Genetic and Epigenetic Regulation of Human lincRNA Gene Expression

Konstantin Popadin,[1,2,3,6] Maria Gutierrez-Arcelus,[1,2,4,6] Emmanouil T. Dermitzakis,[1,2,4,5,*] and Stylianos E. Antonarakis[1,2,*]

Large intergenic noncoding RNAs (lincRNAs) are still poorly functionally characterized. We analyzed the genetic and epigenetic regulation of human lincRNA expression in the GenCord collection by using three cell types from 195 unrelated European individuals. We detected a considerable number of *cis* expression quantitative trait loci (*cis*-eQTLs) and demonstrated that the genetic regulation of lincRNA expression is independent of the regulation of neighboring protein-coding genes. lincRNAs have relatively more *cis*-eQTLs than do equally expressed protein-coding genes with the same exon number. lincRNA *cis*-eQTLs are located closer to transcription start sites (TSSs) and their effect sizes are higher than *cis*-eQTLs found for protein-coding genes, suggesting that lincRNA expression levels are less constrained than that of protein-coding genes. Additionally, lincRNA *cis*-eQTLs can influence the expression level of nearby protein-coding genes and thus could be considered as QTLs for enhancer activity. Enrichment of expressed lincRNA promoters in enhancer marks provides an additional argument for the involvement of lincRNAs in the regulation of transcription in *cis*. By investigating the epigenetic regulation of lincRNAs, we observed both positive and negative correlations between DNA methylation and gene expression (expression quantitative trait methylation [eQTMs]), as expected, and found that the landscapes of passive and active roles of DNA methylation in gene regulation are similar to protein-coding genes. However, lincRNA eQTMs are located closer to TSSs than are protein-coding gene eQTMs. These similarities and differences in genetic and epigenetic regulation between lincRNAs and protein-coding genes contribute to the elucidation of potential functions of lincRNAs.

Introduction

The human genome encodes many thousands of long intergenic noncoding RNAs (lincRNAs), which have been annotated via transcript evidence and chromatin signatures of actively transcribed genes without protein-coding potential (6,020 lincRNA genes; Gencode version 17). The functional information on lincRNAs remains limited and, based on a small number of well-studied cases, are involved in X chromosome inactivation, genomic imprinting, cell-cycle regulation, apoptosis, and establishment of cell identity.[1–6] Despite the fact that the number of functionally annotated lincRNAs is rapidly growing, the question of whether the majority of the lincRNAs per se has a biological role is still unanswered.

lincRNAs can physically associate with chromatin regulatory proteins[7] and their promoters are considered to be target sites for key transcriptional factors.[8] Recent systematic loss-of-function experiments on all expressed lincRNAs in mouse embryonic stem cells[9] have shown that the knockdown of the vast majority of lincRNAs has a strong effect on gene expression patterns, similar to the effect of knocking down the expression of well-known regulatory proteins. The prevalent effect is in *trans* and the majority of lincRNAs maintains the pluripotent status of ESCs or represses lineage-specific gene expression programs. It has been suggested that lincRNAs could make cell-type-specific "flexible scaffolds" where distinct sets of transcribed lincRNAs interact with regulatory protein complexes and modify cell-type-specific gene expression programs.[9] Additionally, it has been proposed that lincRNAs could act as enhancers, regulating gene expression in *cis*.[10,11]

Although experimental analyses of individual lincRNAs provide direct evidence of functionality, studies of the entire gene class can provide global conclusions about function. For example, lincRNAs are more evolutionary conserved than are introns[8] and are subject to purifying selection.[12] Moreover, there is a positive correlation between the conservation and expression level of lincRNAs,[13] implying that highly expressed lincRNAs are subject to a more effective purifying selection as a result of the deleterious effects of mutations falling at their genic sequence. Gene expression studies demonstrate temporal- and spatial-specific expression of lincRNAs[14–16] and a transcript stability study revealed that lincRNA half-lives vary over a wide range, suggestive of their complex metabolism.[17]

Genetic variation can strongly affect gene expression in *cis* (*cis* expression quantitative trait loci [*cis*-eQTLs]).[18–23] Many of these variants act by modifying chromatin accessibility and transcription factor binding.[24] Initial studies have assessed the epigenetics of gene expression in a population context by studying associations between

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel-Servet, 1211 Geneva, Switzerland; [2]Institute of Genetics and Genomics in Geneva (iGE3), 1211 Geneva, Switzerland; [3]Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow 127994, Russia; [4]Swiss Institute of Bioinformatics (SIB), 1211 Geneva, Switzerland; [5]Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[6]These authors contributed equally to this work
*Correspondence: emmanouil.dermitzakis@unige.ch (E.T.D.), stylianos.antonarakis@unige.ch (S.E.A.)
http://dx.doi.org/10.1016/j.ajhg.2013.10.022.

DNA methylation and gene expression (expression quantitative trait methylation [eQTMs]) and their causal relationships.[25] Furthermore, adaptive changes in gene regulation are important determinants of gene expression variation between and within species.[26] Indeed, eQTLs are frequently found in regions of the human genome that have undergone recent positive selection.[27] Hence, investigating the patterns of variation of the genetic and epigenetic regulation of lincRNAs may provide additional evidence regarding their functionality.

Here we analyze the natural variation of lincRNA gene expression by using the GenCord collection[22,25] of three cell-types (primary fibroblast cells, immortalized lymphoblastoid cell lines, and primary T cells) from 195 unrelated European individuals for which transcriptome, genotype, and DNA methylation data are available. By comparing the genetic and epigenetic regulation between lincRNAs and protein-coding genes and by utilizing the advantages offered by this multilayered, multiple-cell-type data set, we have discovered several interesting properties of lincRNAs. Compared to protein-coding genes, we find that lincRNAs have an excess of *cis*-eQTLs, which are located closer to the TSS and have higher effect sizes, implying that lincRNA expression levels could be less constrained than those of protein-coding genes. We discover an influence of lincRNA *cis*-eQTLs on expression level of nearby protein-coding genes and an enrichment of expressed lincRNA promoters in enhancer marks that together suggest an involvement of lincRNAs in the regulation of transcription in *cis*. Finally, comparing epigenetic regulatory patterns between lincRNAs and protein-coding genes reveals mainly similarities, but analogous to eQTLs, DNA methylation sites associated with expression are closer to the TSS of lincRNAs than are protein-coding genes.

Material And Methods

Data Used

Genotype, RNA-seq, and DNA methylation data were used, processed, and analyzed as described previously.[25] In brief, umbilical cord and cord blood samples of 204 newborn individuals of European descent were collected in order to derive three cell types: primary fibroblasts, primary T cells, and lymphoblastoid cell lines (LCLs). Individuals were genotyped with Illumina 2.5M Omni chip. Filtered genotypes were imputed into the 1000 Genomes European panel SNPs of the Phase 1 release with Beagle.[28] This yielded 5,209,348–5,278,330 SNPs with minor allele frequencies >5%. Expression levels for all cell types and samples were measured with RNA-seq. Libraries were selected for polyadenylated transcripts and were sequenced as 49 bp paired-end reads in either HiSeq2000 or Genome Analyzer II machines. A median of 16 million reads was mapped to merged exons from the Gencode v.10 annotation.[10] Scaled exon counts were further normalized by correcting the effects of GC content, run date, primer index, and insert size mode by linear regression. We considered expressed exons as those for which there is at least one mapped read in at least 90% of individuals. This yielded sets of 70,800–76,870 exons belonging to 12,265–12,863 genes. DNA methylation levels were measured for all cell types for a subset of samples by using the 450K Illumina Infinium HD Methylation Assay. Probes containing SNPs were removed, yielding 416,118 CpG sites to analyze. Data were quantile normalized and the β-value[29] was used for DNA methylation levels, which represents the percentage of methylation per site.

Association analyses were performed by Spearman rank correlation and multiple testing corrections via permutations methods. At 10% FDR, the following discoveries were made: 2,115–3,372 eQTLs found in 183–186 samples (1 Mb window to either side of the TSS), 14,189–32,318 mQTLs found in 66–111 samples (5 kb window to either side of the CpG site), and 596–3,838 eQTM genes involving 970–6,846 CpG sites in 66–118 samples (50 kb window to either side of the TSS).

