SUPPLEMENTARY INFORMATION

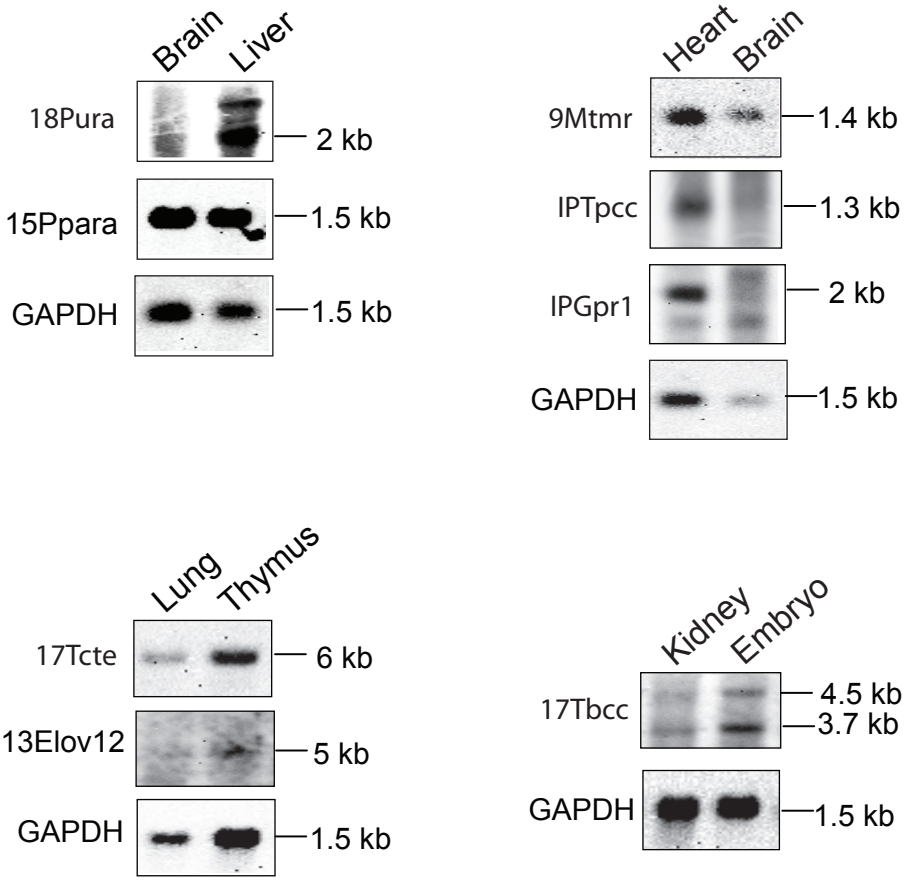


Figure 1: Northern blot validation of lincRNAs. RNA blot analysis was performed on 7 tissues (Brain, Liver, Lung, Thymus, Heart, Embryo, and Kidney). Hybridization of 12 lincRNAs are shown for randomly selected lincRNAs. A reference name for each lincRNA is shown on the left, the predicted sizes are indicated on the right. For each tissue, GAPDH was hybridized as a control.

