Published in final edited form as:
Nature. 2019 June 06; 571(7766): 510–514. doi:10.1038/s41586-019-1341-x.

Developmental dynamics of lncRNAs across mammalian organs and species

Ioannis Sarropoulos[1], **Ray Marin**[2], **Margarida Cardoso-Moreira**[#1,2], **Henrik Kaessmann**[#1]

[1]Center for Molecular Biology of Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, D-69120 Heidelberg, Germany [2]Center for Integrative Genomics, University of Lausanne, CH-1015 Lausanne, Switzerland

[#] These authors contributed equally to this work.

Abstract

While many long noncoding RNAs (lncRNAs) have been identified in human and other mammalian genomes, there has been limited systematic functional characterization. In particular, the contribution of lncRNAs to organ development remains largely unexplored. Here we analyze the expression patterns of lncRNAs across developmental timepoints in seven major organs, from early organogenesis to adulthood, across seven species (human, macaque, mouse, rat, rabbit, opossum, and chicken). Our analyses identified ~15,000-35,000 candidate lncRNAs in each species, most of which show species specificity. We characterized expression patterns of lncRNAs across developmental stages, and found many with dynamic expression patterns across time that show signatures of enrichment for functionality. During development, there is a transition from broadly expressed and conserved lncRNAs towards an increasing number of lineage- and organ-specific lncRNAs. Our study provides a resource of candidate lncRNAs and their patterns of expression and evolutionary conservation across mammalian organ development.

Previous studies identified numerous long noncoding RNAs (lncRNAs) in human1–4 and other mammals5–8. However, molecularly characterized cases are limited9 and the functionality of most loci remains uncertain10. Cross-species genomic comparisons provide a powerful framework for the large-scale identification of putatively functional lncRNAs, as these should carry signatures of evolutionary constraint11,12. Although the physical proximity13,14 and co-expression of lncRNAs with developmental regulators6, together with individual paradigms15–17, have long suggested a contribution of lncRNAs to

Correspondence and requests for materials should be addressed to M.C.M. (m.moreira@zmbh.uni-heidelberg.de) or to H.K. (h.kaessmann@zmbh.uni-heidelberg.de).

Author Contributions

M.C.M. and H.K. conceived and organized the study based on an original design by H.K. R.M. performed the lncRNA annotation and orthology assignment. I.S. performed all other analyses, under the supervision of M.C.M. and H.K. I.S., M.C.M. and H.K. wrote the manuscript, with input from R.M.

Competing interests

The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

mammalian development, data limitations confined previous evolutionary studies of lncRNAs to adult organs6–8. Here, we utilize a dataset of bulk transcriptomes in seven major organs across developmental stages, from early organogenesis to adulthood, across seven species (reported in an accompanying publication, 18) to examine the contribution of lncRNAs to developmental expression programs.

Developmental lncRNA atlases

To assess the relevance of lncRNAs for mammalian organ development we analyzed an RNA-seq dataset18 covering the development of seven major organs (forebrain/cerebrum — hereinafter referred to as “brain”, cerebellum, heart, kidney, liver, ovary, and testis) from early organogenesis to adulthood in seven species (human, macaque, mouse, rat, rabbit, opossum, and chicken; Fig. 1a). Using this dataset we annotated candidate lncRNAs as spliced and long transcripts (>200 nucleotides) with no detectable protein-coding potential and reconstructed homologous lncRNA families based on sequence similarities between species (Methods; Fig. 1b; Extended Data Fig. 1a-b; Supplementary Data 1; Supplementary Tables 1-8).