The causative model analysis was performed by using Bayesian Network construction and relative likelihood to determine which of the following models is the most likely given the data. We tested the SME model, in which the SNP affects methylation and methylation affects expression; the SEM model, in which the SNP affects expression and expression affects methylation; and the INDEP model, in which the SNP independently affects both methylation and expression. The R package bnlearn[30] was used to calculate the maximum likelihood of each network and the Akaike Information Criterion (AIC) score. This score was then used to determine the relative likelihood of each network with respect to the others. This approach was tested for all SNP-exon-CpG triplets, with at least two out of the three pairwise correlations being significant at 10% FDR. For the particular case of the lincRNA genes, results from all three cell types were merged given the small number of cases that we were able to test.

lincRNA Genes

We used the following criteria to select lincRNA exons from the Gencode annotation:[10,31] gene_type = 'lincRNA' and transcript_type = 'lincRNA', which resulted in 4,746 lincRNA genes with exons satisfying these criteria. lincRNAs are, by definition, intergenic genes. Our analysis of Gencode 10 annotation of these 4,746 genes has revealed that the vast majority of lincRNAs (4,259 out of 4,746) do not overlap with protein-coding genes, though some lincRNAs overlap protein-coding genes on the opposite (n = 407), the same (n = 63), or both (n = 17) strands. We therefore eliminated these overlapping genes and performed all our analyses on the subset of lincRNAs that do not overlap with protein-coding genes.

Estimation of Expression Level

RPKM data for each exon were obtained as the median number of reads mapped to an exon and normalized to exon length. Only exons expressed in at least 40 samples (~20% of samples) were taken into account. RPKM values for each gene were obtained as the median value of RPKM data for exons.

Normal-Transformation of Exon Expression Level

The expression level of exons was transformed via the *rntransform* function in the GenABEL R package.

Generation of the Matched Data Set

To compare lincRNAs with protein-coding genes, for each cell type we identified a subset of protein-coding genes matched to lincRNA genes based on the number of exons and expression level. For each lincRNA, we selected up to 15 protein-coding genes with the same number of exons and expression level that did not deviate more

Table 1. Number of lincRNA and Protein-Coding Genes Expressed in Each Cell Type

Gene Type	Cell Types		
	Fibroblasts	Lymphoblastoid Cell Lines	T Cells
Number of Genes, Expressed in at least 20% of Samples			
lincRNA	562	666	743
Protein-coding	15,501	15,386	15,963
Number of Genes, Expressed in at least 90% of Samples			
lincRNA	153	210	206
Protein-coding	12,785	12,357	12,938
Number of Genes with *cis*-eQTLs			
lincRNA	30	50	40
Protein-coding	2,386	3,300	2,057
Number of Positive/Negative eQTMs			
lincRNA	5/4	40/59	19/55
Protein-coding	755/776	6,278/8,154	5,349/ 11,802

Only genes expressed at least in 90% of samples were used for *cis*-eQTL and eQTM calls.

than 1% from that of the lincRNA. All selected protein-coding genes were then combined and duplicates were removed.

Enhancer Enrichment Analysis

Enhancer mark coordinates were downloaded from the UCSC genome browser tables.[28] These coordinates were obtained from ChIP-seq experiments of the ENCODE project and particular groups.[32–34] ChIP-seq data from the NHLF lung fibroblast cell line was used for our fibroblast analyses, and the GM12878 lymphoblastoid cell line was used for LCLs and for T cells given that it was the closest cell line with available data. We defined the promoter as the region spanning −1 kb to +2 kb of the transcription start site.

All statistical analyses were performed in R.

Results

Patterns of lincRNA Expression and Location

We identified 562, 666, and 743 lincRNA genes expressed in at least 20% of samples in fibroblasts (F), lymphoblastoid cell lines (L), and T cells (T), respectively (Table 1; Figure S1 available online). With these genes we sought to assess the general patterns of lincRNA expression, conservation, and location. In line with previous works,[14–16,35] we have shown that lincRNAs are less frequently expressed, are expressed at lower levels, and are more tissue specific than are protein-coding genes (Figures S1–S3). We confirm the recently described positive correlation between the expression level and conservation score of lincRNAs,[13] and we have demonstrated that the conservation score is associated with the level of tissue specificity of lincRNAs and gradually increases from non-expressed lincRNAs to those expressed in one, two, and three investigated tissues (Figure S4). Similarly with the observation made in mouse and zebrafish genomes,[3,8,35] we observed a nonrandom localization of expressed lincRNAs in the human genome. The lincRNAs expressed in our study colocalize with genes involved in zinc ion binding (Table S1 and Figures S5 and S6), suggesting an involvement in transcriptional control, given that 40% of zinc binding proteins in the human proteome are transcription factors.[36] Overall, these results confirm lincRNA properties previously described in other species and cell types.

Genetic Regulation of lincRNA Expression Variation

To assess the patterns of genetic regulation of lincRNAs, we analyzed exons expressed in at least 90% of samples: 153, 210, and 206 lincRNA genes in F, L, and T, respectively (Table 1). We defined a *cis*-eQTL as the most significant SNP located within a 1 Mb window around the transcription start site (TSS) that is associated with the expression of at least one exon (see Material and Methods). We found *cis*-eQTLs for 30, 50, and 40 lincRNA genes in F, L, and T, respectively (Table 1). We then compared *cis*-eQTLs of lincRNAs with *cis*-eQTLs of protein-coding genes in terms of their abundance, location, and effect size.

lincRNA *cis*-eQTL Abundance

We found that 19%–24% of the tested lincRNA genes have *cis*-eQTLs, which is similar to that of protein-coding genes (16%–27%) (Table 1). However, the small number of expressed exons found in lincRNA genes and their low expression level can introduce a bias into the comparison at the gene level. Because the majority of lincRNAs express only one exon (70% of the expressed lincRNAs in our study), we compared fractions of *cis*-eQTLs between single-expressed-exon lincRNA genes and single-expressed-exon protein-coding genes. We have observed 1.5-, 1.2-, and 1.9-fold excess of *cis*-eQTLs for lincRNAs versus protein-coding genes in F, L, and T, respectively. The fraction of *cis*-eQTLs in lincRNAs is higher, irrespective of the set p value thresholds (all p values < 0.004, Mann-Whitney paired U-test; Figure 1A) and the expression level of genes (all p values < 0.002, Mann-Whitney paired U-test; Figure 1B). We further confirmed this trend by creating a matched data set of protein-coding genes that have both a similar number of expressed exons and similar expression levels as do lincRNA genes (see Material and Methods). Comparison of lincRNAs with the matched data set of protein-coding genes confirms excess of *cis*-eQTLs among lincRNAs (p values $= 0.043$, 0.077, and 0.032; odds ratios $= 1.50$, 1.30, and 1.45; one-sided Fisher test). The combined analysis from all cell types also confirmed an excess of *cis*-eQTLs among lincRNAs versus matched protein-coding genes (odds ratio $= 1.41$, p value $= 0.001$; one-sided Fisher test). The excess of lincRNA *cis*-eQTLs that we observed is opposite to the findings reported recently in a study using microarray expression

