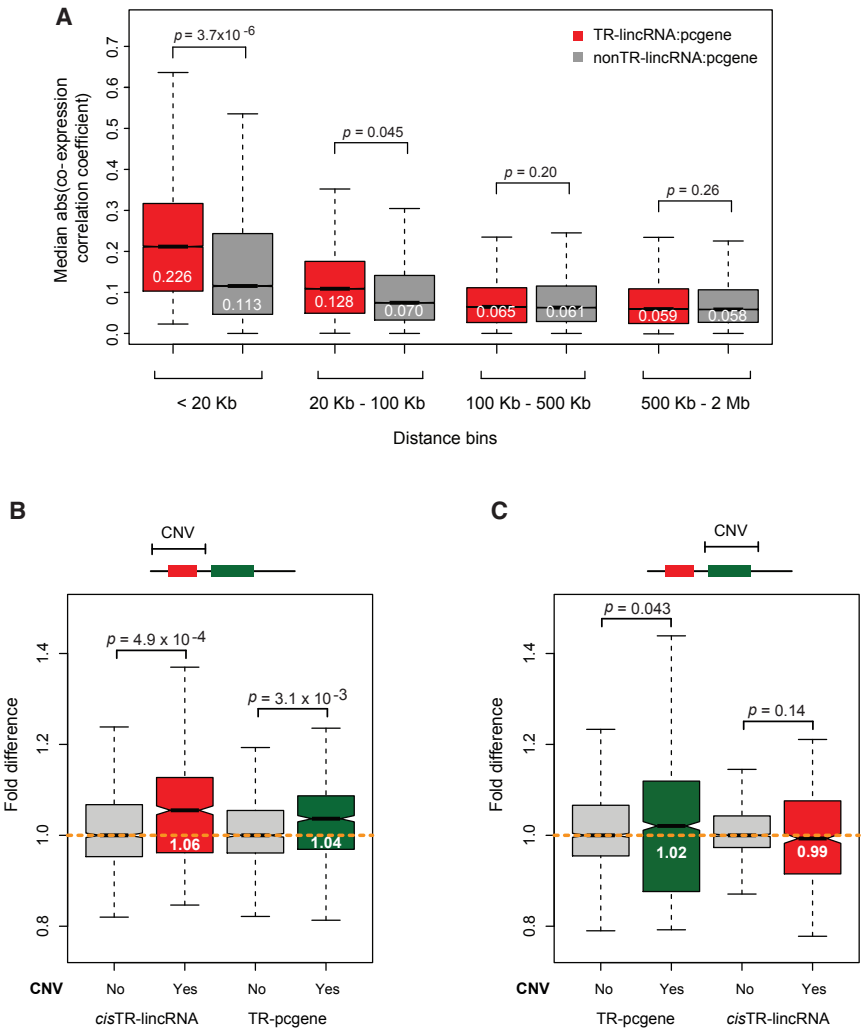


Figure 3. TR-lincRNAs Are Enriched at TAD Boundaries and Regulate Proximal TR-pcgenes in *cis*, Likely by Modulating Chromatin Architecture

(A) Distribution of median absolute correlation coefficient between expression levels in LCLs of TR-lincRNAs (red) or other LCL-expressed lincRNAs (gray) and nearby protein-coding genes. Pairs are split into bins based on their genomic distance (<20 kb, 20–100 kb, 100–500 kb, and 500 kb to 2 Mb).

(B and C) Absolute fold difference in expression levels across individuals that carry copy-number variants (CNVs) (1000 Genomes Project Consortium et al., 2012) that encompass (B) *cis*TR-lincRNAs (red) or (C) TR-pcgenes (green) and that of the nearby trait-relevant protein-coding genes or lincRNAs, respectively, relative to the expression of the loci in individuals without CNVs (gray). Differences between groups were tested using a two-tailed Mann-Whitney *U* test, and p values are indicated.

See also Tables S3, S4, S5, S6, and S7.

expression between individuals with or without CNVs and found that variations in *cis*TR-lincRNA copy number are associated with significant changes in the levels of TR-pcgenes ($p < 0.05$, two-tailed Mann-Whitney *U* test; Figure 3B). In contrast, no significant difference in the levels of *cis*TR-lincRNAs was observed when CNVs encompassed TR-pcgenes ($p = 0.14$, two-tailed Mann-Whitney *U* test; Figure 3C). Together, these observations provide preliminary evidence that *cis*TR-lincRNAs contribute to the regulation of the levels of TR-pcgenes in their vicinities.

Trait-Relevant lincRNAs Are Associated with Local Chromosomal Architecture

TADs are genomic regions where DNA-DNA interactions are frequent (Dixon et al., 2012). These genomic structures have been proposed to modulate gene transcription through increased accessibility to shared local regulatory elements (Nora et al., 2013). This hypothesis is supported by evidence of frequent co-expression between genes within the same TAD (Le Dily et al., 2014; Neems et al., 2016). We investigated whether frequent localization within the same TAD would explain the co-expression between pairs of trait-relevant coding and noncoding loci. First, we found that cisTR-lincRNAs are enriched within LCL TADs that also contain TR-pcgenes (fold enrichment = 3.2, $p < 1 \times 10^{-3}$; Experimental Procedures). Interestingly, when we analyzed the location of *cis*TR-lincRNAs within sub-compartments of TADs, we found them to be significantly enriched at the boundaries and depleted at the center of these genomic units (Figure 4A). Such enrichment at TAD boundaries is specific to *cis*TR-lincRNAs, as no preferential location was found when we analyzed the distribution of other LCL-expressed lincRNAs. To assess the relevance of *cis*TR-lincRNAs to local chromosomal architecture, we investigated the correlation between their expression levels and intra-TAD DNA-DNA contact density (Experimental Procedures). We found that the density of chromosomal contacts is significantly higher for TADs containing *cis*TR-lincRNAs (9.1 times, $p < 5 \times 10^{-3}$, two-tailed Mann-Whitney *U* test; Figure 4B) relative to those containing other LCL-expressed lincRNAs. Interestingly, this difference appears to be specific to LCLs, supporting cell-type-specific functions of *cis*TR-lincRNAs ($p > 0.05$, two-tailed Mann-Whitney *U* test; Figure S4A). Strikingly, we found a significant positive correlation between the levels of *cis*TR-lincRNAs and DNA-DNA contacts within their associated TADs relative to other LCL-expressed lincRNAs ($r = 0.163$, Spearman's correlation, $p < 0.05$; Figure 4C). Importantly, this association is also cell-type-specific and restricted to TR-lincRNAs (Figures S4B–S4D), strongly supporting the role of these loci in the modulation of chromosomal architecture.