We identified ~15,000-35,000 candidate lncRNAs of various genomic classes in each species (Extended Data Fig. 2a). We recovered ~50% of the human and murine lncRNA and antisense transcripts from Ensembl19, and detected 24,951 and 21,263 novel lncRNAs, respectively (Fig. 1c). The distribution of genomic classes and spatial expression patterns is indistinguishable between the newly identified and the previously annotated lncRNAs, suggesting our repertoire extensions are unbiased (Extended Data Fig. 2b). While most of our lncRNAs are species-specific6–8, we identified 8,953 conserved human lncRNAs (Fig. 1b). The sensitivity of our lncRNA family detection was similar to previous studies8 and synteny conservation was comparable to that of protein-coding genes (Extended Data Fig. 2c-e). The lncRNA expression profiles and gene models can be explored interactively: lncrnas.kaessmannlab.org.

Depending on the species, 35-60% of robustly expressed lncRNAs (i.e., RPKM ≥ 1 in at least one sample) show maximal expression in the testis (Extended Data Fig. 2f), confirming the disproportional contribution of this organ to lncRNA repertoires1,6–8. However, this peculiarity is limited to the adult organ; the number of lncRNAs expressed in the developing testis is indistinguishable from the remaining organs (Extended Data Fig. 2g).

Features of developmentally dynamic lncRNAs

We identified lncRNAs with significant differential expression through time (termed “developmentally dynamic”) using a regression approach (maSigPro; Methods)20. The ability to detect dynamic expression depends on multiple factors, some of which have been associated with increased functional relevance of lncRNAs. These include robust expression levels21, transcript stability22, reproducibility between biological replicates, and consistent changes in expression across developmental stages10 (Extended Data Fig. 3a). While dynamic expression is not sufficient to claim functionality, we reasoned that it would allow us to enrich for functionally relevant lncRNAs. As the disproportionate lncRNA expression

of the adult testis6,12,23 is associated with a pervasive chromatin environment that also allows the transcription of putatively non-functional elements23, we excluded post-puberty testis samples from this estimation.

Most protein-coding genes (73-90% depending on the species) but only a fraction of lncRNAs (16-38%) show developmentally dynamic expression (Extended Data Fig. 3b). Contrary to the highly variable total number of lncRNAs, the numbers of developmentally dynamic lncRNAs are similar across species (Fig. 2a). Notably, large proportions of dynamic lncRNAs in human (2,998, 51%) and mouse (4,188, 74%) are not annotated in Ensembl19. While most dynamic lncRNAs (51-63%) are differentially expressed in a single organ they show broader and higher expression than non-dynamic lncRNAs (Extended Data Fig. 3c-f).

Developmentally dynamic lncRNAs show an overrepresentation of divergent, downstream sense and antisense transcripts, which results in a closer proximity to protein-coding genes (Extended Data Fig. 4a-b). However, all genomic classes substantially contribute to the total number of dynamic lncRNAs (Extended Data Fig. 4a) and the differences between classes mostly disappear after controlling for maximum expression (Extended Data Fig. 4c). Dynamic lncRNA transcripts are also longer (Fig. 2b) and contain more exons (Extended Data Fig. 4d), suggesting selection for splice sites and against premature polyadenylation signals, as well as a higher capacity to accommodate modular RNA domains that facilitate interactions with proteins or other nucleic acids24.

Evolutionary conservation provides a strong line of evidence for lncRNA functionality6,11–13. We observed a significant increase in the fraction of developmentally dynamic genes for older lncRNA groups (Fig. 2c; $P < 0.01$, two-sided Fisher's exact test). As the overlap with protein-coding genes and regulatory elements can lead to the overestimation of lncRNA evolutionary age, we repeated this analysis excluding antisense and divergent transcripts, and lncRNAs that overlap transcribed enhancers25, with similar results (Extended Data Fig. 4e; $P < 0.05$, two-sided Fisher's exact test). The enrichment of dynamic transcripts amongst older lncRNAs also remained significant after controlling for maximum expression, even for lowly expressed lncRNAs (Extended Data Fig. 4f-g; $P < 0.05$, two-sided Fisher's exact test). Overall, our analyses suggest a clear association between developmentally dynamic expression and evolutionary conservation.