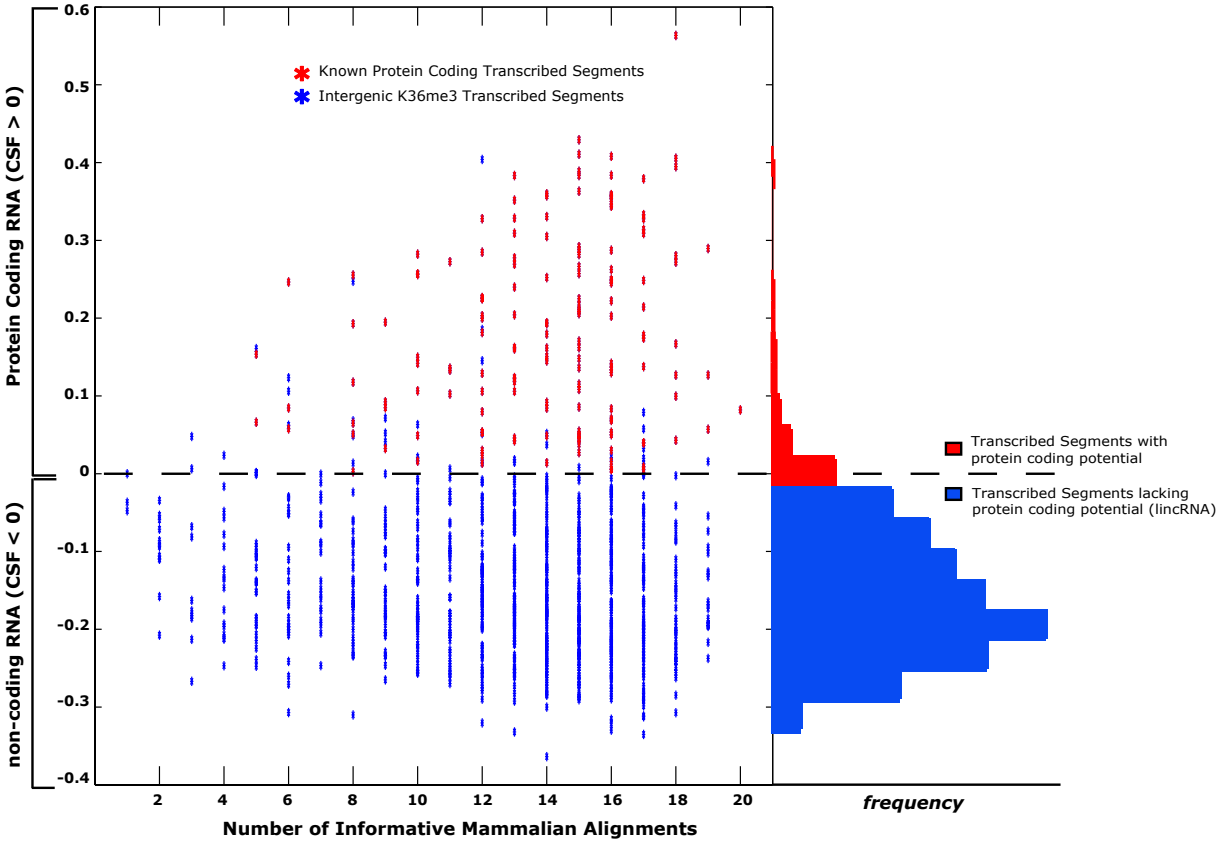


Figure 2: Transcribed Segments in intergenic K4/K36 domains do not have coding potential. The normalized CSF score for each exon determined from our tiling microarray is plotted. Red dots represent known protein coding controls and blue represents novel exons both determined by hybridization to a tiling array. The y-axis represent the normalized CSF score (CSF<0, noncoding, CSF>0 protein coding). The x-axis is the number of mammalian species that contributed to the determination of coding potential. The frequency graph of the right indicates the proportion of novel intergenic exons that were seen at each value. The horizontal dashed line indicates the threshold for protein-coding and non-coding determination.

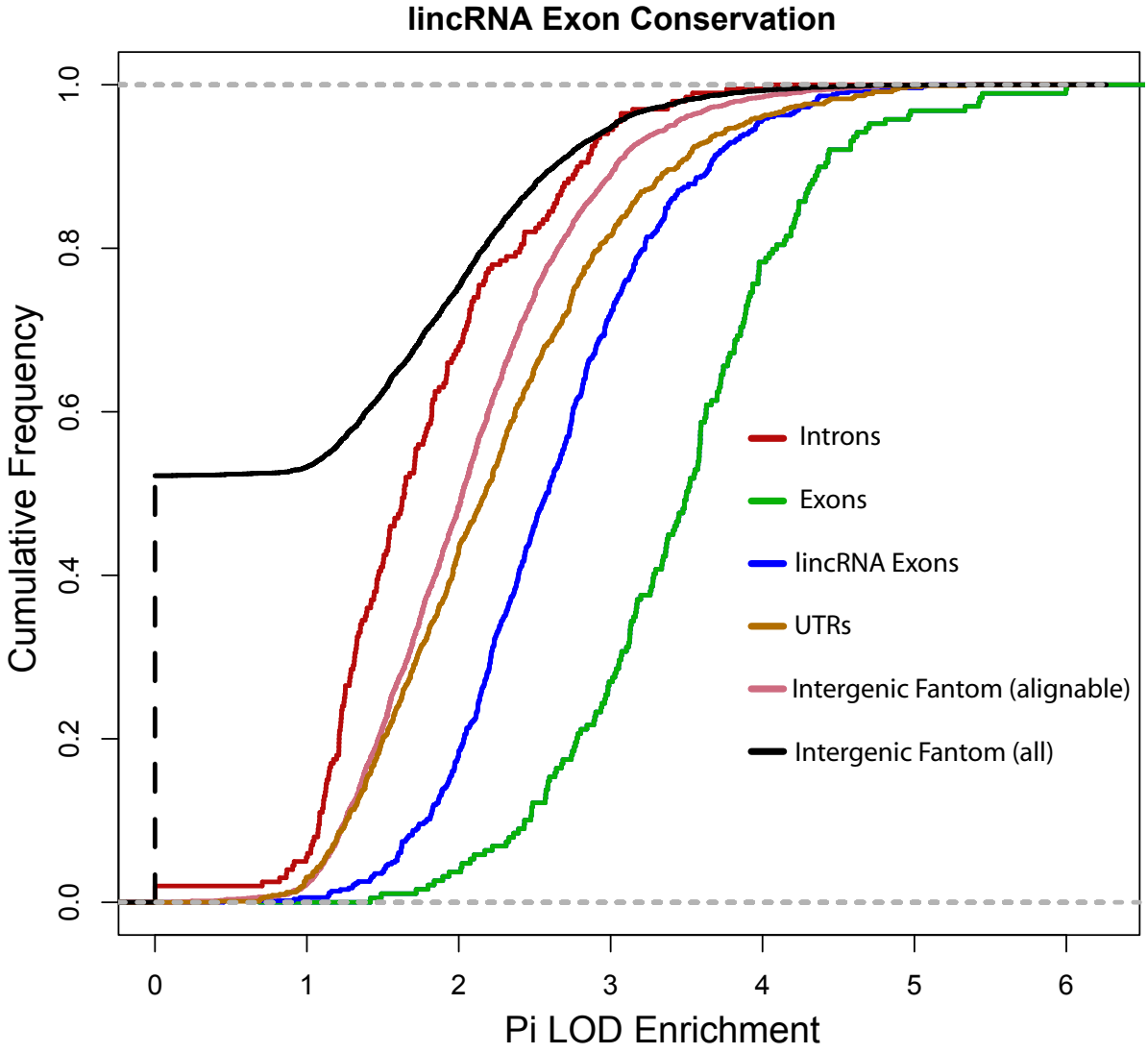
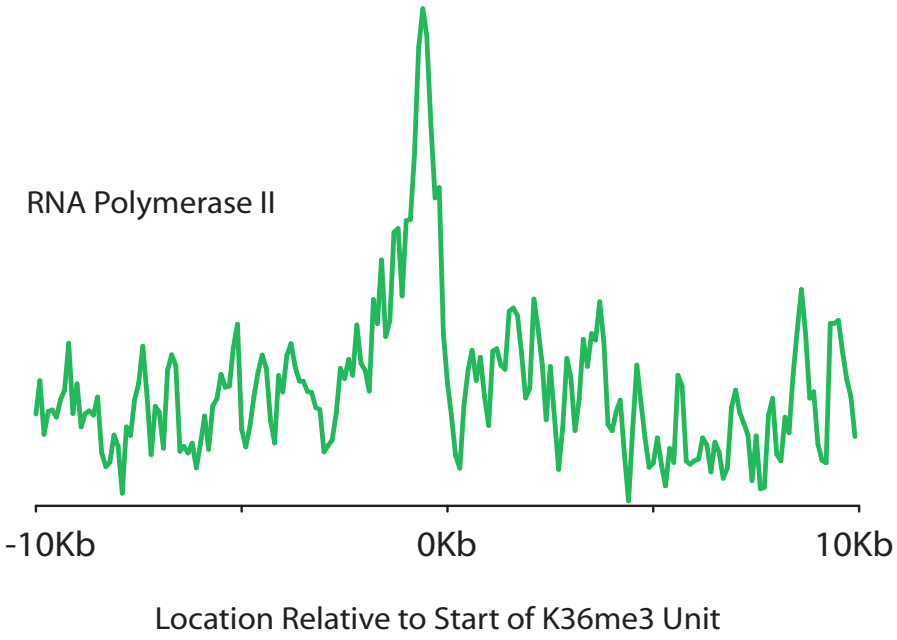