Previous studies have demonstrated that active enhancer-like regulatory elements are enriched at the boundaries of TADs (Huang et al., 2015). Interestingly, transcription at these

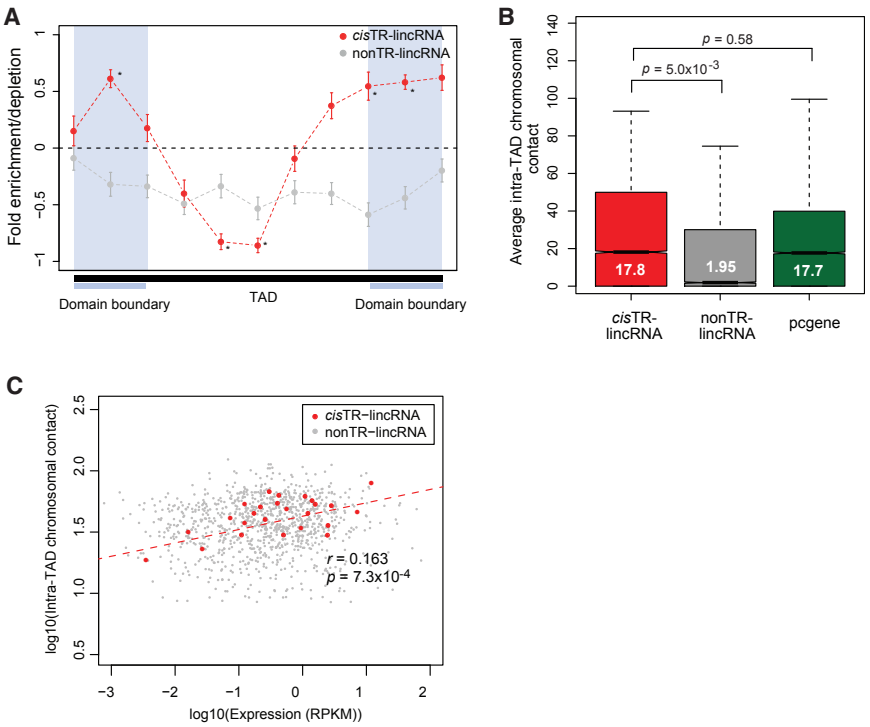


Figure 4. TR-lincRNAs Are Enriched at TAD Boundaries and Regulate Proximal TR-pcgenes in *cis*, Likely by Modulating Chromatin Architecture

(A) Fold enrichment or depletion of *cis*TR-lincRNA (red) and other LCL-expressed lincRNAs (gray) at fractional positions within LCL TADs (GM12878, black bar; Rao et al., 2014) and at TAD boundaries (light blue bar, area shaded in light blue). Significant fold differences are denoted with an asterisk, and SD is shown with error bars ($p < 0.05$, permutation test).

(B) Average chromosomal contacts within TAD that contain *cis*TR-lincRNAs (red), other LCL-expressed lincRNAs (gray), and pcgenes (green) in LCLs (GM12878; ENCODE Project Consortium, 2012). Differences between groups were tested using a two-tailed Mann-Whitney *U* test, and p values are indicated.

(C) Correlation (Spearman's) between expression levels of *cis*TR-lincRNAs ($r = 0.163$, $p = 7.3 \times 10^{-4}$, red) and other LCL-expressed lincRNAs ($r = 0.105$, $p = 0.53$, gray) with the average chromosomal contacts within their residing TADs in LCLs (GM12878; ENCODE Project Consortium, 2012).

See also Figure S4 and Tables S3, S4, S5, and S6.

enhancers is widespread in humans (Andersson et al., 2014), and a large fraction of lincRNA transcription has been previously shown to originate at enhancers (Marques et al., 2013). We investigated whether TR-lincRNAs were enhancer associated. We found that relative to other LCL-expressed lincRNAs, the promoters of *cis*TR-lincRNAs are enriched in mono- versus trimethylation of histone H3K4, a well-established signature of enhancer elements ($p < 0.05$, two-tailed Mann-Whitney *U* test; Figures 5, S5A, and S5B), indicating their likely enhancer origin. Interestingly, we found that the syntenic regions in mouse of our *cis*TR-lincRNA putative promoters are also significantly enriched in enhancer-associated chromatin marks (murine LCLs [CH12 cells]; Mouse ENCODE Consortium et al., 2012) relative to other LCL-expressed lincRNAs ($p < 0.05$, two-tailed Mann-Whitney *U* test; Figure S5C), suggesting their associated enhancer activity is conserved between species at some of these loci. These *cis*TR-lincRNAs are also more enriched in the nucleus versus the cytoplasm relative to other LCL-expressed lincRNAs ($p < 0.05$, two-tailed Mann-Whitney *U* test; Figure S5D), which is as expected and consistent with their role in transcriptional regulation.