Next, we assessed the extent of spatiotemporal expression similarity between species. Human lncRNAs with a dynamic mouse ortholog are more likely to also be dynamic (Extended Data Fig. 4h), and lncRNAs that are dynamic in both species show almost as high expression similarity as protein-coding genes, even after excluding antisense and divergent lncRNAs (Fig. 2d; Extended Data Fig. 4i). To assess the effect of evolutionary age on the conservation of lncRNA spatiotemporal profiles across a wider phyletic range, we analyzed lncRNAs dynamic in mouse and rat. We observed an increase in expression similarity with lncRNA age (Extended Data Fig. 4j), in agreement with the slow turnover of transcription and tissue-specificity of conserved lncRNAs7,26.

Finally, we sought to more directly examine the functional relevance of dynamic lncRNAs. In a set of molecularly characterized, functional lncRNAs from lncRNAdb27, the fraction of dynamic genes is 76%, four times higher than among all human lncRNAs and close to the fraction of dynamic protein-coding genes (86%, Fig. 2c). This enrichment remained significant after controlling for maximum expression (Extended Data Fig. 4f; $P = 0.037$, two-sided Fisher's exact test). To account for ascertainment biases, like the preferred experimental characterization of broadly expressed and conserved lncRNAs, we also examined a set of lncRNAs associated with cell proliferation phenotypes based on an unbiased CRISPRi screen in human cell lines21. Despite the different physiological context and high cell line specificity of the reported results, we found that lncRNAs in the screen libraries that were also present in our annotation had an increased likelihood of exhibiting a cell proliferation phenotype ("hit", Extended Data Fig. 4k; $P = 2.2 \times 10^{-16}$, two-sided Fisher's exact test). Moreover, we observed a significant, albeit small, enrichment of hits among dynamic lncRNAs (Fig. 2e; Extended Data Fig. 4k; $P = 0.02$, two-sided Fisher's exact test) further supporting their enrichment for functional loci.

Regulatory landscape of dynamic lncRNAs

We next investigated whether the developmentally dynamic expression of lncRNAs is also reflected in more complex transcriptional regulation (Methods). As expected6,28 the promoters of protein-coding genes contain the most transcription factor (TF) binding sites29 (Fig. 2f). However, the promoters of dynamic intergenic lncRNAs are bound by more TFs than those of non-dynamic lncRNAs, suggesting a stronger and more complex transcriptional regulation (Fig. 2f; Extended Data Fig. 5a).

To assess the relevance of this increased transcriptional regulation during development, we estimated the fraction of dynamic lncRNA promoters bound by each TF (termed "binding frequency"). We identified three major classes: TFs with high binding frequencies for lncRNAs dynamic in the nervous tissues, heart or liver (Fig. 2g; Extended Data Fig. 5b). For tissue-specific TFs, we observed a high concordance between the organ where the TFs are maximally expressed and the binding frequency for lncRNAs dynamic in that organ (Fig. 2g; Extended Data Fig. 5c). Despite their ubiquitous expression, well-established regulators of cardiac development30, such as *Nkx2-5, Mef2d* and *Gata4*, also predominantly bind to promoters of lncRNAs dynamic in the heart (Fig. 2g). Overall, these results show that the increased transcriptional regulation of dynamic lncRNAs matches the organ in which they are expressed.

Expression patterns during organ development

Organ development is punctuated by periods when large numbers of protein-coding genes change their expression levels18. These periods are associated with the establishment of organ identity early in development and with the transition to mature organ-specific functions around birth18. Strikingly, the stages where dynamic lncRNAs show the greatest differential expression coincide with these periods of greater transcriptional change, even when only considering lncRNAs located more than 100 kb away from the closest protein-coding gene (Fig. 3a; Extended Data Fig. 6a-b). Although we cannot exclude a contribution

Europe PMC Funders Author Manuscripts

from proximal developmental enhancers31, the enrichment of dynamic lncRNAs for functionally relevant features (Fig. 2) argues against the prevalence of non-autonomous expression for most loci.