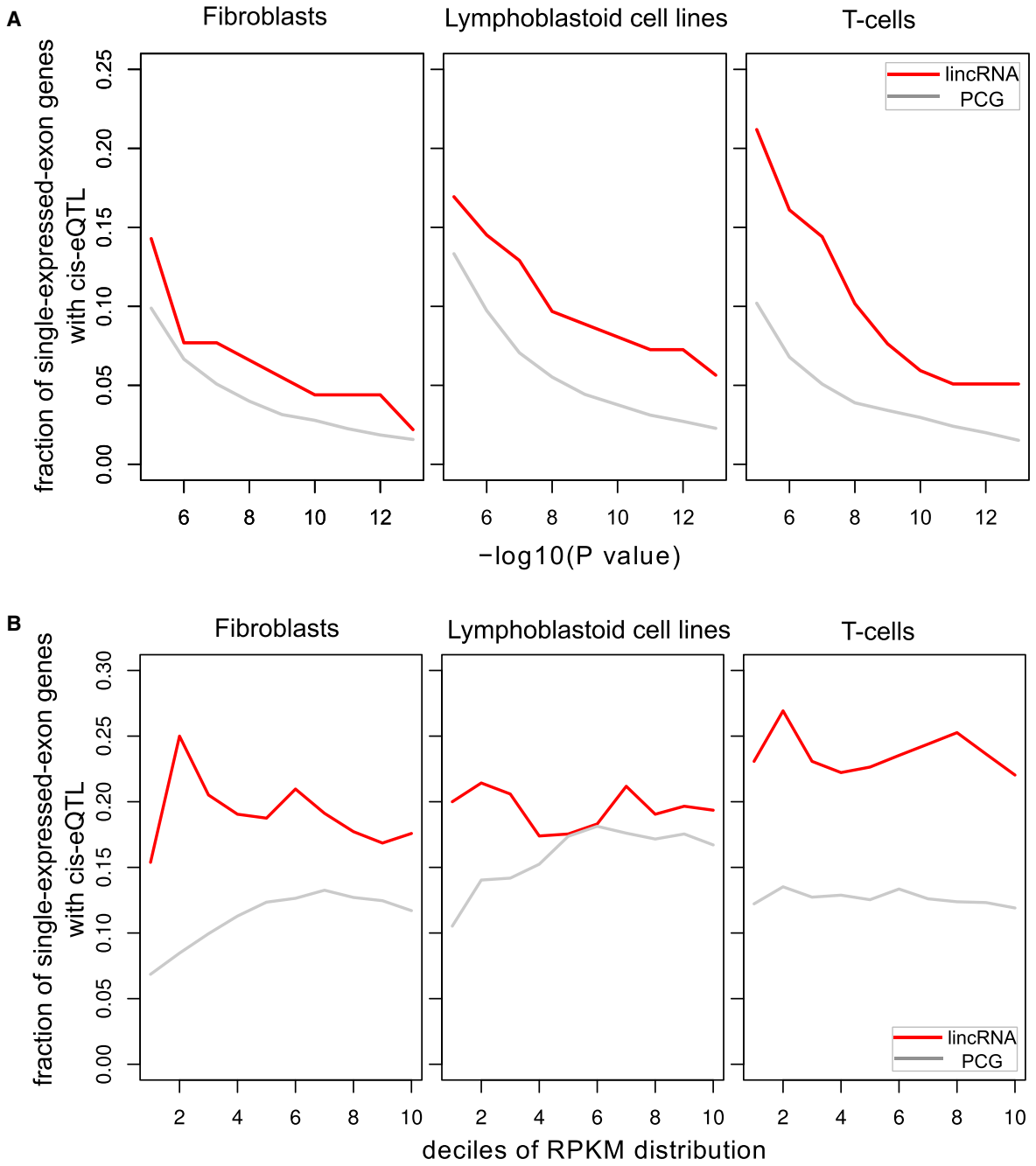

Figure 1. lincRNAs Have a Larger Abundance of *cis*-eQTLs than Do Protein-Coding Genes
(A) Fraction of single-expressed-exon genes with *cis*-eQTLs (y axis) for protein-coding genes (PCG, gray) and lincRNA genes (red), called at different p value thresholds (x axis). We observe a larger fraction of single-expressed-exon genes with *cis*-eQTLs in lincRNAs versus PCGs for all p value thresholds.
(B) Fraction of single-expressed-exon genes with *cis*-eQTLs (y axis) for protein-coding genes (PCG, gray) and lincRNA genes (red), at different expression levels based on deciles of the expression level distribution of protein-coding genes (x axis). From the distribution of expression levels of PCGs, we have split the genes into ten deciles. By using the decile breakpoints, we have determined ten lincRNA subsets, corresponding to ten PCG deciles. For each decile we have calculated a fraction of *cis*-eQTL-associated genes. We observe a larger fraction of single-expressed-exon genes with *cis*-eQTLs in lincRNAs versus PCGs irrespective of expression level.

levels.[37] We believe our findings are robust because our study is based on high-resolution RNA-seq data and takes advantage of the wider dynamic range of expression afforded by this technology, and also because we observed this trend in three different cell types. Overall, the high abundance of DNA polymorphisms influencing lincRNA expression levels could suggest that lincRNA expression is less constrained than protein-coding gene expression.

lincRNA *cis*-eQTL Effect Size

If lincRNA genes indeed allow more changes in their expression levels, we expect eQTL effect sizes to be higher

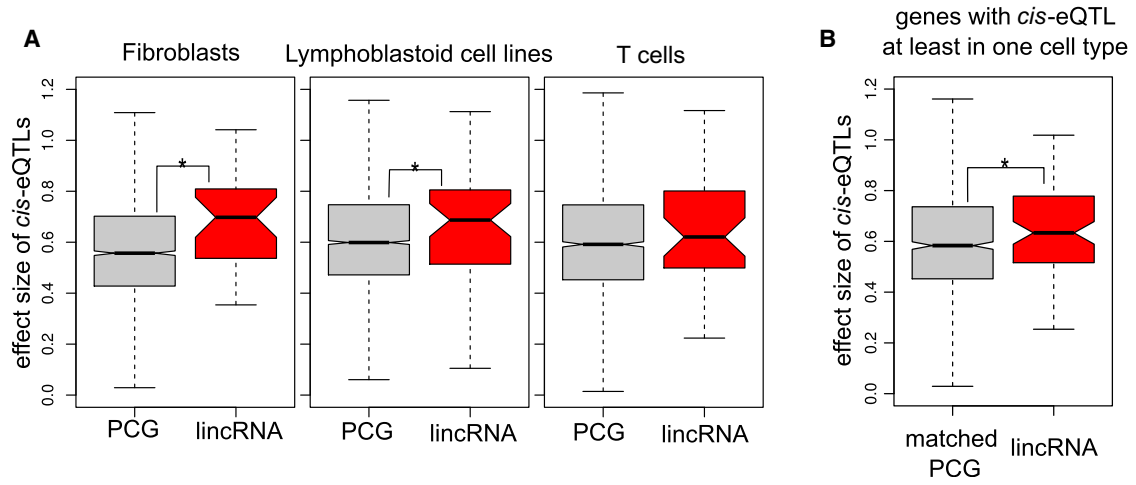

Figure 2. cis-eQTL Effect Sizes are Larger in lincRNAs than in Protein-Coding Genes
(A) *cis*-eQTL effect size comparison between lincRNA genes and protein-coding genes (PCGs). In fibroblasts and lymphoblastoid cell lines, lincRNA *cis*-eQTLs have significantly higher effect sizes than do PCG *cis*-eQTLs. Star indicates statistical significant difference with $p \leq 0.021$, one-sided Mann-Whitney U-test.
(B) Effect sizes of *cis*-eQTLs of lincRNAs are higher than *cis*-eQTLs of protein-coding genes, integral analysis. Star indicates statistical significant difference with $p = 0.014$, one-sided Mann-Whitney U-test. For each gene with *cis*-eQTL at least in one tissue, we have estimated average (among tissues) effect size of its *cis*-eQTLs. Only protein-coding genes, matched to lincRNAs in respect to their number of expressed exons and expression level, have been used for this analysis.
Effect size was calculated as the difference in median scaled expression levels between heterozygous individuals and individuals homozygous for the major allele. To get scaled expression levels, we subtracted the mean from expression values and divided by the standard deviation.