The cohesin protein complex, known to be enriched at active enhancer elements and TAD boundaries, has been previously shown to be important for intra-TAD gene regulation in a cell-type-specific manner (Merkenschlager and Odom, 2013). For example, cohesin depletion is associated with disrupted promoter-enhancer interactions within TADs (Kagey et al., 2010; Seitan et al., 2011). Another central player in the regulation of chromatin architecture and gene expression is the CTCF transcription factor (reviewed in Merkenschlager and Odom, 2013). Unlike cohesin, which is involved in cell-specific intra-TAD interactions, CTCF is important for the spatial segregation of topological domains (Zuin et al., 2014) with binding sites that are often conserved and shared across different species and cell types (Kim et al., 2007). We observed that cohesin binding sites are significantly enriched at *cis*TR-lincRNAs loci (fold enrichment = 1.43, $p < 0.05$). In contrast, CTCF binding sites are depleted at these noncoding RNA loci (fold depletion = −0.86, $p < 0.05$; Experimental Procedures) relative to intergenic regions of the human genome. These observations suggest that rather than acting to establish TAD architecture, TR-lincRNAs are more likely to be involved in cell-type-specific regulation of enhancer-promoter interactions within TADs.

Taken together, (1) the positive co-expression of a large proportion of trait-relevant lincRNAs with their proximal TR-pcgenes, (2) the contribution to their nearby TR-pcgene GWAS *cis*-eQTL, (3) enrichment at TAD boundaries and cohesin binding sites, and (4) enrichment in enhancer-like RNA properties are all compatible with enhancer origins and local regulatory roles of TR-lincRNAs.

DISCUSSION

Since the discovery of pervasive lincRNA transcription in humans (Carninci et al., 2005), extensive research efforts have strived to establish what might be their contribution, if any, to organismal phenotypes (Marx, 2014). Previous studies (Kumar et al., 2013; Lappalainen et al., 2013; McDowell et al., 2016; Popadin et al., 2013) have led to the identification of lincRNAs associated with complex human traits and diseases, often through *cis*-eQTL analysis. This wealth of information comes with a new and challenging question: what might be the functions of

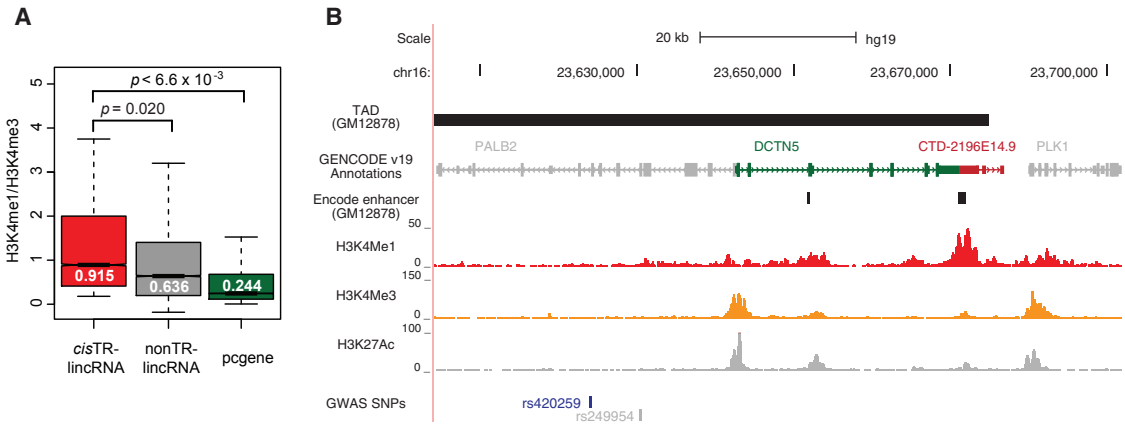


Figure 5. TR-lincRNA Promoter Regions Are Enriched in Enhancer-Associated Chromatin Marks

(A) Ratio of the number of H3K4me1 to H3K4me3 sequencing reads mapped to the putative promoter regions (1 kb upstream and downstream of the TSS) in LCLs (GM12878; ENCODE Project Consortium, 2012) for *cis*TR-lincRNAs (red), other LCL-expressed lincRNAs (gray), and protein-coding genes (green). Differences between groups were tested using a two-tailed Mann-Whitney *U* test, and p values are indicated.

(B) UCSC genome browser view of one *cis*TR-lincRNA, CTD-2196E14.9 (ENSG00000260482, chr16: 23,681,332–23,684,448, red), and a neighboring TR-pcgene, DCTN5 (ENSG00000166847, green), which is associated with the same GWAS *cis*-eQTL (rs420259, blue). Non-trait-associated protein-coding genes between CTD-2196E14.9 and COG7 are colored in gray. Arrows within introns indicate direction of transcription. CTD-2196E14.9 overlaps predicted enhancer elements in a lymphoblastoid cell line (GM12878, vertical black bars; ENCODE Project Consortium, 2012) at the boundary of a TAD (GM12878, horizontal dark gray bar; Rao et al., 2014), and its transcription start site has a high H3K4me1 (red track) over H3K4me3 (yellow track) ratio.

See also Figure S5 and Tables S3, S4, S5, and S6.

these candidates, and how might they contribute to phenotype? Given the heterogeneity of the known molecular mechanisms underlying lincRNA functions and the current lack of approaches to predict them, genetic dissection of these trait-associated candidates is challenging and has only been achieved for a handful of transcripts thus far (for example, Ishii et al., 2006; Jendrzejewski et al., 2012).