Figure 3. lincRNA Exon Conservation Compared with FANTOM and UTRs. Sequence conservation across 21 mammalian species is plotted cumulatively across each exon in the lincRNA transcript (Blue), protein coding exons (Green), and introns of protein coding genes (Red), as well as alignable FANTOM exons (pink), all FANTOM exons (black), and UTRs (orange). The X-axis is the enrichment of the log odds score of the Pi estimator (see methods) normalized by random genomic regions, thus larger LOD scores are more highly conserved.



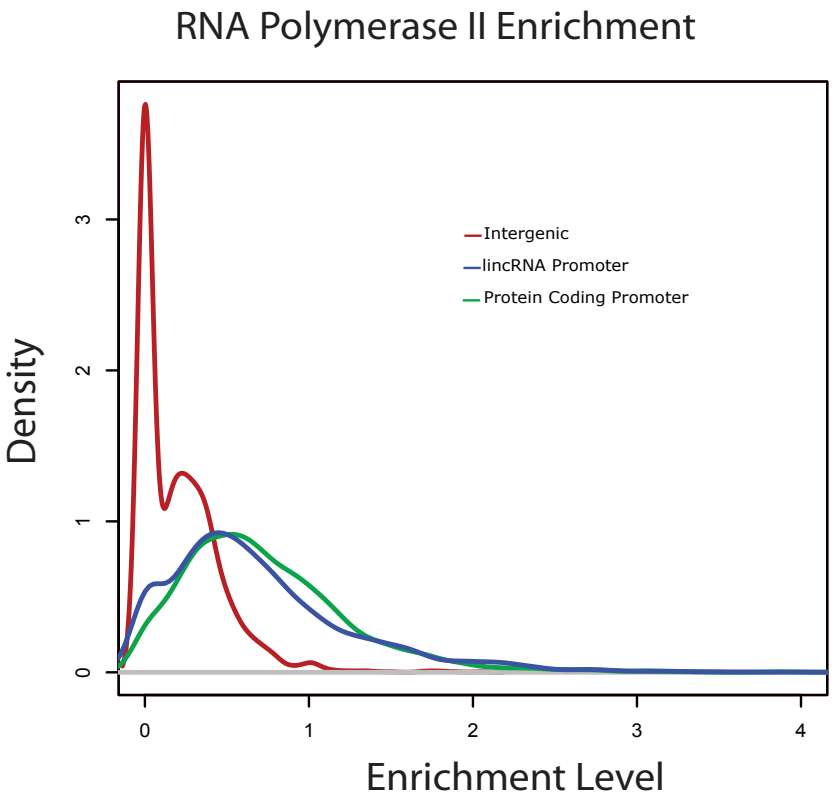


Figure 4. lincRNA Promoters Are Enriched for RNA Polymerase II binding. (**A**) The average enrichment of RNA Polymerase II as a function of the distance from the start of the K36me3 unit. (**B**) Distibutions of the enrichment of RNA Polymerase II over the promoter regions of lincRNAs (blue), protein coding genes (green), and random (non-masked) genomic regions (red).