for lincRNAs than for protein-coding genes. In order to test this, we estimated the effect size of each *cis*-eQTL as the difference in median scaled expression levels between heterozygous individuals and individuals homozygous for the most frequent allele. By using scaled expression levels (subtracting the mean and dividing by the standard deviation), the effect size is measured as the number of standard deviations affected by an allele change. Our data show that the absolute magnitude of *cis*-eQTL effect sizes is higher for lincRNAs than for protein-coding genes (median *cis*-eQTL effect sizes for lincRNAs are 0.70, 0.69, and 0.62 for F, L, and T; median values for protein-coding genes are 0.56, 0.60, and 0.59 for F, L, and T), with statistical significance found in two of the three cell types (p values = 0.001, 0.021, and 0.149 for F, L, and T, one-sided Mann-Whitney U-test; Figure 2A). By analyzing all cell types together (averaging effect sizes when more than one *cis*-eQTL was found per gene), we have confirmed a significantly higher average effect size in lincRNAs compared to the matched set of protein-coding genes (p value = 0.014, one-sided Mann-Whitney U-test; Figure 2B). Together, these results further support the conclusion that lincRNA genes tolerate more gene expression changes than do protein-coding genes.

lincRNA *cis*-eQTL Location

It has been previously demonstrated that highly significant *cis*-eQTLs of protein-coding genes are located proximally to the TSS of the gene, whereas less significant *cis*-eQTLs are distributed more distantly.[23] For lincRNA *cis*-eQTLs, we have observed a similar pattern (Figure 3A). However, we noted that the majority of lincRNA *cis*-eQTLs are located preferentially closer to the TSS than those of protein-coding genes. Indeed, the median distances for lincRNA *cis*-eQTLs are 2.5–4.5 times lower than the distances for protein-coding gene *cis*-eQTLs ($p = 1.5 \times 10^{-3}$; $p = 3.2 \times 10^{-4}$; $p = 2.9 \times 10^{-3}$ for F, L, and T; Mann-Whitney U-test; Figure 3B). Analysis of the protein-coding genes matched to lincRNA genes by the number of exons and expression level confirmed this trend (p values = 0.016, 0.065, and 0.027 for F, L, and T, one-sided Mann-Whitney U test), as did integrating the data from all cell types (taking average distance when more than one *cis*-eQTL is found per gene; p value = 0.030, one-sided Mann-Whitney U-test). Overall, these results show that lincRNA *cis*-eQTLs, compared to those found for protein-coding genes, are closer to TSSs, suggesting a deficit of distant regulatory elements for lincRNA genes.

cis-eQTLs Common for lincRNA and Protein-Coding Exons

cis-eQTLs significantly associated with expression levels of both protein-coding and lincRNA exons can be used to test whether there is an independent regulation of lincRNA expression by the *cis*-eQTL or whether lincRNAs are likely to be a byproduct of protein-coding gene expression. We identified 15, 48, and 30 pairs of genetically coregulated lincRNA-protein-coding exon pairs in F, L, and T. In order to distinguish the influence of each

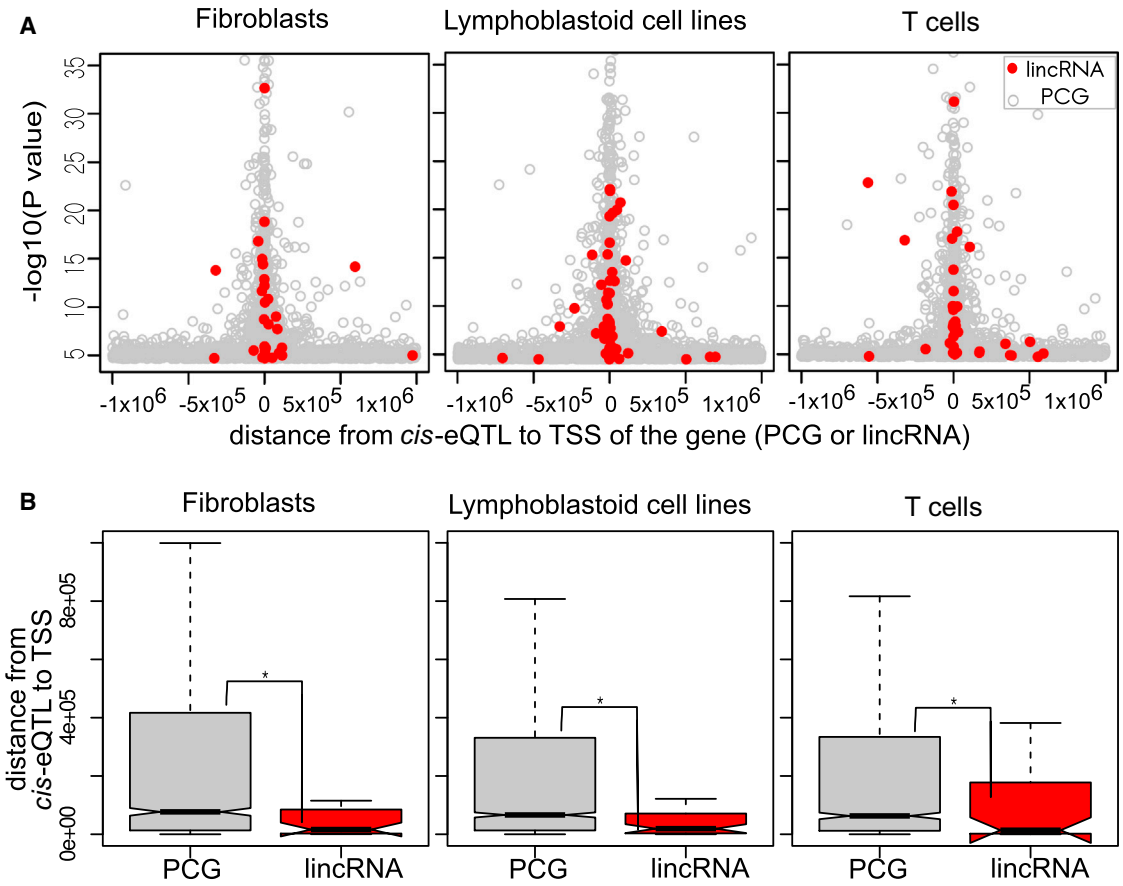

Figure 3. lincRNA *cis*-eQTLs Are Located Closer to the Transcription Start Site than Are Protein-Coding Gene *cis*-eQTLs
(A) The location of *cis*-eQTLs of protein-coding genes (PCGs, gray) and *cis*-eQTLs of lincRNAs (red) relative to the transcription start site (TSS, x axis), by their level of significance based on –log(p value) (y axis). The most significant *cis*-eQTLs are located closer to the TSS for both lincRNA and protein-coding genes, in a symmetric manner.
(B) Comparison of location of *cis*-eQTLs of protein-coding genes (PCGs, gray) and *cis*-eQTLs of lincRNAs (red). *cis*-eQTLs of lincRNAs are located significantly closer to transcription start site (TSS) than are *cis*-eQTLs of protein-coding genes. Star indicates statistical significant difference with $p \leq 2.9 \times 10^{-3}$, Mann-Whitney U-test.

cis-eQTL on the expression levels of lincRNA and protein-coding exons, we performed multiple linear regression analysis as follows: the *cis*-eQTL genotype is treated as the dependent variable and the normal-transformed exon expression levels of the associated lincRNA and protein-coding genes as two independent variables (see Material and Methods). If an association between a *cis*-eQTL and the expression level of a lincRNA is a byproduct of regulation of a protein-coding gene by the *cis*-eQTL, the p value obtained for lincRNA in this model should not be significant. We found that the predominant influence of *cis*-eQTLs on lincRNA or protein-coding exons changes between cell types (Figures 4A and 4B). *cis*-eQTLs have a stronger influence on protein-coding exons in fibroblasts (p value $< 2.2 \times 10^{-16}$, Mann-Whitney paired U-test), but the influence of *cis*-eQTLs is stronger on lincRNA exons in LCL and T cells (p values $< 2.2 \times 10^{-16}$, Mann-Whitney paired U-test). Interestingly, when we perform an integrative analysis by combining information from all cell types, we have observed a slightly higher influence of *cis*-eQTLs on lincRNAs versus protein-coding genes (median −log10(p values) for lincRNA is 4.58 versus 3.37 for protein-coding genes; p value $< 2.2 \times 10^{-16}$, Mann-Whitney paired U-test). Overall, these results suggest independent regulation of many lincRNA genes and reject the model that lincRNA expression is mainly a regulatory byproduct of protein-coding genes.