Our genome-wide analysis of a stringent set of TR-lincRNAs suggests that these loci often associate with *cis* regulation of nearby trait-associated protein-coding genes and provides a working hypothesis for how lincRNAs can contribute to human complex traits. While co-expression between loci in close genomic proximity is common (McDowell et al., 2016), we show this phenomenon is stronger between TR-lincRNAs and protein-coding genes in their vicinity than between pairs of non-trait-associated loci. Furthermore, we provide evidence that changes in TR-lincRNA copy number are specifically associated with changes in the levels of nearby TR-pcgenes, consistent with the roles of these lincRNAs in the regulation of proximal TR-pcgene expression levels. Recent studies have shown that boundary elements are key to maintaining TAD organization and that mutations in these boundary elements disrupt regulatory interactions and influence phenotypes, specifically during development (Guo et al., 2015; Lupiáñez et al., 2015). The preferential location of TR-lincRNAs at TAD boundaries and their frequent and evolutionarily conserved enhancer origin suggest that TR-lincRNA transcription affects the levels of trait-relevant genes in their vicinity, likely by modulating local chromosomal organization, thus impacting complex normal and disease phenotypes in humans. The correlation observed between TR-lincRNA expression and intra-TAD DNA-DNA interactions in LCLs provides genome-wide support for this hypothesis.

Our results suggest that lincRNAs are generally lowly expressed (Cabili et al., 2011), which is likely to limit their ability to regulate the expression of mRNAs in *trans*. In contrast, regulation of gene expression in *cis* through the modulation of chromosomal architecture is likely to require fewer transcript copies or merely the act of transcription. Therefore, we propose that this mechanism of enhancer-associated lincRNA transcription is likely not restricted to trait-relevant lincRNAs.

While further work is still required to dissect the biological role of individual TR-lincRNAs, our genome-wide results provide the much needed mechanistic insights into their functions, furthering the understanding of the intricate genetic networks underlying complex human traits and diseases.

EXPERIMENTAL PROCEDURES

cis-eQTL Analysis

Mapped RNA-sequencing reads of Epstein-Barr virus (EBV)-transformed LCLs derived from 373 individuals of European descent (Utah Residents with Northern and Western Ancestry [CEU], British in England and Scotland [GBR], Finnish in Finland [FIN], and Toscani in Italy [TSI]) and the corresponding processed genotypes were downloaded from EBI ArrayExpress (EBI: E-GEUV-1) (Lappalainen et al., 2013).

eQTL analysis was performed for genome-wide significant ($p < 5 \times 10^{-8}$; 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 corrected and transformed gene expression levels and trait-associated SNP genotypes. A detailed description of the *cis*-eQTL identification process is provided in Supplemental Experimental Procedures.

Enhancer-Associated TR-lincRNAs

Coordinates of ENCODE-predicted enhancer elements and H3K4me1 and H3K4me3 chromatin immunoprecipitation (ChIP) sequencing reads in human

GM12878 and mouse CH12 LCLs (ENCODE Project Consortium, 2012; Mouse ENCODE Consortium et al., 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). Details on defining putative promoter regions of TR-lincRNAs in human and mouse LCLs are provided in Supplemental Experimental Procedures.

Spatial Chromosomal Architecture Analysis

Intra-chromosomal 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 Mapping Quality (MAPQ) score above 30. Spearman's correlation was estimated between gene expression levels and the average density of contacts 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). Details on data normalization and estimation of average intra-TAD contacts are described in Supplemental Experimental Procedures.

Additional materials and methods are described in Supplemental Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2017.02.009.

AUTHOR CONTRIBUTIONS

J.Y.T. and A.C.M. designed the study. J.Y.T., A.A.T.S., M.F.d.S., C.M.-D., R.R., R.S., and D.D. performed analyses. J.Y.T., Z.K., S.B., and A.C.M. conceived methods and discussed the results. A.C.M. supervised the analysis. J.Y.T. and A.C.M. wrote the manuscript. All authors approved the manuscript.

ACKNOWLEDGMENTS

We thank Chris P. Ponting and members of the Marques group, Dario Bottinelli and Adriano Biasini for valuable comments and discussion. We thank Wilfried Haerty and Chris Rands for discussion on DAF analysis and Mathieu Heulot for discussion on experimental design. This work was funded by the Swiss National Science Foundation (grant PP00P3_150667 to A.C.M., grant FN 31003A-143914 to Z.K., and grant FN 310030_152724/1 to S.B.) and the NCCR in RNA & Disease.

Received: September 17, 2016
Revised: December 16, 2016
Accepted: January 30, 2017
Published: February 28, 2017

REFERENCES

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (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.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al.; FANTOM Consortium (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Choy, E., Yelensky, R., Bonakdar, S., Plenge, R.M., Saxena, R., De Jager, P.L., Shaw, S.Y., Wolfish, C.S., Slavik, J.M., Cotsapas, C., et al. (2008). Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. PLoS Genet. *4*, e1000287.

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

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. *93*, 779–797.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.

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

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Esteller, M. (2011). Non-coding RNAs in human disease. Nat. Rev. Genet. *12*, 861–874.

Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., and Knight, J.C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat. Genet. *44*, 502–510.

Firmann, M., Mayor, V., Vidal, P.M., Bochud, M., Pécoud, A., Hayoz, D., Paccaud, F., Preisig, M., Song, K.S., Yuan, X., et al. (2008). The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovasc. Disord. *8*, 6.

Gilad, Y., Rifkin, S.A., and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. *24*, 408–415.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell *162*, 900–910.