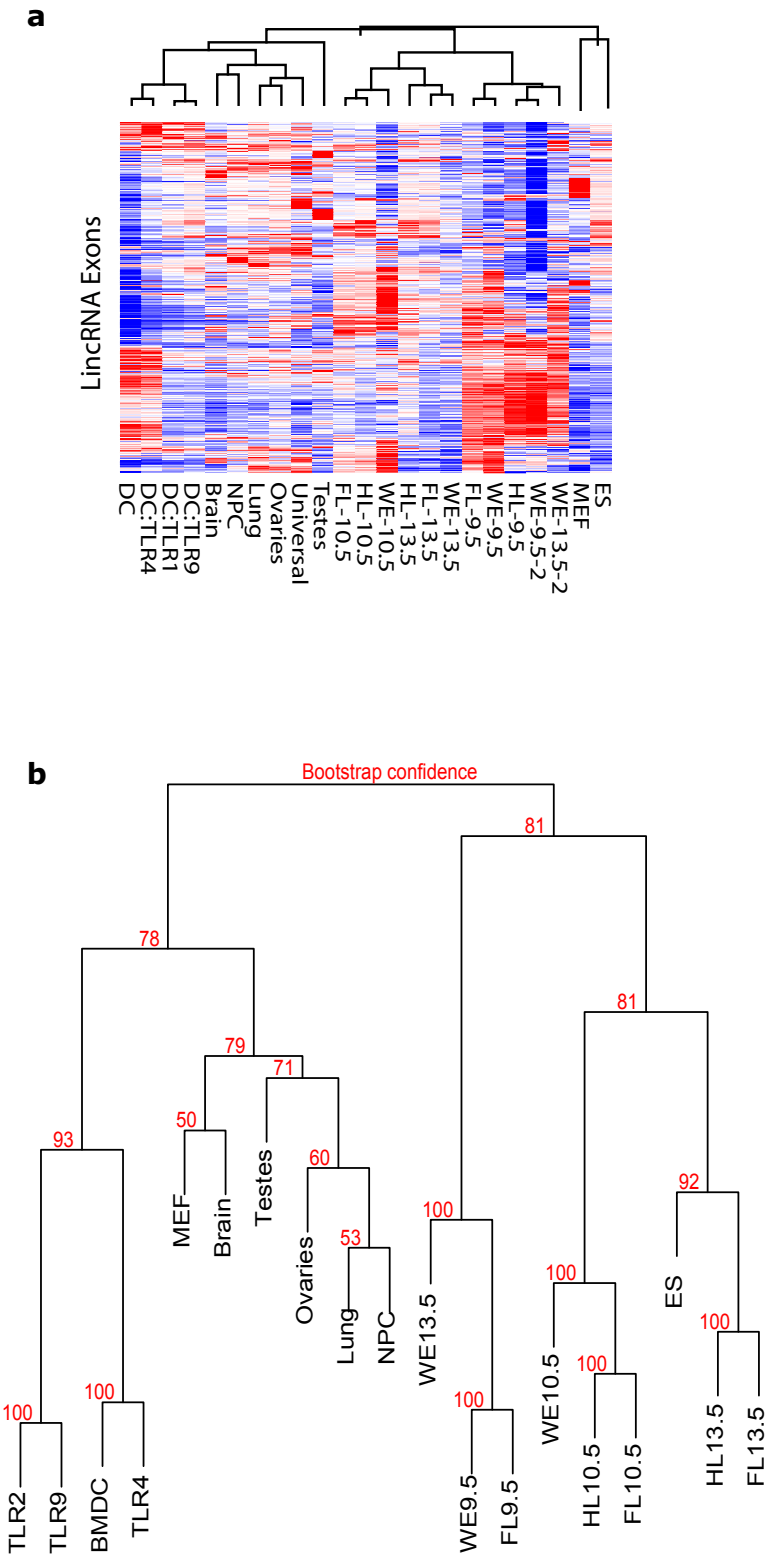


Figure 5. lincRNAs cluster experiments in biologically meaningful ways. A hierarchical clustering heatmap of lincRNAs across 19 conditions (Whole Embryo E9.5, E10.5, E13.5, Hindlimb E9.5, E10.5, E13.5, Forelimb E9.5, E10.5, E13.5, Embryonic Stem Cells, Embryonic Fibroblasts, Neural Progenitor Cells, Ovary, Testis, Lung, Brain, Dendridic Cells, TLR2, TLR4, TLR9) blue represent low expression and red represents high expression. Bootstrapping was performed to determine the reproducibility of these clusters. The hierarchical tree and associated bootstrap values are indicated along the tree (bottom).