Distal Effects of *cis*-eQTLs

The expression level of lincRNAs, when not independent, can be affected by the transcription level of protein-coding genes located upstream. To analyze the influence of upstream genes on the transcription level of genes located downstream, we used *cis*-eQTLs associated with the upstream gene (hereafter referred to as proximal *cis*-eQTL effect) and have estimated the effect of this *cis*-eQTL on the expression level of the downstream gene (hereafter referred to as distal *cis*-eQTL effect) (see Figure 5A). First, we extracted all pairs of annotated genes that are the immediate neighbors located on the same strand. Second, we included in the analysis only the nonoverlapping gene pairs, where both genes are expressed in the investigated cell type. Third, we selected those pairs for which there is a *cis*-eQTL for the first gene and further required

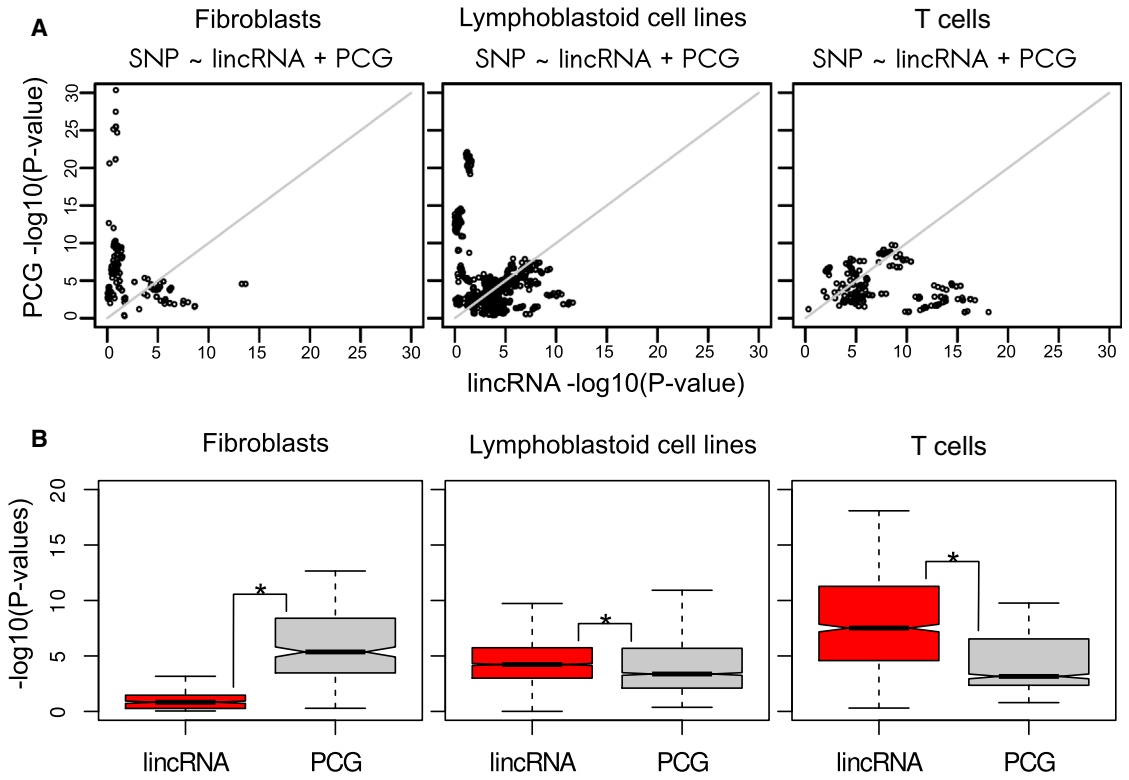

Figure 4. Effect of *cis*-eQTLs Shared between lincRNA and Protein-Coding Genes

(A) p values depicting the effects whether a SNP associated to both lincRNA and protein-coding genes is really reflecting an independent effect on one or the other gene or whether it is significantly affecting both independently. To evaluate independence of effects for *cis*-eQTLs shared between protein-coding and lincRNA genes, we used a multiple linear model where the *cis*-eQTL genotype (SNP) was taken as the dependent variable and the lincRNA expression (lincRNA) and protein-coding gene expression (PCG) were taken as independent variables: SNP ~lincRNA + PCG. Under this scenario, the –log(p value) for each independent variable, lincRNA (x axis) and PCG (y axis), are plotted. Given that we observe cases where the lincRNA p value remains significant despite having the PCG as covariate, we can conclude that for many of the *cis*-eQTLs shared between protein-coding and lincRNA genes, we are capturing a real effect on the lincRNA expression. This is contrary to the hypothesis that lincRNA expression could be just a by-product of protein-coding gene expression.

(B) Box plots depicting the distribution of –log(p values) of lincRNA expression (lincRNA, red) and protein-coding gene expression (PCG, gray) in the context of the multiple linear model used for assessing independent effects of *cis*-eQTLs significant in both classes of genes: SNP ~lincRNA + PCG. Star depicts a significant difference with $p < 2.2 \times 10^{-16}$, Mann-Whitney paired U-test. In fibroblasts most of the shared eQTLs between protein-coding and lincRNA genes probably reflect dominant effects on protein-coding genes. However, in lymphoblastoid cell lines and T cells, effects on lincRNA expression seem stronger.

that this association is more significant for the first than for the second (downstream) gene (see Figure 5A). Next, we split all gene pairs into three groups: P-P pairs (protein-protein), P-L pairs (protein-lincRNA), and L-P pairs (lincRNA-protein). We then compared the distal *cis*-eQTL effects among these three groups (Figure 5A) and found that distal *cis*-eQTL p values are higher (less significant) in P-L pairs compared to P-P pairs (p values 3.3×10^{-5}, 2.2×10^{-7}, 0.1×10^{-3} for F, L, and T, Mann-Whitney U-test), suggesting that lincRNAs are less influenced by upstream genes than are protein-coding genes (Figures 5B and 5C). We also observed that L-P pairs in L and T cell types have more significant distal *cis*-eQTL p values than do P-L pairs (p values = 0.52, 3.0×10^{-6}, 0.006 for F, L, and T, Mann-Whitney U-test), suggesting that lincRNAs exert a stronger influence on the downstream gene compared to protein-coding genes (Figures 5B and 5C). We noted that distal *cis*-eQTL effects in L and T tissues depend on the distance from the *cis*-eQTL to the TSS of the downstream gene (Spearman's rho = 0.013, −0.090, and −0.066 with p = 0.44, 2.4×10^{-8}, and 4.6×10^{-5} for F, L, and T): the longer the distance, the less significant the distal *cis*-eQTL effect. To control for the effect of distance, we estimated median p values for each of the four quartiles of the distance distribution for the different pairs and we observe the same trends. Our data show that lincRNAs are less affected by upstream protein-coding genes and that lincRNAs exert greater influence upon the downstream gene than do protein-coding genes, irrespective of the distance (Figure 5D). Overall, our results demonstrate that lincRNAs are independent units of transcription from the neighboring protein-coding genes and also suggest that lincRNAs may act as common *cis* regulatory elements of downstream protein-coding genes. This further suggests that a considerable fraction of lincRNA *cis*-eQTLs is likely to be enhancer QTLs.