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

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Huang, J., Marco, E., Pinello, L., and Yuan, G.C. (2015). Predicting chromatin organization using histone marks. Genome Biol. *16*, 162.

Hüttenhofer, A., Schattner, P., and Polacek, N. (2005). Non-coding RNAs: hope or hype? Trends Genet. *21*, 289–297.

Ishii, N., Ozaki, K., Sato, H., Mizuno, H., Saito, S., Takahashi, A., Miyamoto, Y., Ikegawa, S., Kamatani, N., Hori, M., et al. (2006). Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. J. Hum. Genet. *51*, 1087–1099.

Jendrzejewski, J., He, H., Radomska, H.S., Li, W., Tomsic, J., Liyanarachchi, S., Davuluri, R.V., Nagy, R., and de la Chapelle, A. (2012). The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. Proc. Natl. Acad. Sci. USA *109*, 8646–8651.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. *9*, e1003470.

Kelley, D., and Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. *13*, R107.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Kumar, V., Westra, H.J., Karjalainen, J., Zhernakova, D.V., Esko, T., Hrdlickova, B., Almeida, R., Zhernakova, A., Reinmaa, E., Võsa, U., et al. (2013). Human disease-associated genetic variation impacts large intergenic noncoding RNA expression. PLoS Genet. *9*, e1003201.

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 Comput. Biol. *12*, e1004714.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Le Dily, F., Baù, D., Pohl, A., Vicent, G.P., Serra, F., Soronellas, D., Castellano, G., Wright, R.H., Ballare, C., Filion, G., et al. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. *28*, 2151–2162.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Manghera, M., and Douville, R.N. (2013). Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors? Retrovirology *10*, 16.

Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S., Higgs, D.R., and Ponting, C.P. (2013). Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol. *14*, R131.

Marx, V. (2014). A blooming genomic desert. Nat. Methods *11*, 135–138.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

McDowell, I.C., Pai, A.A., Guo, C., Vockley, C.M., Brown, C.D., Reddy, T.E., and Engelhardt, B.E. (2016). Many long intergenic non-coding RNAs distally regulate mRNA gene expression levels. bioRxiv. Published online March 19, 2016. http://dx.doi.org/10.1101/044719.

Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat. Protoc. *9*, 989–1009.

Merkenschlager, M., and Odom, D.T. (2013). CTCF and cohesin: linking gene regulatory elements with their targets. Cell *152*, 1285–1297.

Mouse ENCODE Consortium, Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. *13*, 418.

Neems, D.S., Garza-Gongora, A.G., Smith, E.D., and Kosak, S.T. (2016). Topologically associated domains enriched for lineage-specific genes reveal expression-dependent nuclear topologies during myogenesis. Proc. Natl. Acad. Sci. USA *113*, E1691–E1700.

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 Genet. *6*, e1000895.

Nora, E.P., Dekker, J., and Heard, E. (2013). Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? BioEssays *35*, 818–828.

Popadin, K., Gutierrez-Arcelus, M., Dermitzakis, E.T., and Antonarakis, S.E. (2013). Genetic and epigenetic regulation of human lincRNA gene expression. Am. J. Hum. Genet. *93*, 1015–1026.

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., and Aiden, E.L. (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 Res. *43*, D670–D681.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.

Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K., et al. (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. Nature *476*, 467–471.

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 Res. *15*, 1034–1050.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.

Vance, K.W., and Ponting, C.P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. Trends Genet. *30*, 348–355.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. *23*, 1494–1504.

Zheng, J., Huang, X., Tan, W., Yu, D., Du, Z., Chang, J., Wei, L., Han, Y., Wang, C., Che, X., et al. (2016). Pancreatic cancer risk variant in LINC00673 creates a miR-1231 binding site and interferes with PTPN11 degradation. Nat. Genet. *48*, 747–757.

Zuin, J., Dixon, J.R., van der Reijden, M.I., Ye, Z., Kolovos, P., Brouwer, R.W., van de Corput, M.P., van de Werken, H.J., Knoch, T.A., van IJcken, W.F., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. USA *111*, 996–1001.

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.

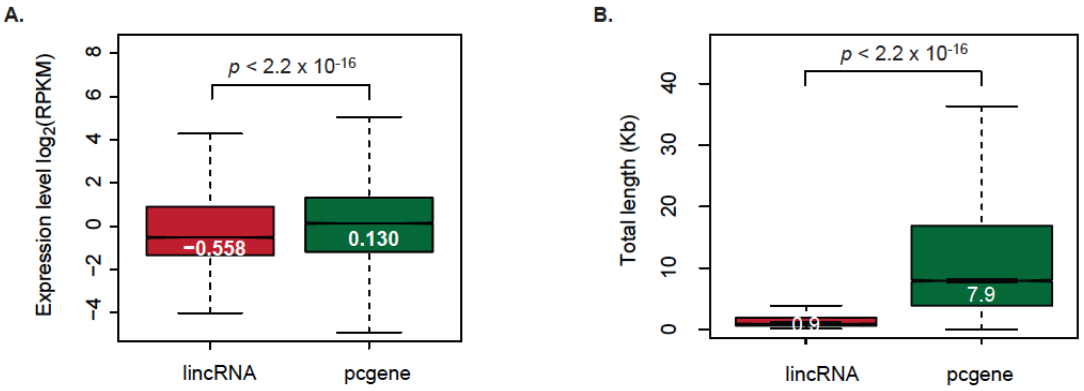


Figure S1. lincRNAs are generally shorter and more lowly expressed than protein-coding genes. Related to Figure 1. (A) Distribution of the expression levels [$\log_2$(RPKM)] and B) transcript length (Kb) of LCL-expressed lincRNAs (red) and protein-coding genes (green). Differences between groups were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

