The American Journal of Human Genetics, Volume *93*

Supplemental Data

Genetic and Epigenetic Regulation of Human lincRNA Gene Expression

Konstantin Popadin, Maria Gutierrez-Arcelus, Emmanouil T. Dermitzakis, and Stylianos E. Antonarakis

Supplemental Figures and Tables

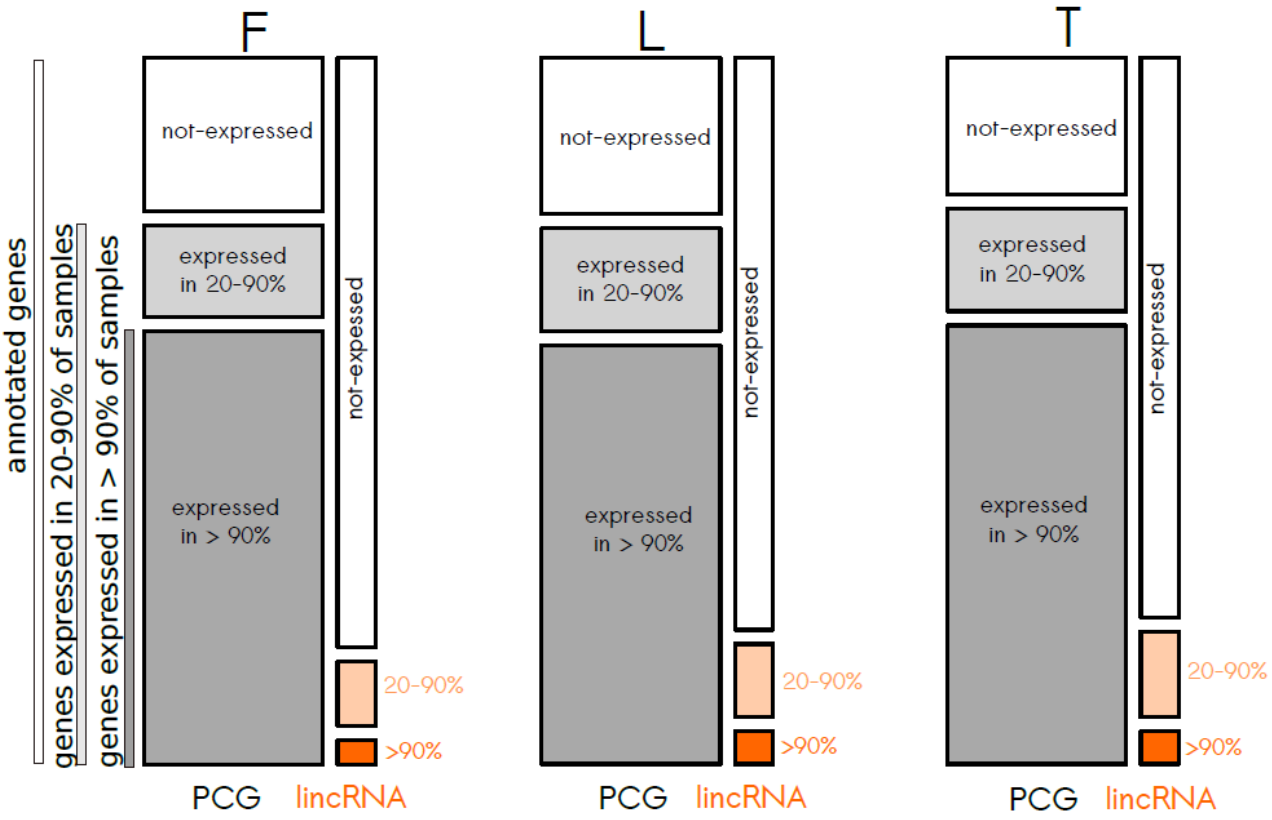

Figure S1. LincRNAs are more rarely-expressed than protein-coding genes. The mosaic plots demonstrating fraction of protein-coding genes (PCG) and lincRNAs, expressed at least in 20% and 90% of samples. We have found that from the number of annotated lincRNA genes (N = 4259), only 13-17% (N = 562-743) are expressed in at least 20% of the individuals and 4-5% (N = 153-210) are expressed in at least 90% of the individuals in fibroblasts (F), LCLs (L) and T-cells (T). These proportions are much smaller than those found for the annotated protein coding genes (N = 20007), in which 77-80% and 62-65% are expressed at least in 20 and 90% correspondingly.