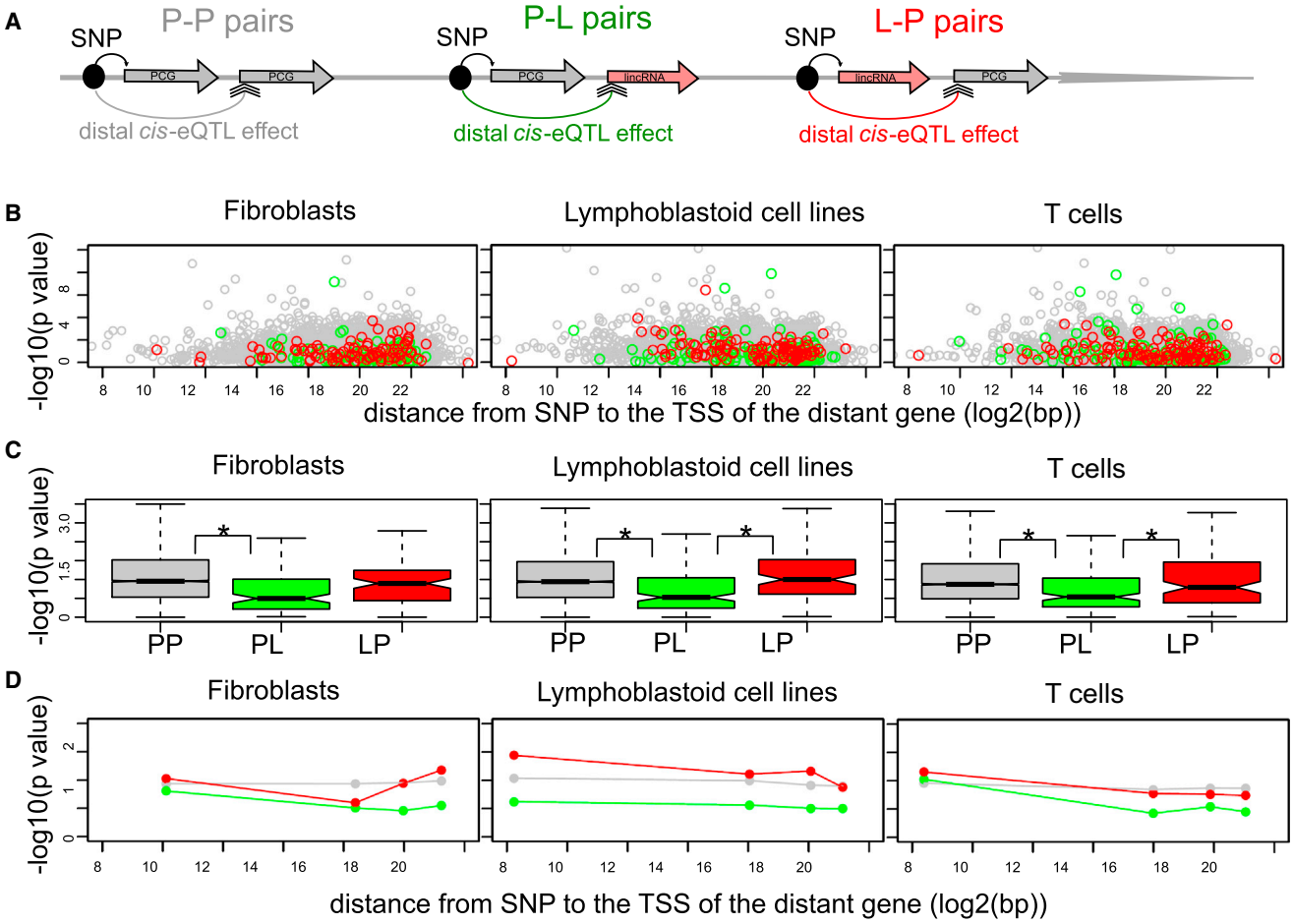

Figure 5. lincRNA *cis*-eQTLs often Influence Expression of Nearby Protein-Coding Genes

(A) A scheme of the analysis of independent transcription of lincRNA genes. The next pairs of expressed neighbor genes are considered: protein-coding-protein-coding (P-P pairs, gray), protein-coding-lincRNA (P-L pairs, green), and lincRNA-protein-coding (L-P pairs, red). The distal *cis*-eQTL effect (effect of *cis*-eQTL on the expression level of the second gene) is estimated and compared between the three types of gene pairs. Only gene pairs that satisfy the following criteria have been analyzed: genes are immediate neighbors located on the same strand; genes are nonoverlapping; both genes are expressed in the investigated cell type; and there is a *cis*-eQTL for the first gene that is more significantly associated with the first than with the second (downstream) gene.

(B) p values, representing distal *cis*-eQTL effect in P-P pairs (gray), P-L pairs (green), and L-P pairs (red) are plotted as a function of distance between SNP and transcription start site (TSS) of the distal (second) gene.

(C) Comparison of distal *cis*-eQTL effects between different pairs of neighbor genes: P-P pairs (gray), P-L pairs (green), and L-P pairs (red). Star depicts a significant difference with $p \leq 0.006$, Mann-Whitney U-test. Box plots depicting the distribution of $-\log10$(p values) for P-P, P-L, and L-P pairs demonstrate that P-L pairs tend to have less significant p values than do P-P and L-P pairs, meaning that lincRNAs are less influenced by distal *cis*-eQTLs than are protein-coding genes. This suggests that lincRNAs are very often not a by-product of protein-coding gene expression.

(D) Distal *cis*-eQTL effects between different pairs of neighbor genes: P-P pairs (gray), P-L pairs (green), and L-P pairs (red) demonstrate less significant p values for P-L pairs irrespective of distance from *cis*-eQTL to the transcription start site (TSS) of the distal gene. Four dots connected by line correspond to median values of $-\log10$(p values) for P-P (gray), P-L (green), and L-P (red) pairs for four quartiles of distance distribution.

lincRNAs as Enhancers

The regulation of downstream protein-coding genes by lincRNAs is compatible with a hypothesis that lincRNAs could act as enhancers.[11,38] To test this hypothesis, we assessed the enrichment of frequently expressed lincRNAs in enhancer regions defined by chromatin marks (see Material and Methods). Specifically, we counted the overlaps between enhancer marks and the promoter regions of expressed lincRNAs and compared these to the overlaps between enhancer marks and the promoter regions of the matched data set of protein-coding genes. We found a significant enrichment of expressed lincRNA promoters in enhancers (Figure 6A) (Fisher's odds ratio = 2.01, 2.39, and 1.67 for F, L, and T; all Fisher's p values $< 1.1 \times 10^{-6}$). Furthermore, because lincRNA expression is highly tissue specific, we asked whether tissue-specific protein-coding gene *cis*-eQTLs could be enriched in expressed lincRNA genes. Despite the low number of data points available, we found a significant enrichment (with respect to a null, see Material and Methods) of tissue-specific protein-coding *cis*-eQTLs in expressed lincRNA genes in LCLs ($p = 0.004$, Fisher's odd