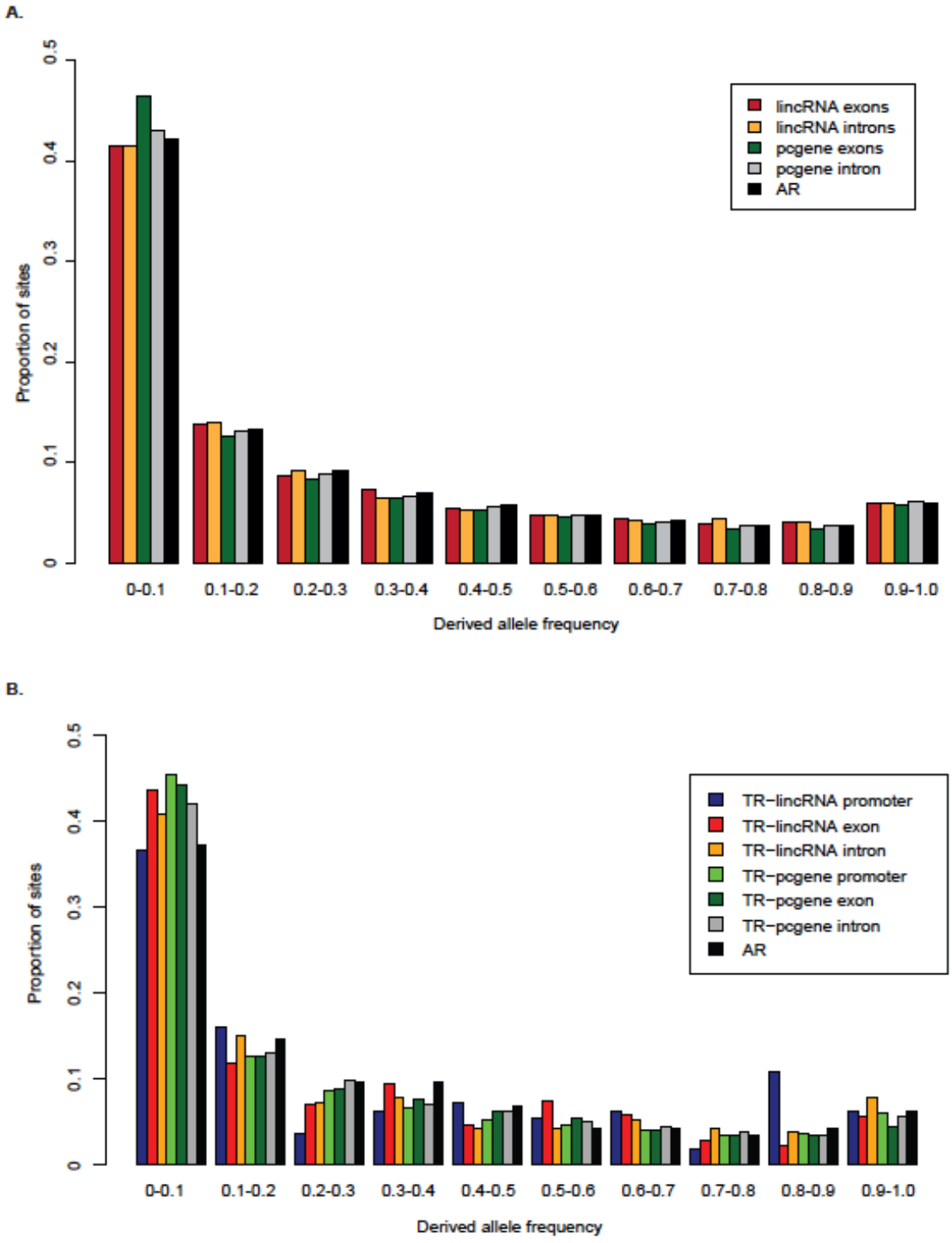


Figure S2. TR-lincRNAs evolved under purifying selection during recent human history. Related to Figure 2. Distribution of derived allele frequency (DAF) for (A) variants within exons and introns of LCL-expressed lincRNAs (exon-red, intron-orange) and protein-coding genes (exon-green, intron-grey) and 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).