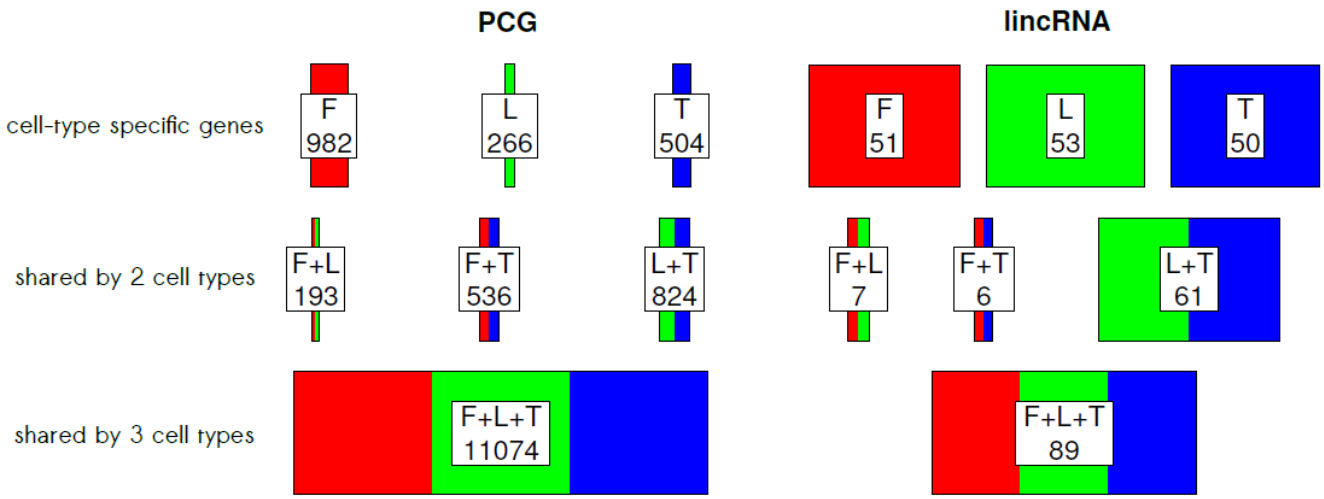

Figure S2. LincRNAs demonstrate an excess of cell-type specifically expressed genes. The intersect diagram demonstrating the expression patterns of protein-coding genes (PCG) and lincRNAs, expressed at least in 90% of samples.