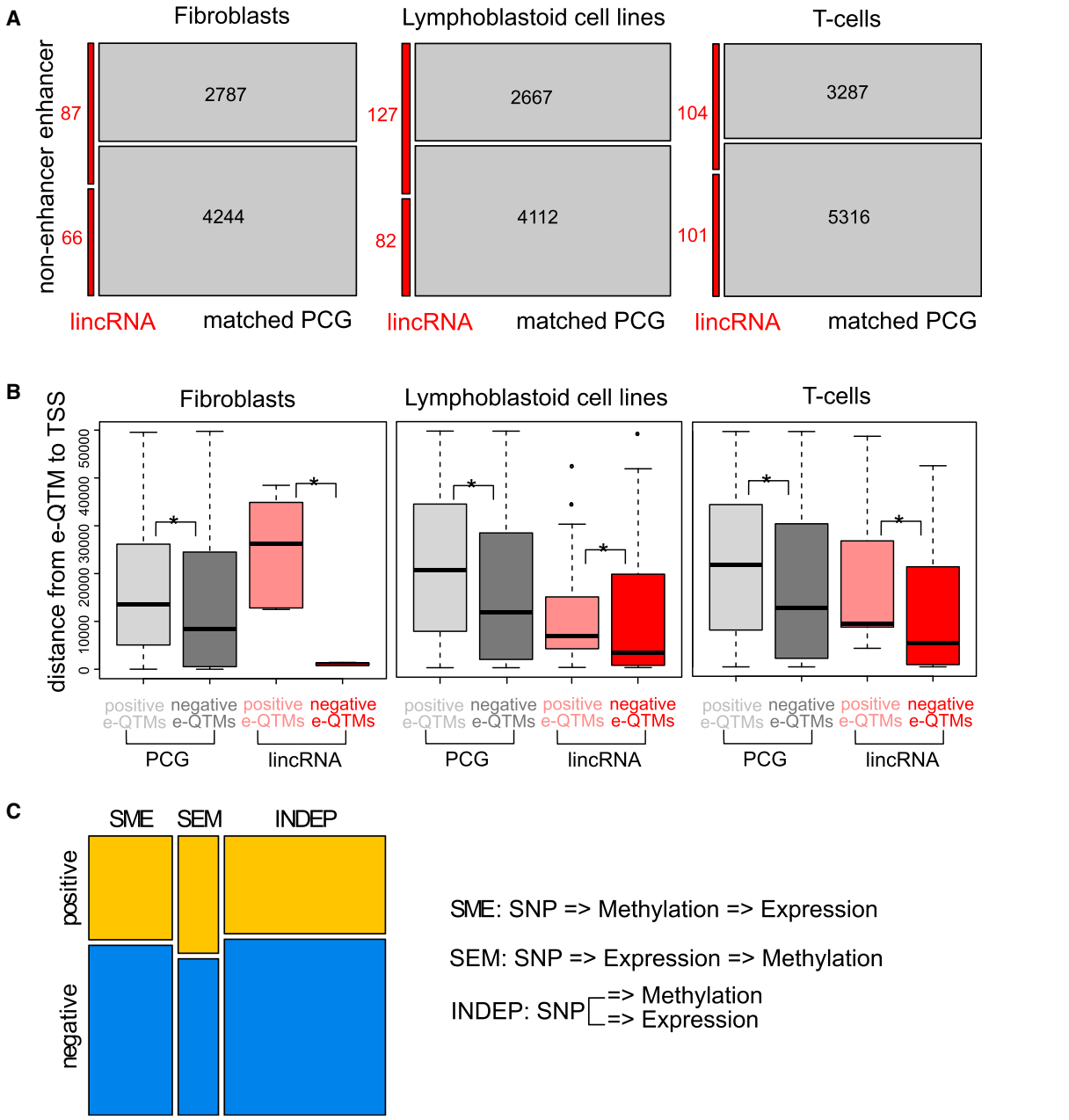

Figure 6. Epigenetic Regulatory Patterns of lincRNAs

(A) Enrichment of expressed lincRNA promoters in enhancer marks. Mosaic plots depict the relative frequency of lincRNA promoters (red) and protein-coding gene promoters matched for number of exons and expression levels (matched PCG, gray; x axis), by the relative frequency of cases overlapping (enhancer) or not overlapping (nonenhancer) enhancer marks (y axis). Enhancer marks are based on ENCODE chromatin marks data (see Material and Methods). Data specific for fibroblasts and LCLs were used for each corresponding cell type. No data for T cells were available at the time of analysis so the data for LCLs were used instead for this cell type. Enrichment of expressed lincRNA promoters on enhancer marks is significant with $p < 1.1 \times 10^{-6}$ for all cell types; Fisher's exact test.

(B) Location of methylation sites associated to gene expression, relative to the transcription start site (TSS). Box plots depict the distance from methylation sites associated to gene expression (expression quantitative trait methylation [eQTMs]) to transcription start sites (TSSs) of lincRNA and protein-coding genes (PCGs). For both PCGs and lincRNAs, negative eQTMs are located closer to TSS than are positive eQTMs. Stars indicate $p < 0.035$, Mann-Whitney U-test.

(C) Inference of mechanistic relationships among genetic variation, DNA methylation, and lincRNA expression. Schemes on the right depict the three causative models tested by constructing Bayesian networks and determining the most likely model given our data with relative likelihood (see Material and Methods). The SME model depicts a scenario in which the SNP affects methylation and methylation affects expression. The SEM model shows a scenario in which the SNP affects expression and expression affects methylation. The INDEP model illustrates a case in which the SNP is independently affecting gene expression and DNA methylation. Mosaic plots depict the relative frequency of each model (x axis), by the relative frequency of cases involving positive (yellow) or negative (blue) associations between DNA methylation and gene expression. Triplets of SNP, DNA methylation site, and lincRNA exon were tested if at least two out of the three pairwise correlations were significant. The mechanistic landscape inferred for lincRNAs looks very similar to that inferred for protein-coding genes.

ratio = 6.98). Overall, these results suggest that many lincRNAs transcripts could be enhancer RNAs (eRNAs) and may contribute to, or mark, the tissue-specific regulation of protein-coding genes in *cis*.

Epigenetic Regulation of lincRNA Gene Expression

Little is known about the patterns of epigenetic regulation of lincRNA expression. Recent studies have begun to explore the correlations between DNA methylation and gene expression in a population context in different cell types.[39–43] It has been observed that DNA methylation in CpG sites located within 50 kb of TSSs can be both positively and negatively associated with gene expression levels (expression quantitative trait methylation [eQTMs]), with negative eQTMs being significantly enriched in promoter regions compared to positive eQTMs.[25] Here we compared epigenetic patterns of gene expression regulation between protein-coding genes and lincRNAs. We found similar proportions of positive and negative eQTMs for lincRNAs and protein-coding genes (Table 1; Fisher's exact p values > 0.3). Furthermore, in both lincRNA and protein-coding genes, negative eQTMs are significantly closer to the TSS than positive eQTMs (Figure 6B, all p values < 0.035 and all p values $< 3 \times 10^{-12}$ in lincRNA and protein-coding genes, respectively; Mann-Whitney U-test). Interestingly, overall eQTMs are closer to the TSS in lincRNAs compared to protein-coding genes in LCLs and T cells ($p = 8.11 \times 10^{-7}$ and $p = 8.72 \times 10^{-4}$, respectively), this being maintained for negative eQTMs ($p = 9.86 \times 10^{-4}$ and $p = 8.23 \times 10^{-4}$, respectively) and for positive eQTMs only in LCLs ($p = 4.25 \times 10^{-4}$). It is possible that we did not observe the same signal in fibroblasts because of the reduced number of eQTMs we were able to analyze ($n = 9$ and $n = 1{,}531$ for lincRNA and protein-coding genes, respectively). Overall, these results indicate that lincRNAs are subject to a similar epigenetic regulation as are protein-coding genes. However, in a manner similar to what is observed for genetic regulation, the epigenetic correlations with lincRNA gene expression tend to manifest themselves in a smaller distance around the TSS.

By correlating DNA methylation and gene expression levels, it is impossible to know the causal direction of effects, i.e., whether DNA methylation changes alter gene expression, whether gene expression changes alter DNA methylation, or whether gene expression and DNA methylation are correlated given that they are independently affected by a common factor. By utilizing genetic variation as an anchor, the relative likelihoods of the three above-mentioned scenarios can be inferred by Bayesian Networks construction.[25] This methodology has allowed the inference of the proportion of passive and active participation of DNA methylation in gene regulation. The INDEP model (see scheme Figure 6C) depicts a scenario in which a SNP independently affects gene expression and DNA methylation (passive role for DNA methylation). The SME model occurs when the SNP affects methylation and methylation affects expression (active role). Finally, in the SEM model the SNP affects expression and expression affects methylation (passive role). It has been observed in protein-coding genes that in general the INDEP model tends to be the more likely model, followed by the SME and SEM models. In addition, by examining the proportion of positive and negative correlations between DNA methylation and gene expression found in each of the three models, it was observed that the SEM model presents a higher proportion of positive correlations compared to the SME model. In order to assess whether the same proportion of epigenetic regulatory patterns observed in protein-coding genes would be found in lincRNA genes, we inferred the most likely model for SNP-methylation-exon triplets in which at least two out of the three pair-wise correlations were significant, as previously described.[25] Because of the small number of testable lincRNAs, we merged the results of the three cell types. As expected, lincRNAs displayed similar proportions for the three different models compared to the pattern generally observed in protein-coding genes (Figure 6C). In our data, the INDEP model is the most frequent pattern, followed by the SME model and the SEM model. Furthermore, there is a higher proportion of positive eQTMs in the SEM model compared to the SME. Together, these results further support the observation that the epigenetic regulatory mechanisms present in lincRNAs are similar to those participating in protein-coding genes.