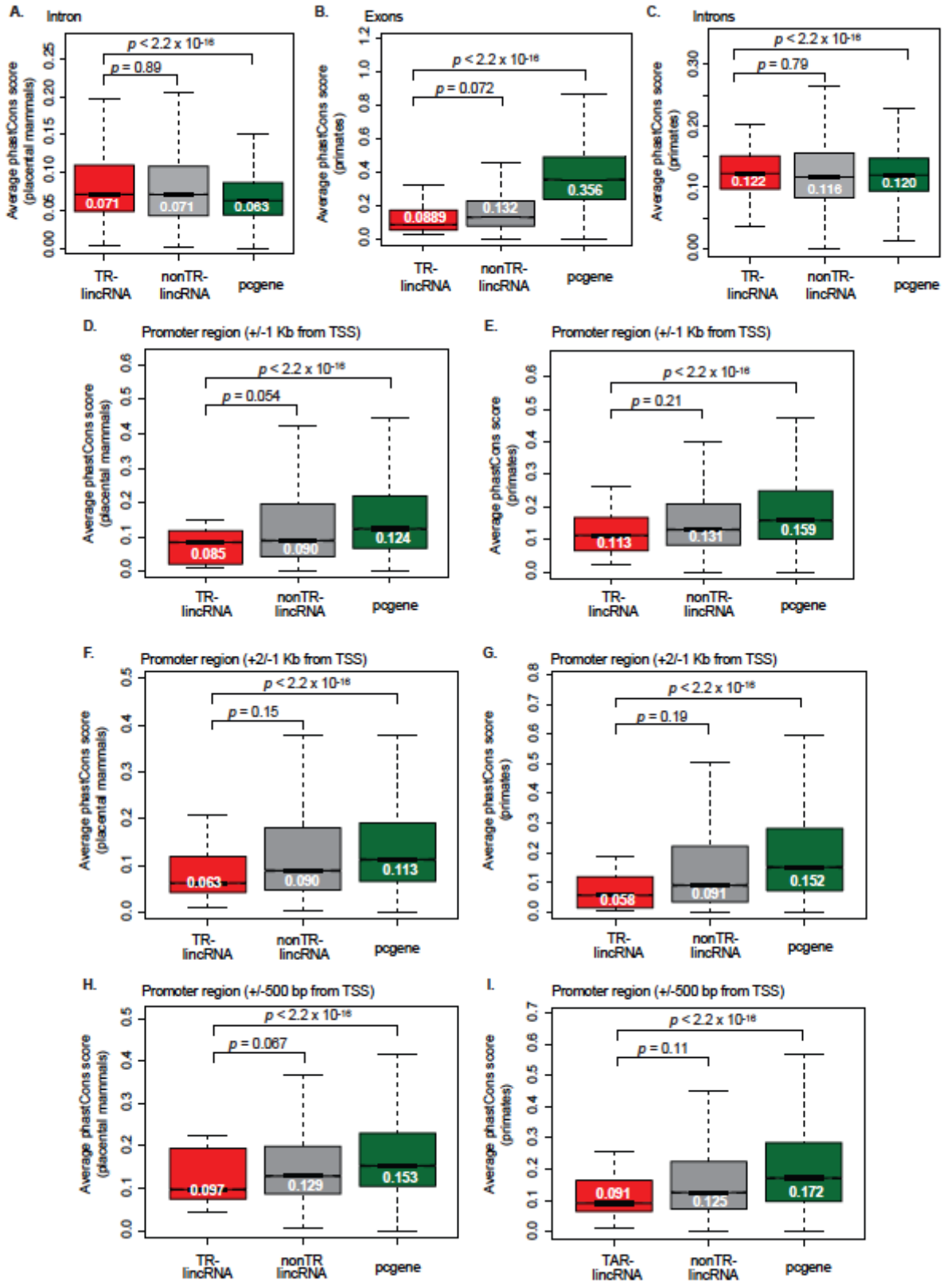


Figure S3. No evidence of constraint during TR-lincRNA evolution across mammals and primates. Related to Figure 2. Distribution of the average phastCons score of TR-lincRNA (red), other LCL-expressed lincRNA (grey), and protein-coding gene (green) for A) introns in placental mammals, (B) exons and (C) 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.