Motivated by the similar temporal dynamics between protein-coding genes and lncRNAs, we assigned putative functions to dynamic lncRNAs based on their co-expression with protein-coding genes, i.e. through "guilt by association"6,13. Across organs, the co-expression clusters with the highest fraction of lncRNAs consistently showed similar developmental trajectories and were associated with developmental functions and adult organ physiology (Extended Data Fig. 7; Supplementary Tables 9-10). By contrast, lncRNAs contributed the least to clusters associated with housekeeping genes, in agreement with the hypothesis that few lncRNAs are involved in essential cellular functions10.

Early vs. late development

In the developmental period studied here, the transcriptomes of different organs share strong commonalities at the earliest stages and then gradually diverge into distinct, organ-specific developmental programs18. In parallel with this divergence of gene expression programs, the number of dynamic lncRNAs expressed in each organ steadily increases (Fig. 3b). In contrast, the fraction of lncRNAs showing selective preservation (i.e., those with an age > 80 million years) decreases with time (Fig. 3c). Consistently, the expression similarity between lncRNAs dynamic in both human and mouse, also declines during development (Extended Data Fig. 8a). Thus, although the absolute number of lncRNAs expressed during early organ development is lower than in postnatal stages, these genes have been under stronger selective constraints. Notably, levels of lncRNA sequence and expression conservation are particularly high in nervous tissues and lower in liver and gonads (Extended Data Fig. 8b-c), as observed for protein-coding genes18.

Early-expressed protein-coding genes also show higher sequence and expression conservation, which was suggested to result from the higher pleiotropy (broader spatiotemporal expression) of early-expressed genes and associated increased functional constraints18. Consistently, we found that lncRNAs expressed early in development are more broadly expressed across organs than lncRNAs expressed late (Fig. 3d; Extended Data Fig. 8d). We also found that lncRNAs expressed earlier in development are more likely to be characterized as functional by lncRNAdb27 and to result in a cell proliferation phenotype in the CRISPRi screen21 (Fig. 3e-f; Extended Data Fig. 8e-f). This enrichment is consistent with our ‘guilt-by-association’ analysis, which associated early-expressed lncRNAs with broad cellular functions (Extended Data Fig. 7). On the other hand, late-expressed dynamic lncRNAs still retain signatures of functional enrichment when compared to non-dynamic lncRNAs (Extended Data Fig. 8g). Their organ-specific expression (Fig. 3d; Extended Data Fig. 8d) suggests they may be involved in more specialized functions and thus under weaker functional constraints than early-expressed lncRNAs.

Collectively, our analyses revealed a distinction between lncRNAs expressed early and late in organ development. While fewer lncRNAs are expressed during early stages, these genes are more pleiotropic and are under stronger evolutionary constraint at the sequence and

expression levels, consistent with broader functions. By contrast, most lncRNAs are expressed in later stages and are characterized by higher organ- and lineage-specificity, suggesting milder effects on developmental programs and phenotypes.

Co-expression with adjacent protein-coding genes

Several well-characterized lncRNAs, such as *XIST* and *Airn*, are known to act *in cis*, regulating the expression of their immediate neighbors9. However, the extent of such effects at the genomic scale remains unresolved1,2,26,32. We examined this question within the context of organ development using our set of dynamic lncRNAs. We observed a significantly higher expression correlation between dynamic lncRNAs and their adjacent protein-coding genes compared to mRNA-mRNA controls (Fig. 4a; $P = 2.2 \times 10^{-16}$, two-sided Wilcoxon's signed-rank test; Methods; Extended Data Fig. 9a-b). Although the distance between genes impacts the degree of their correlation, we found an excess of positive correlations for lncRNA-mRNA pairs for distances up to 100 kb (Extended Data Fig. 9c-d). We obtained similar results excluding bidirectional and antisense lncRNAs, as few protein-coding genes are transcribed in such orientations (Extended Data Fig. 9e; $P = 2.2 \times 10^{-16}$, two-sided Wilcoxon's signed-rank test). Protein-coding genes significantly correlated with their neighboring lncRNA were enriched for developmental genes (Fig. 4b; Extended Data Fig. 9f), supporting the biological significance of the enrichment of lncRNAs near developmental regulators13. Consistently, our set of co-expressed lncRNAs is enriched for a set of "positionally conserved" lncRNAs that are linked to chromatin organization structures and are co-expressed with their adjacent developmental protein-coding genes in adult primary tissues and cancer samples33 (Fig. 4c; $P < 10^{-11}$, two-sided Fisher's exact test).