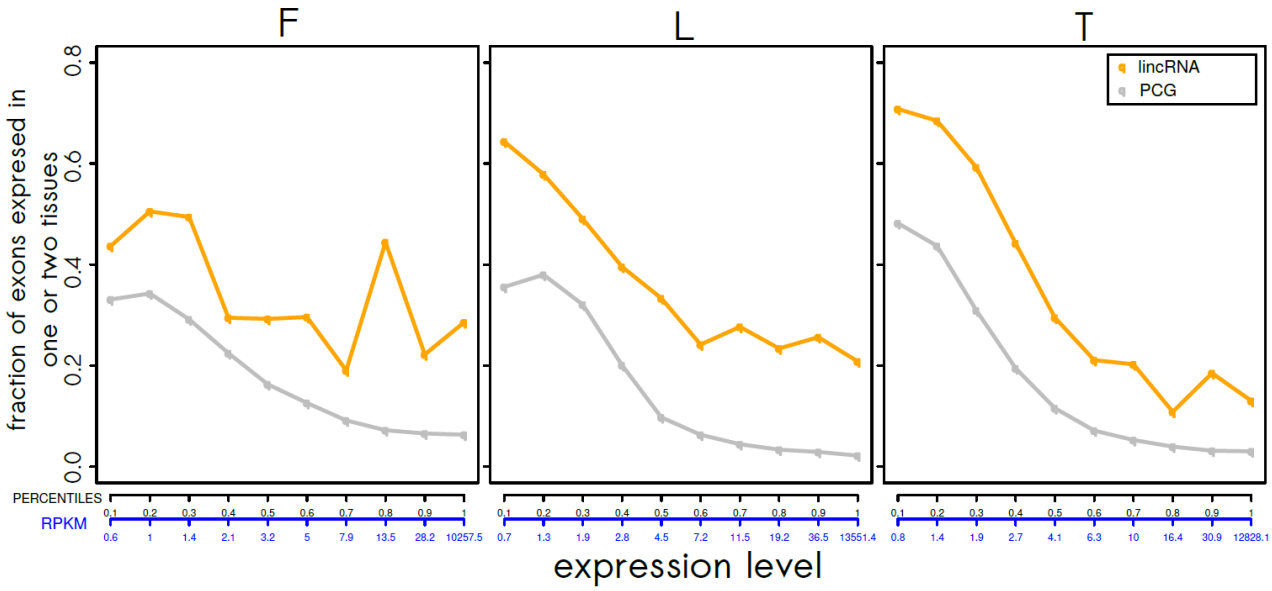

Figure S3. LincRNAs are more tissue-specific irrespectively of expression level. LincRNAs present about 5-fold lower levels of expression than protein-coding genes (all p-values < 2.0E-16, Mann-Whitney U-test). So, increased tissue-specificity of lincRNAs can be the result of their low expression level. In order to control for the expression level we compared tissue-specificity (fraction of exons, expressed in one or two tissues) of protein-coding (gray line) and lincRNA (orange line) exons, coming from the same decile of the distribution of expression levels of all exons. The upper X axis marks deciles of the distribution of expression levels of all exons, the lower axis provides absolute values of median expression level of exons (RPKM) for each decile. We observe that irrespectively of expression level the fraction of tissue-specific exons is 20% higher for lincRNAs than for the protein coding genes.