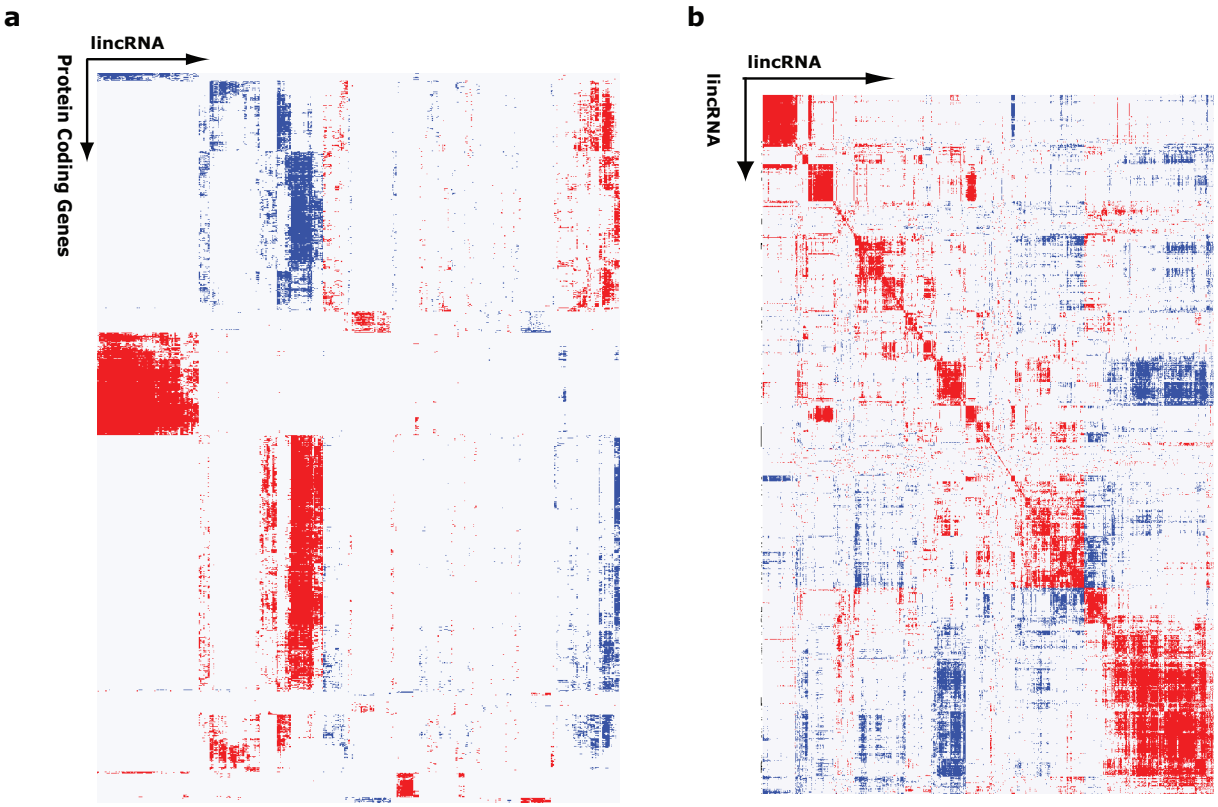


Figure 6. lincRNAs are correlated with groups of protein coding genes. (**A**) A hierarchically clustered heatmap of the correlation between lincRNAs and protein coding genes across 19 conditions is shown. Blue represent negative correlation, red represents positive correlation, and white indicates no correlation. lincRNAs are indicated along the columns and protein coding genes along the rows. (**B**) Correlation matrix showing numerous blocks of lincRNAs with correlated expression (red) as well as blocks of lincRNAs with anticorrelated expression (blue).

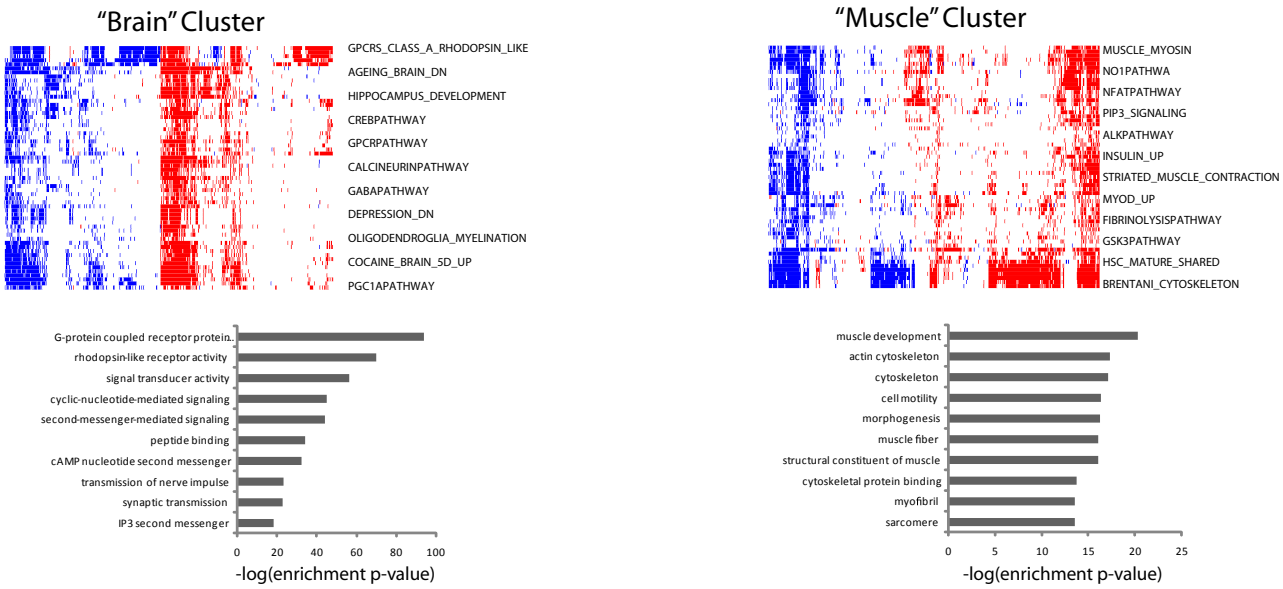


Figure 7. lincRNAs are associated with many biological processes. We performed biclustering on lincRNAs by functional terms and identify many significant biological processes including (**a**) a brain cluster (**b**) muscle development cluster and (**c**) miRNA regulated cluster. Heatmaps of these biclusters, lincRNAs are indicated by columns and the Gene Set terms that they associate with are indicated in rows. Blue boxes represent negative association, Red boxes represent positive association, and white boxes represent no association. Representative functional terms are indicated on the right of each heatmap. Gene Ontology was used to determine general categories of each bicluster and the results are plotted as a bar graph below each cluster. The length of each bar represent the –log(p-value) for the enrichment of each term._