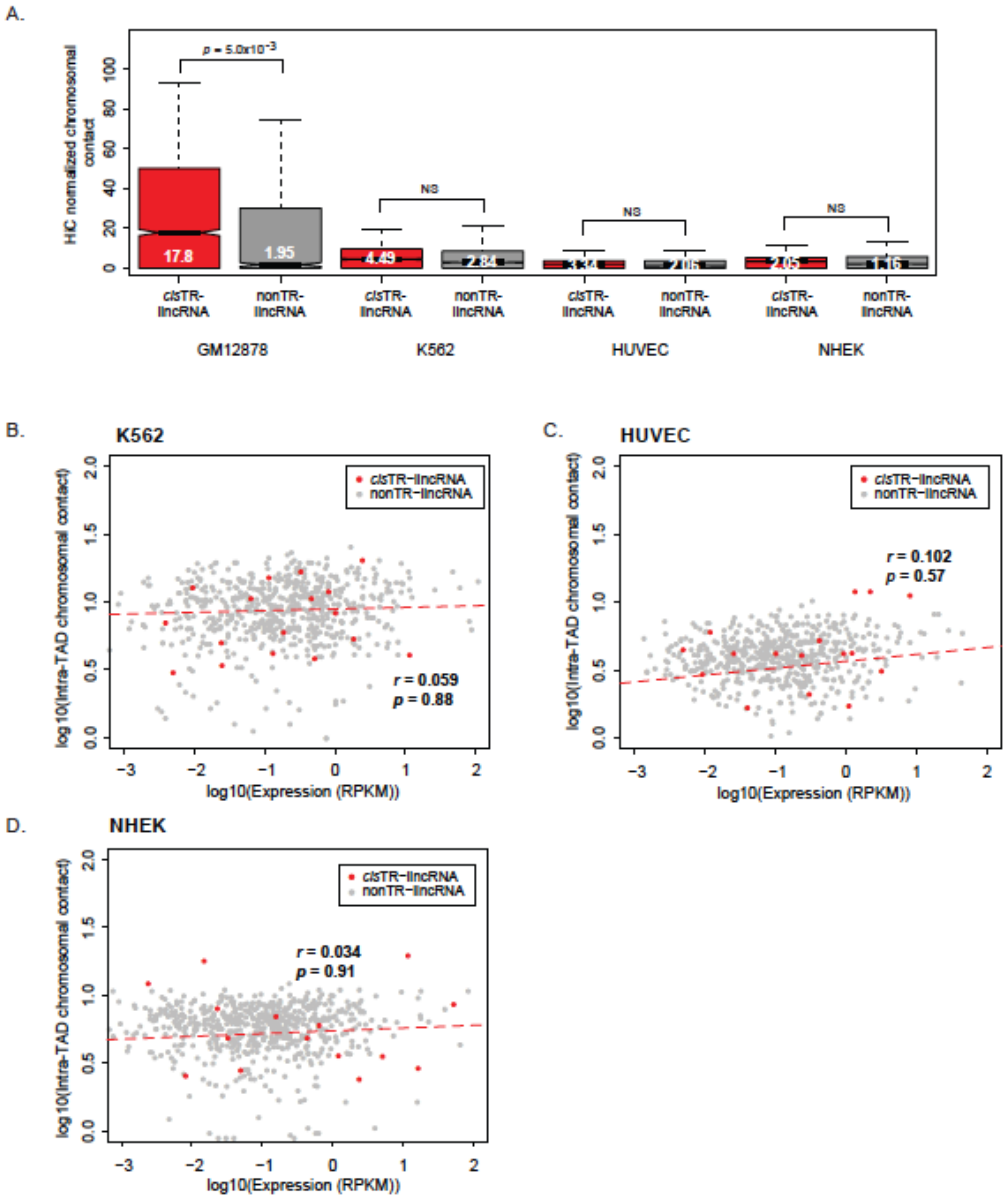


Figure S4. TR-lincRNAs regulate proximal TR-pcgenes in *cis* likely by modulating chromatin architecture. Related to Figure 4. (A) Average chromosomal contacts within TAD containing *cis*TR-lincRNAs (red) and other LCL-expressed lincRNAs (grey) in GM12878, K562, HUVEC and NHEK cell lines. Differences 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.

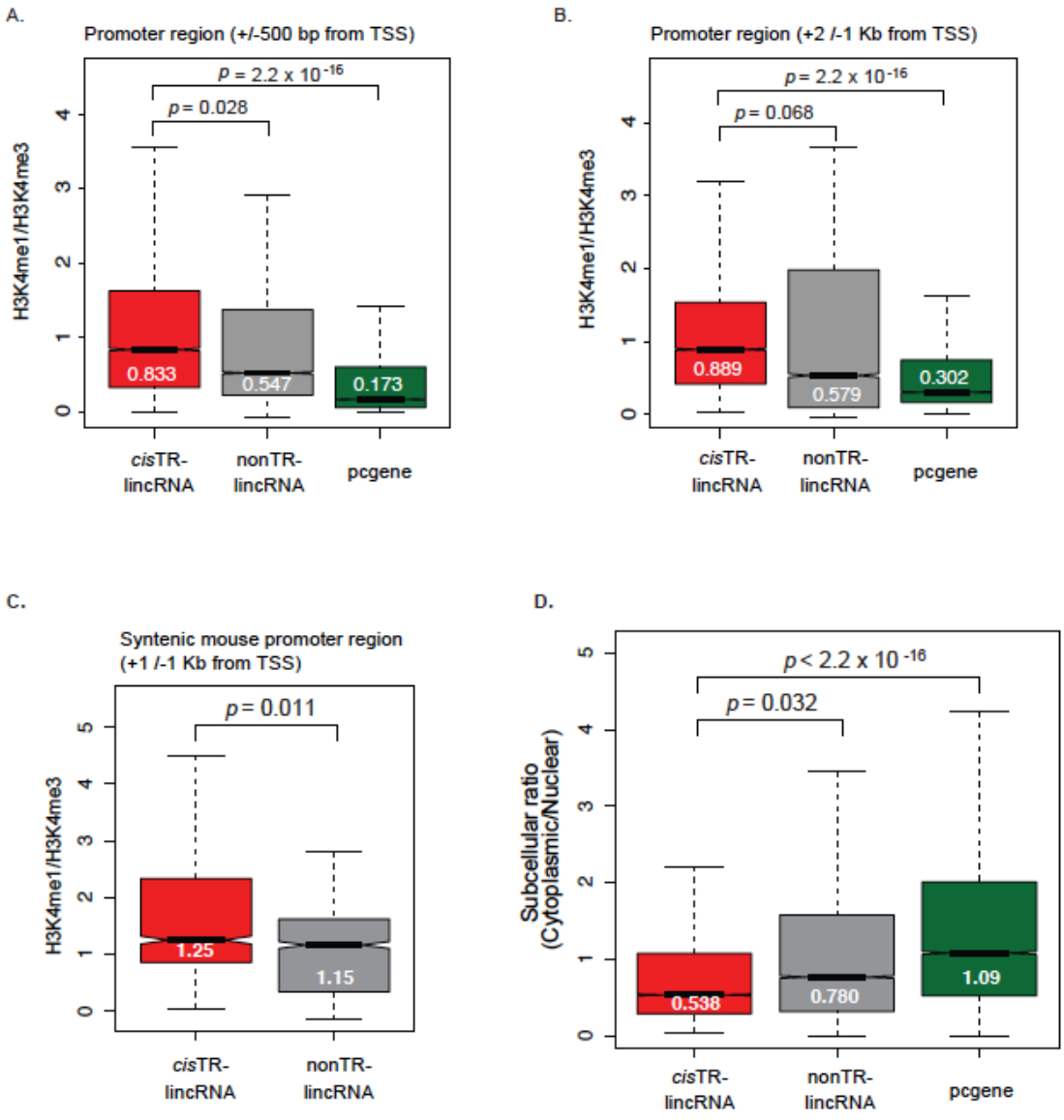


Figure S5. TR-lincRNAs promoter regions are enriched in enhancer-associated chromatin marks. Related to Figure 5. (A-C) Ratio of the number of H3K4me1 to H3K4me3 sequencing reads mapped to the putative promoter regions [A] 500 bp upstream and downstream of TSS and B) 2 Kb upstream and 1 Kb downstream 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.