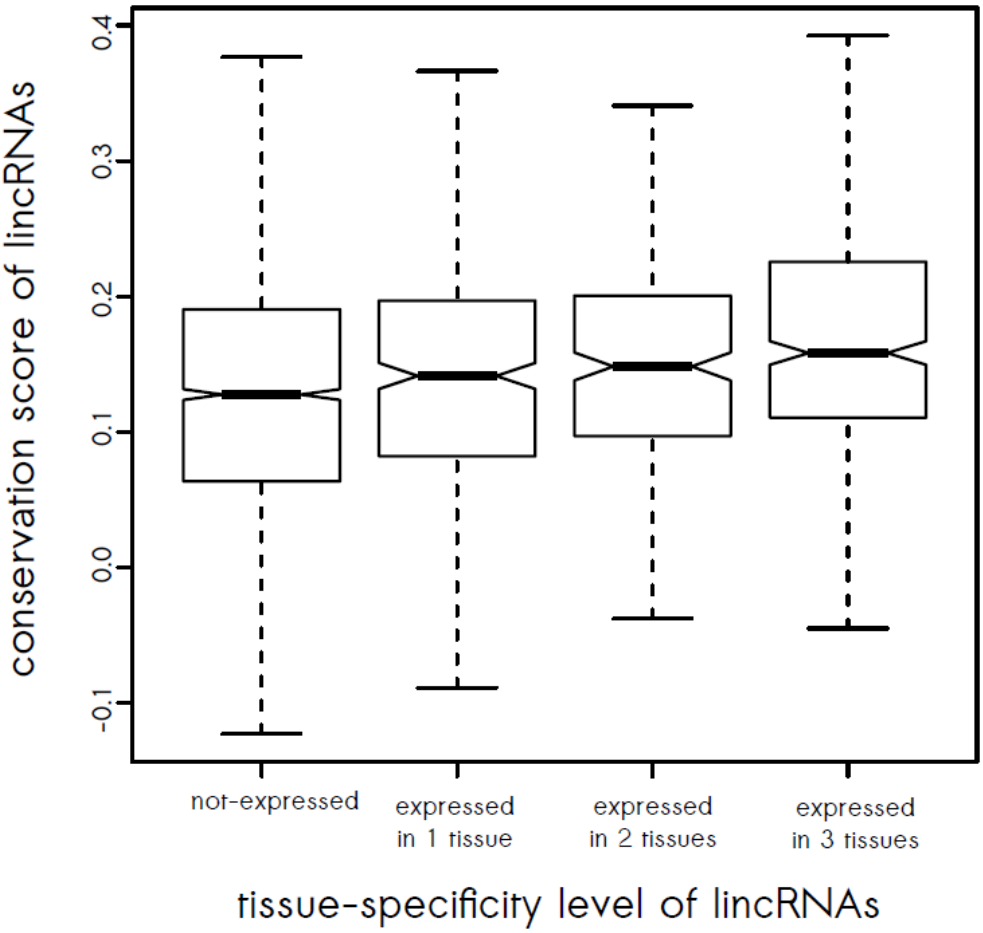

Figure S4. Conservation score is increased in ubiquitously expressed lincRNAs: the conservation score is gradually increasing from non expressed lincRNAs, to those expressed in one, two and three investigated tissues. We have estimated lincRNA conservation score as a median of nucleotide-based phyloP scores of lincRNA exons from primate multiple alignment using UCSC genome browser (http://genome.ucsc.edu/). High conservation score of ubiquitously expressed lincRNAs assumes stronger evolutionary constraints of lincRNAs expressed in many tissues versus lincRNAs specifically expressed in a few particular tissues.