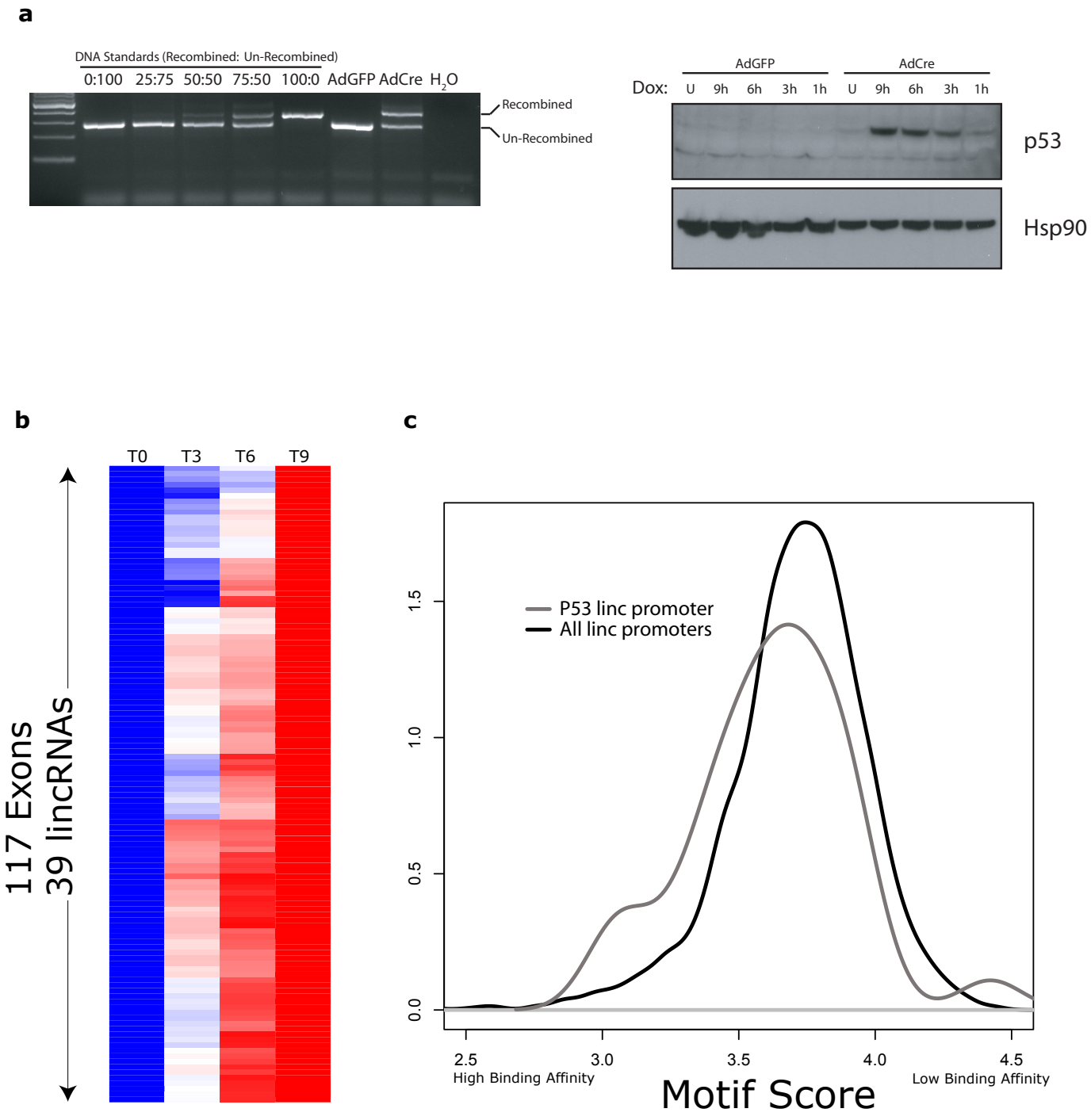
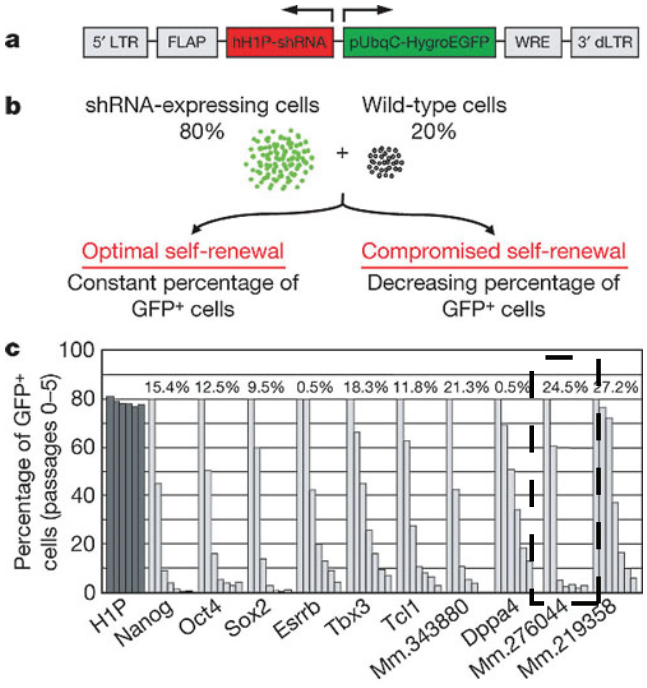


Figure 8. p53 regulated lincRNAs . (**a**) DNA Gel and Western blot showing the accurate reactivation of p53 as described27 (**b**) Shows a heatmap of 39 lincRNAs that show a temporal induction across a p53 induction time course. (**c**) These RNAs are enriched for the p53 binding motif. p53 induced lincRNA promoters (red) compared with all lincRNA promoters (blue).