We identified 77 protein-coding genes co-expressed with an adjacent lncRNA in both human and mouse (Fig. 4d), a significant enrichment relative to the fraction of 1:1 orthologous protein-coding genes co-expressed with a lncRNA in each species ($P = 2.2 \times 10^{-16}$, hypergeometric test; Supplementary Tables 11-12). Compared to all co-expressed pairs, those detected in both species show an even stronger association with organ development (38% involved in the development of at least one organ; $P = 0.0002$, hypergeometric test; Methods). Thus, co-expression between developmental regulators and their adjacent lncRNAs is a feature shared between species.

We note that the observed correlations are not sufficient to infer regulatory functions for lncRNAs, which requires experimental scrutiny9. Nonetheless, our results are consistent with studies suggesting that some mammalian lncRNAs act by influencing the expression of their adjacent genes9,33,34, having identified several lncRNAs (i.e., *GAS6-AS2*35, *DEANR1*33,36, *SSTR5-AS1*37, *EMX2OS*38 and *Dlx1as*39) previously implicated in the regulation of their neighboring protein-coding genes. The co-expressed lncRNA-mRNA pairs represent a reference set to facilitate future efforts for the experimental characterization of the *cis*-regulatory potential of lncRNAs.

Discussion

We utilized a dataset of transcriptomes across seven major organs and developmental stages to provide uniformly processed annotations and expression profiles for thousands of candidate lncRNAs. This extensive resource will facilitate future investigations of lncRNA biology (lncrnas.kaessmannlab.org). We also identified a set of developmentally dynamic lncRNAs that show multiple signatures of functional enrichment. We cannot exclude that some of our observations for dynamic lncRNAs might be explained by proximal or overlapping regulatory sequences, although these are typically transcribed into short-lived, unspliced and non-polyadenylated transcripts40, which are not included in our annotations. Furthermore, the enrichment of dynamic lncRNAs for longer and more complex transcripts argues against them being transcriptional or splicing by-products of regulatory sequences. Our analyses identified important differences in the contribution of lncRNAs to different stages of organ development and associated dynamic lncRNAs with putative functions. Future studies utilizing emerging technologies, such as single-cell41 or long-read RNA sequencing42, will further refine the annotations and expression profiles of mammalian developmentally dynamic lncRNAs.

Online Methods

Annotation of transcribed regions and identification of lncRNAs

We used a transcriptomic dataset covering the development of seven major organs (forebrain/cerebrum, hindbrain/cerebellum, heart, kidney, liver, ovary and testis) across seven amniote species (human, rhesus macaque, mouse, rat, rabbit, opossum, chicken), comprising a total of 1,993 (strand-specific) RNA-seq libraries18. Data for ovary development in rhesus macaque were not available. Genomic read alignments (BAM files)18 were filtered from reads partially mapping outside a contig or chromosome, mapping to more than 50 locations or having more than 50 nt with a phred score below 20 using samtools (0.1.18)43. The processed BAM files from the same species, organ and developmental stage (i.e., replicates) were merged to increase coverage and detection power. The merged BAM files were then used to identify transcribed regions for each sample (species, organ and developmental stage) with stringtie (1.2.3)44 using the following parameters:

```
stringtie <sample.bam> -o <sample.gtf> -p 2 -f 0.50 -m 200 -a 10 -j 3 -c 0.1
-g 10
```