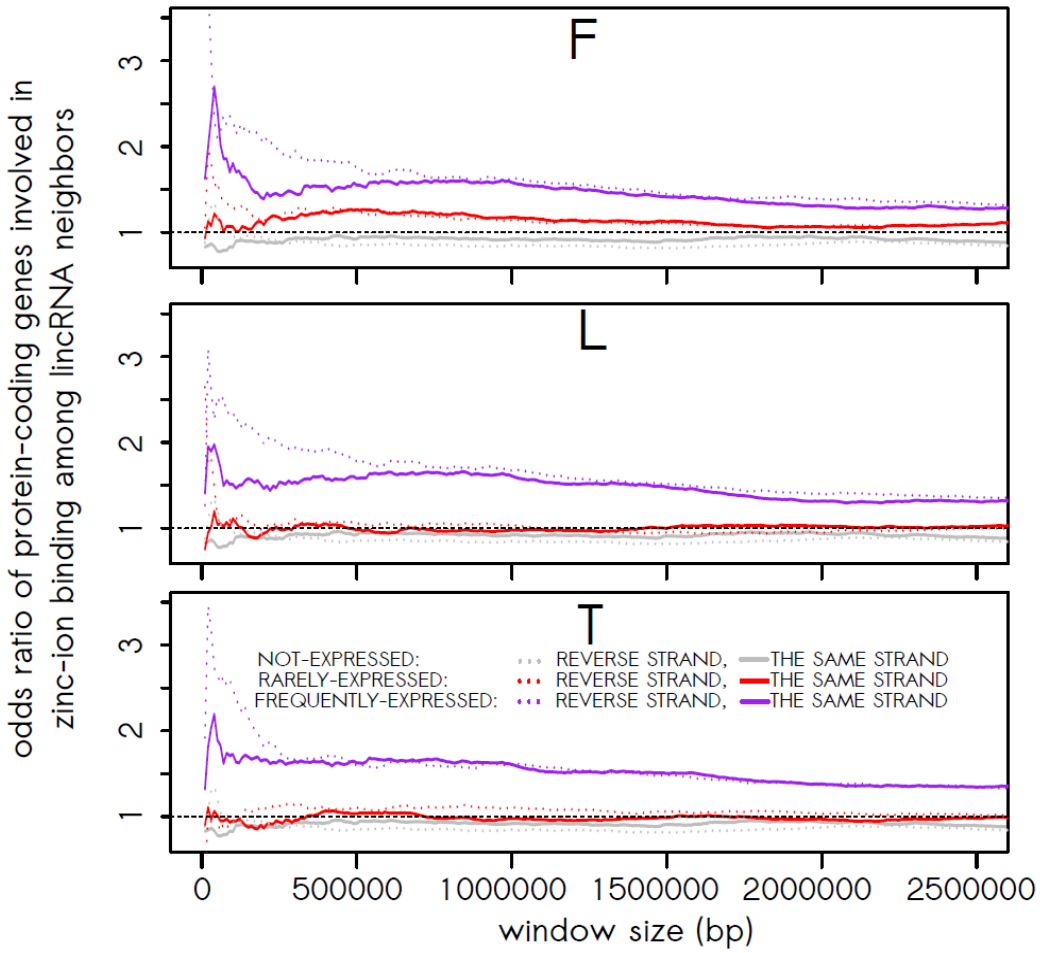

Figure S5. Enrichment of lincRNA's protein-coding neighbors by GO category 0008720 as a function of window size (upstream and downstream distance around lincRNA), level of expression of lincRNA (frequently- (>90% of samples), rarely- (20-90% of samples), not-expressed) and the strand (the same or reverse). Among four annotated (not necessary expressed) immediate protein-coding neighbors of frequently-expressed lincRNAs we have revealed an excess of protein-coding genes from GO category '0008270 - zinc-ion binding' (all Fisher's odds ratios > 1.21, all P-values < 0.050). Since protein-coding genes, involved in zinc-ion binding are common neighbors of expressed/frequently-expressed lincRNAs on transcriptome/genome level, we have investigated the distribution of these neighbors around lincRNAs using different window sizes. We discovered that (i) frequently-expressed lincRNAs have the most enriched neighbors, while not-expressed lincRNAs have a deficit of these genes, (ii) the highest enrichment is observed for neighbors located closest to lincRNAs and (iii) neighbors located on the opposite strand to a particular lincRNA tend to be more enriched.