Ivanova et al. 2006

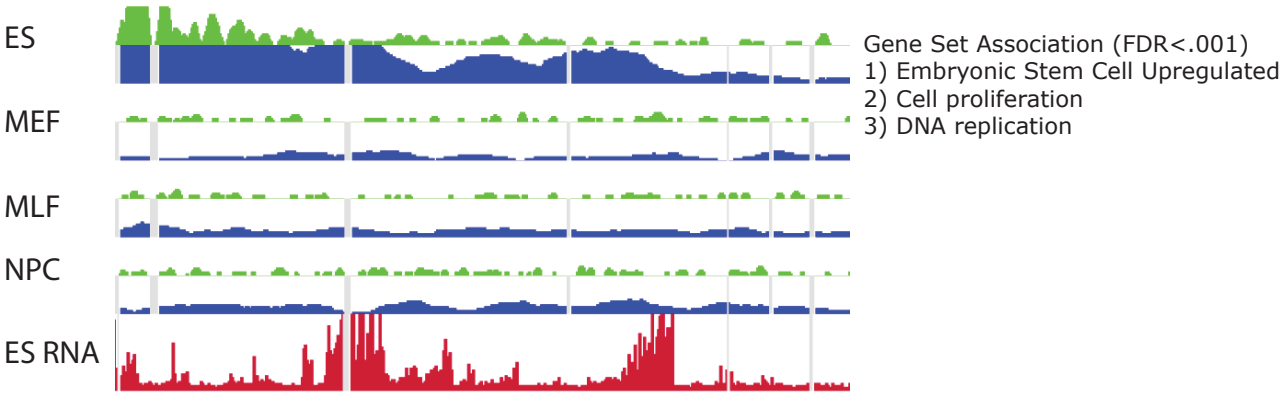


Figure 9. A lincRNA functionally associated with cell proliferation is ES cells is required to maintain proper cell proliferation rates in ES cells. A figure from Ivanova et al. 2006 describing the screen they performed to identify genes involved in ES pluripotency (top). One of the top 10 hits is a lincRNA expressed only in ES cells and functionally associated with cell proliferation in ES cells (boxed gene). The K4-K36 across ES, MEF, MLF, and NPC is shown along with the RNA peaks identified in ES (red).