Discussion

In this study we uncovered interesting properties concerning the genetic regulation of lincRNAs and compared the epigenetic (DNA methylation) regulatory mechanistic landscape between lincRNAs and protein-coding genes. These findings provide insights into lincRNA functionality.

From a mechanistic point of view, we observed that lincRNA *cis*-eQTLs tend to influence neighbor downstream protein-coding genes that, in combination with the excess of enhancer marks in lincRNA promoter regions, may suggest an involvement of lincRNAs as enhancer-like *cis* regulators of transcription. This hypothesis is supported by the tissue-specific expression patterns of lincRNAs (as we and others observe), because enhancers often drive tissue-specific expression. Interestingly, the observations we report on eQTLs and eQTMs being closer to the TSS of lincRNAs also support the enhancer hypothesis. Under this scenario, we can speculate that lincRNAs may present a lack of distant associations because they are the distant regulators themselves, and hence they would be subject only to local, enhancer-like promoter regulation. Alternatively, another potential explanation for a lack of distant lincRNA *cis*-eQTLs is their young age, according to which we may expect that lincRNAs haven't had sufficient time to acquire long-distance regulatory elements. However, future studies will need to address these aspects in more detail.

From an evolutionary perspective, we found that lincRNAs are more tolerant to changes in gene expression levels than are protein-coding genes. It has been shown that the primary mode of selection acting on expression level of genes is stabilizing selection.[44] Thus, we interpret the excess of *cis*-eQTLs in lincRNAs, together with their larger effect sizes and closer proximity to TSSs, mainly as a signature of relaxed purifying selection acting on regulatory regions of lincRNAs. Indeed, because the expression level of lincRNAs can frequently be affected by DNA polymorphism, it is possible that the function of lincRNAs is not as essential and/or is less sensitive to expression levels. However, taking into account that lincRNAs are young genes and thus have an increased rate of evolution and variable selection pressure compared to old genes,[45–47] we hypothesize that a significant fraction of lincRNA *cis*-eQTLs can be under positive selection. Additional studies are needed to resolve the question of whether this excess of lincRNA *cis*-eQTLs is explained by relaxed purifying selection or by positive selection acting on lincRNA expression levels.

Supplemental Data

Supplemental Data include six figures and one table and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We acknowledge X. Bonilla for editing the manuscript, F. Santoni and M. Garieri for discussion of statistical analyses, S. Nikolaev for discussion of potential roles of lincRNAs, and S. Lukowski for the help in the preparation of the manuscript. This study has been supported by SNF grant 144082 and ERC grant 249968 to S.E.A. and by SNFRNA grant 31003A_130342, ERC grant 260927 POPRNASEQ, and Louis-Jeantet foundation to E.T.D. K.P. was supported by EMBO long-term fellowship program ALTF 527-2010 and Novartis grant N°3A15. The computations were performed at the Vital-IT Center for high-performance computing of the SIB Swiss Institute of Bioinformatics.

Received: August 23, 2013
Revised: October 11, 2013
Accepted: October 21, 2013
Published: November 21, 2013

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://browser.1000genomes.org
BEAGLE, http://faculty.washington.edu/browning/beagle/beagle.html
Gencode v.17, http://www.gencodegenes.org
R statistical software, http://www.r-project.org/
Vital-IT, http://www.vital-it.ch

References

1. Lee, J.T. (2012). Epigenetic regulation by long noncoding RNAs. Science *338*, 1435–1439.
2. Latos, P.A., Pauler, F.M., Koerner, M.V., Şenergin, H.B., Hudson, Q.J., Stocsits, R.R., Allhoff, W., Stricker, S.H., Klement, R.M., Warczok, K.E., et al. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science *338*, 1469–1472.
3. Ponjavic, J., Oliver, P.L., Lunter, G., and Ponting, C.P. (2009). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet. *5*, e1000617.
4. Pauli, A., Rinn, J.L., and Schier, A.F. (2011). Non-coding RNAs as regulators of embryogenesis. Nat. Rev. Genet. *12*, 136–149.
5. Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.
6. Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26–46.
7. Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. USA *106*, 11667–11672.
8. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.
9. Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.
10. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.
11. Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.
12. Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. *17*, 556–565.
13. Managadze, D., Rogozin, I.B., Chernikova, D., Shabalina, S.A., and Koonin, E.V. (2011). Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. Genome Biol. Evol. *3*, 1390–1404.
14. Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. USA *105*, 716–721.
15. Dinger, M.E., Amaral, P.P., Mercer, T.R., Pang, K.C., Bruce, S.J., Gardiner, B.B., Askarian-Amiri, M.E., Ru, K., Soldà, G., Simons, C., et al. (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res. *18*, 1433–1445.
16. 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.
17. Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome Res. *22*, 885–898.

18. Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M., and Burdick, J.T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. Nature *437*, 1365–1369.
19. Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.
20. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.
21. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavaré, S., et al. (2005). Genome-wide associations of gene expression variation in humans. PLoS Genet. *1*, e78.
22. Dimas, A.S., Stranger, B.E., Beazley, C., Finn, R.D., Ingle, C.E., Forrest, M.S., Ritchie, M.E., Deloukas, P., Tavaré, S., and Dermitzakis, E.T. (2008). Modifier effects between regulatory and protein-coding variation. PLoS Genet. *4*, e1000244.
23. 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.
24. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNaseI sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.
25. Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., et al. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife *2*, e00523.
26. Fay, J.C., and Wittkopp, P.J. (2008). Evaluating the role of natural selection in the evolution of gene regulation. Heredity (Edinb) *100*, 191–199.
27. Kudaravalli, S., Veyrieras, J.-B., Stranger, B.E., Dermitzakis, E.T., and Pritchard, J.K. (2009). Gene expression levels are a target of recent natural selection in the human genome. Mol. Biol. Evol. *26*, 649–658.
28. Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223.
29. Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E.W., Wu, B., Doucet, D., Thomas, N.J., Wang, Y., Vollmer, E., et al. (2006). High-throughput DNA methylation profiling using universal bead arrays. Genome Res. *16*, 383–393.
30. Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. J. Stat. Softw. *35*, 1–22.
31. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.
32. 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.
33. Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.
34. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311–322.
35. Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.
36. Andreini, C., Banci, L., Bertini, I., and Rosato, A. (2006). Counting the zinc-proteins encoded in the human genome. J. Proteome Res. *5*, 196–201.
37. 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 non-coding RNA expression. PLoS Genet. *9*, e1003201.
38. Ørom, U.A., and Shiekhattar, R. (2011). Long non-coding RNAs and enhancers. Curr. Opin. Genet. Dev. *21*, 194–198.
39. Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y., and Pritchard, J.K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. *12*, R10.
40. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. *6*, e1000952.
41. Kulis, M., Heath, S., Bibikova, M., Queirós, A.C., Navarro, A., Clot, G., Martínez-Trillos, A., Castellano, G., Brun-Heath, I., Pinyol, M., et al. (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat. Genet. *44*, 1236–1242.
42. van Eijk, K.R., de Jong, S., Boks, M.P.M., Langeveld, T., Colas, F., Veldink, J.H., de Kovel, C.G.F., Janson, E., Strengman, E., Langfelder, P., et al. (2012). Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics *13*, 636.
43. Zhang, D., Cheng, L., Badner, J.A., Chen, C., Chen, Q., Luo, W., Craig, D.W., Redman, M., Gershon, E.S., and Liu, C. (2010). Genetic control of individual differences in gene-specific methylation in human brain. Am. J. Hum. Genet. *86*, 411–419.
44. Gilad, Y., Oshlack, A., and Rifkin, S.A. (2006). Natural selection on gene expression. Trends Genet. *22*, 456–461.
45. Albà, M.M., and Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. Mol. Biol. Evol. *22*, 598–606.
46. Wolf, Y.I., Novichkov, P.S., Karev, G.P., Koonin, E.V., and Lipman, D.J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc. Natl. Acad. Sci. USA *106*, 7273–7280.
47. Vishnoi, A., Kryazhimskiy, S., Bazykin, G.A., Hannenhalli, S., and Plotkin, J.B. (2010). Young proteins experience more variable selection pressures than old proteins. Genome Res. *20*, 1574–1581.