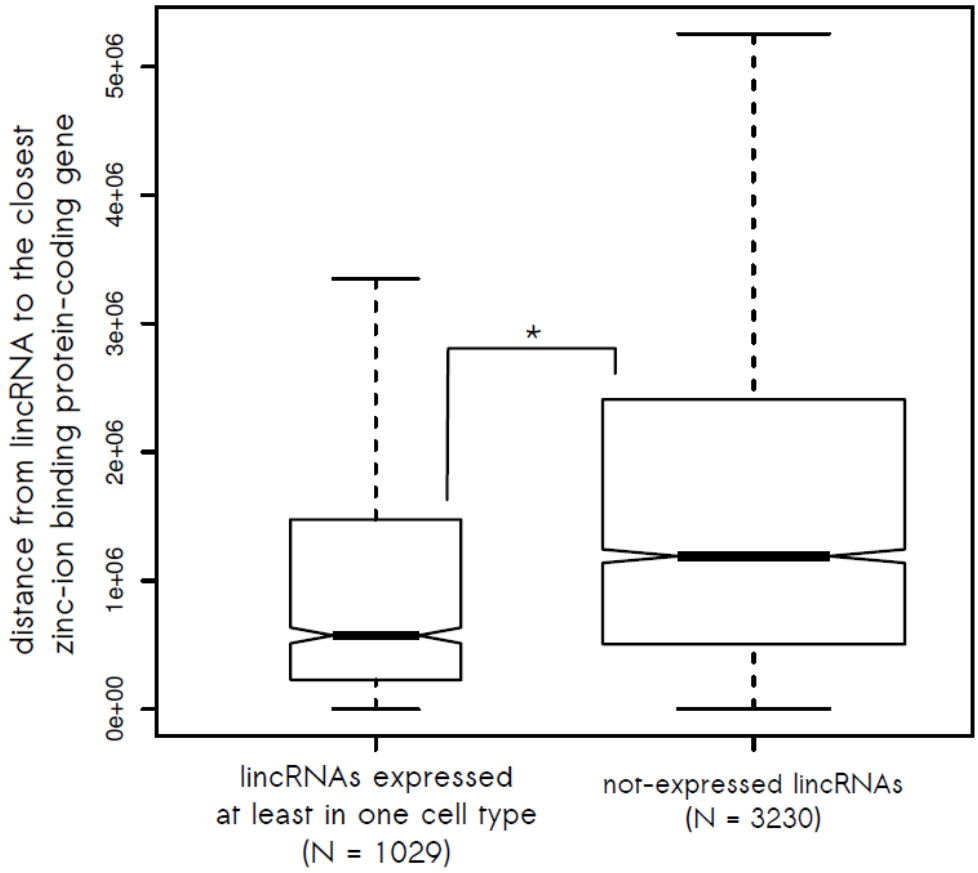

Figure S6. Expression status of lincRNA (expressed at least in one of three cell types or not-expressed at all) depends on distance of lincRNA to the closest protein-coding gene, involved in zinc ion binding (GO # 0008720). To corroborate the importance of the location of genes from the zinc ion binding (GO # 0008720) category on expression level of lincRNAs we have demonstrated that lincRNAs expressed at least in one tissue are located significantly closer to the proximal representative of the GO category than lincRNAs not expressed in any of the tissues. The predominant location of expressed lincRNAs close to genes involved in zinc ion binding may imply an involvement of lincRNA in transcriptional control, since 40% of zinc binding proteins in human proteome are transcription factors (Andreini, C., Banci, L., Bertini, I., and Rosato, A. (2006). Journal of Proteome Research *5*, 196–201). This preferred location of the frequently-expressed lincRNAs might be a result of selection, which optimized the position of the most important lincRNA genes across the genome facilitating cis regulation between lincRNAs and zinc-binding protein-coding genes and thus supporting the “flexible scaffold” model in *cis*.

GO clusters	F	L	T
0008270: zinc ion binding	5.7e-6	1.8e-8	6.7e-9
0046914: transition metal ion binding	7.3e-6	9.4e-7	6.2e-8
0003677: DNA binding	2.5e-8	1.2e-8	1.6e-7
0006355: regulation of transcription	3.0e-7	6.7e-8	5.9e-5
0006351: transcription	9.1e-8	9.1e-8	1.1e-4
0051252: regulation of RNA metabolic process	5.9e-5	5.6e-5	1.6e-4
0043169: cation binding	4.7e-2	2.8e-2	2.9e-3
0046872: metal ion binding	4.3e-2	3.0e-2	3.6e-3
0043167: ion binding	4.6e-2	3.1e-2	4.0e-3
0003700: transcription factor activity	6.3e-1	8.8e-1	9.0e-1

Table S1. Functional annotation of protein-coding genes, transcriptionally co-localized with lincRNAs. In order to assess what functional classes of protein coding genes are enriched near expressed lincRNAs we ranked all expressed protein-coding genes according to their distance to the closest expressed lincRNA (irrespectively of the strand and the direction of transcription). Functional annotation of the top 10% of protein-coding genes located closer to expressed lincRNAs (1278, 1235 and 1293 protein-coding genes located closer than 364, 343 and 349 kb for F, L and T) revealed an over-representation of protein-coding genes associated with DNA binding, transcription, regulation of transcription and regulation of RNA metabolic processes in all 3 tissues. GO terms, coming from the first functional cluster of DAVID (Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Nature Protocols *4*, 44–57) with enrichment scores 11.75, 12.19 and 11.57 for F,L and T, and their Benjamini adjusted p values for each tissue are presented.