The multiexonic transcripts from each sample were combined into a single assembly for each species using the tool cuffmerge from the Cufflinks package (2.2.1)45:

```
cufflinks -o <outprefix> -F 0.0 -q --overhang-tolerance 200 --library-
type=transfrags -A 0.0 --min-frags-per-transfrag 0 --no-5-extend --overlap-
radius 1 -p 20 <assembly_list.txt>
```

We first removed from each species' annotation the genes that overlap with Ensembl protein-coding genes in the same strand or that are shorter than 200 nts (Extended Data Fig.

1a). We then removed all genes with evidence for coding potential. The coding potential of our lncRNA candidates was estimated in three ways: using CPAT (1.2)46, RNAcode (0.3)47 and similarity with known proteins. CPAT uses an alignment-independent logistic regression model to detect lncRNAs based on sequence features46. To select a cutoff for the classification, we used a training set of randomly selected 10,000 protein-coding genes and 10,000 intronic regions. We selected the cutoff of 0.8 for mouse, rat and rabbit; 0.75 for human; and 0.70 for macaque, opossum and chicken. RNAcode uses multiple-species alignments to infer coding probability based on the rate of synonymous to non-synonymous mutations47. We generated customized whole genome alignments for each species in our dataset against seven other species (Supplementary Table 13), which we used to estimate coding potential. Transcripts with an open reading frame in the same strand, $P \leq 10^{-5}$ and alignment length ≥ 10 aminoacids were considered to be putatively coding (termed 'new putative-coding'). Finally, we used blastx (2.4.0)48 to translate each lncRNA in all possible six frames, which we then compared to known proteins in the databases UniProt (2016_04)49 and PFAM (v29)50. Transcripts with $E \leq 10^{-3}$, alignment length ≥ 10 aminoacids and identity ≥ 95% were considered 'new putative-coding'. Only genes that successfully passed all three filters for all their isoforms were included in our lncRNA annotation (Supplementary Data 1).

Gene expression quantification, specificity indexes and dynamic expression

For each species, we merged our lncRNA annotation with Ensembl's (v75 for human and v77 for all other species), after removing from the latter all genes overlapping lncRNAs in the same strand. We generated read counts using HTSeq (0.6.1)51, only allowing for uniquely mapped reads and only for the alignments of the 1,893 libraries that had a Spearman's correlation with its biological replicates ≥ 0.918. Since the samples used to quantify gene expression are a subset of the dataset used for the annotation of lncRNAs (see above), we removed lncRNAs showing no detectable expression in this smaller dataset. We calculated expression levels as cpm (counts per million) or RPKM (reads per kilobase of exon model per million mapped reads) (Supplementary Data 2) after normalizing the count data using the method TMM from the package edgeR (3.14.0)52. We also generated variance stabilized counts, using the respective transformation (VST) implemented in the package DESeq2 (1.12.4)53.

We estimated time- and tissue-specificity indexes using the Tau metric of tissue-specificity54. Tissue-Tau was calculated as previously described54, using for each organ the maximum expression observed during development. For time-specificity, we applied the same metric to the expression across developmental stages of the same organ. Time-specificity indexes were only calculated for the organs in which a gene is robustly expressed (i.e., RPKM > 1). Because time-specificity is highly correlated between organs18, we used the median time-specificity in our analyses unless otherwise noted. The median time-specificity only takes into consideration the organs where lncRNAs are expressed. Both time- and tissue-specificity indexes range from 0 (broad expression) to 1 (restricted expression).