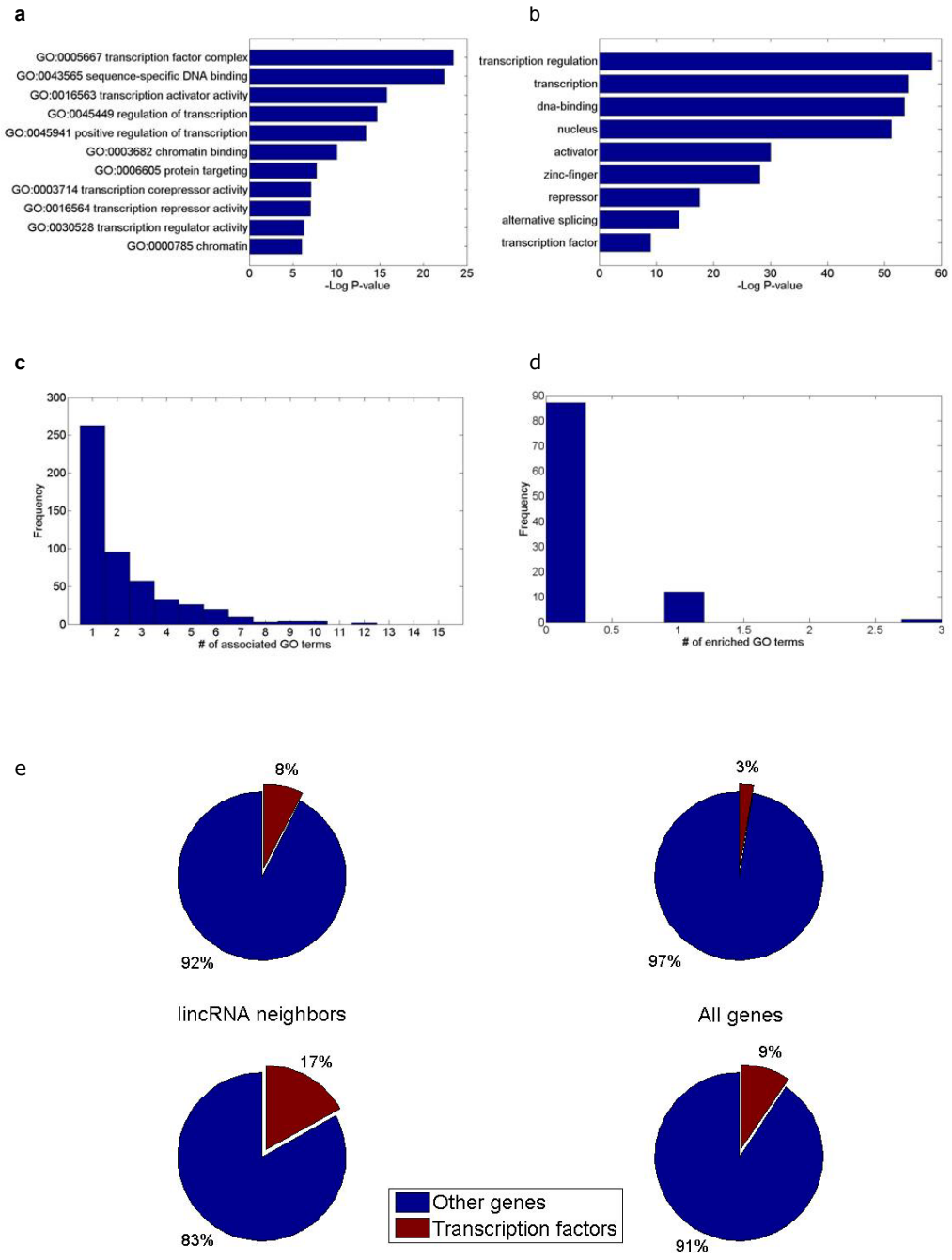


Figure 10. lincRNAs are positionally enriched near transcription factors. (**a**) Gene Ontology (GO) enrichment was computed for all lincRNA neighbors and are plotted as the –log(p-value). (**b**) Swiss-Prot key words from DAVID (http://david.abcc.ncifcrf.gov/) and enriched domains are plotted as the –log(p-value). (**c**) The distribution of enriched GO terms for protein coding genes that neighbour lincRNAs are displayed. (**d**) The distribution of enriched GO terms for randomized lincRNA datasets (methods) and their associated gene neighbours is indicated. (**e**) Pie charts show the proportion of genes annotated as transcription factors (red) and other genes (blue). The proportion of lincRNA neighbour genes are shown on the left as compared to all known protein-coding genes (right).

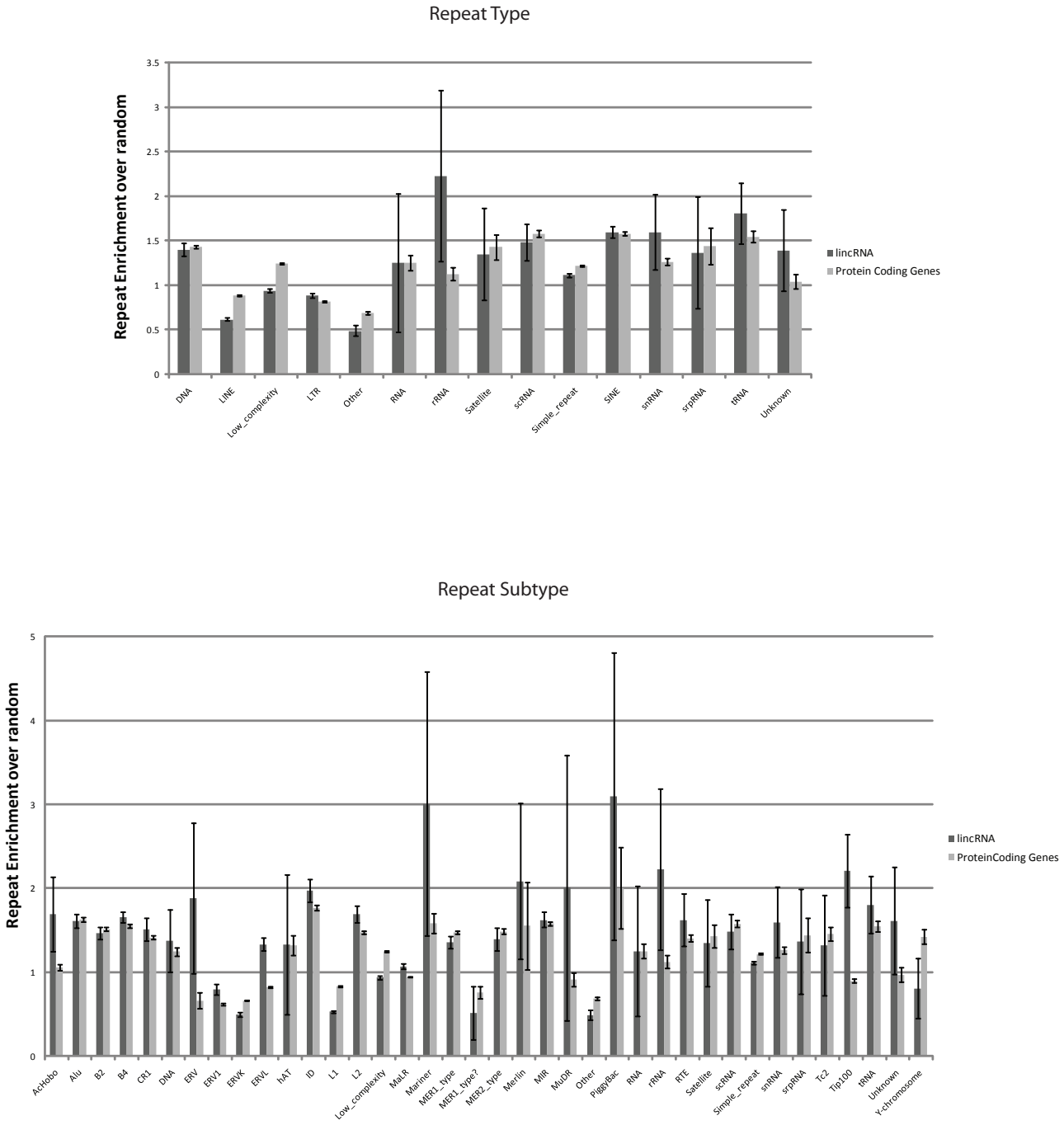


Figure 11. Repeat Content Enrichment in lincRNAs. A barchart plotting the enrichment of repeat elements of various types (top) and subtypes (bottom) are plotted for lincRNAs (dark gray bars) and protein-coding genes (light gray). The average enrichment is indicated by the height of each bar and the error bars indicate the standard deviation of the estimate. Enrichment is defined as the number of repeats divided by the expected number of repeats for random regions of equal size.