Developmentally dynamic gene expression (i.e., significant temporal changes during organ development) was detected using masigPro20, an R package designed for the analysis of transcriptomics time-courses, as previously described18. Briefly, expression values in cpm were given as input to calculate a goodness-of-fit (R^2) metric for each organ. Genes with an $R^2 > 0.3$ in an organ were classified as developmentally dynamic in that organ. Consequently, developmentally dynamic genes in our dataset reach an $R^2 > 0.3$ in at least one organ. Due to the extensive transcription associated with the permissive chromatin environment of the sexually mature testis23, we excluded these samples from the calculation of the R^2 index for the testis. Differences between species in the number of identified developmentally dynamic genes can most likely be attributed to technical aspects, such as the number of assayed developmental stages and the similarity between biological replicates. The latter is influenced by the amount of genetic diversity across sampled individuals and the developmental interval spanned by replicates (e.g., hours for rodents, days-years for humans)18.

Orthology assignment and lncRNA age estimation

We used a Markov clustering algorithm to reconstruct homologous lncRNA families based on sequence similarity55 (Extended Data Fig. 1b). For each species, we merged all exonic regions for each lncRNA or ‘new putative-coding’ locus (newly identified transcribed regions that failed one of the coding potential filters, see above). ‘New putative-coding’ loci were included in this analysis because some lncRNAs have been shown to originate from protein-coding genes through pseudogenization56,57. We used blastn to search for similarity with exonic sequences of the same or different species, following soft-masking for repeats from RepeatMasker (4.0.6)58 (within and between species blastn (2.4.0)48). We filtered our alignments for identity ≥ 10% or a minimum length ≥ 50 nts and additionally required an E-value ≤ 10^{-3}. We then used reciprocal best hits between pairs of species and significant self-hits to cluster genes into homologous families with OrthoMCL (2.0)55, a method allowing for recent paralogs (duplicate genes arising after speciation) to be incorporated into families55. We allowed up to one member of each lncRNA family to be classified as ‘new putative-coding’ and required at least one lncRNA member of each family to show evidence of detectable transcription (> 1 RPKM in at least one sample). Finally, we removed all 1,324 multimember lncRNA families (families with recent paralogs, Supplementary Table 14) because manual inspection revealed that many of the identified paralogous relations were driven by repeats, low complexity regions or the split of a single lncRNA into two genes during annotation. For example, the lncRNA *Xist* is detected in all eutherian mammals in our dataset (marsupials and birds have different dosage compensation systems) but appears as two separate lncRNAs, *Rab_XLOC_042762* and *Rab_XLOC_042763*, in rabbit. The two lncRNAs are directly adjacent to each other, transcribed from the same strand and show similar female-specific spatiotemporal expression. Both align to the human *XIST* but not to each other, thus clearly representing a case where our genome annotation pipeline artificially split one lncRNA into two loci. To avoid incorrectly estimating the evolutionary age of these ambiguous multimember families, we only used our 18,459 high-confidence 1:1 orthologous lncRNA families (Supplementary Table 8) to infer the minimum evolutionary age for each lncRNA with parsimony, based on the phylogenetic relationships of the species where each lncRNA was transcribed, as previously described6. To account for the asymmetric

distribution of species in our dataset, we classified the age of lncRNAs shared between chicken and no more than two other species as “ambiguous”.

Evaluation of the lncRNA orthology assignment

We used the identity of neighboring protein-coding genes to assess the specificity of our lncRNA family definitions, as orthologous loci are often found in conserved synteny across vertebrates8,11,14,59. For each human lncRNA we identified the closest upstream and downstream protein-coding gene using bedtools (2.25.0) closest with the options –id and –iu, respectively60. We calculated distances based on gene bodies and allowed for assignment to an overlapping transcript. We repeated the procedure for lncRNAs in three more species in our dataset representing various evolutionary distances (macaque - 25 million years ago, Mya, mouse - 90 Mya, opossum - 180 Mya). We then estimated the fraction of lncRNA orthologs in each species pair that had at least one conserved neighbor (in the same orientation)8. Protein-coding gene orthologs were retrieved from Ensembl v75. As a control, we used protein-coding genes to estimate the expected degree of synteny conservation across these evolutionary distances. As the presence of antisense lncRNAs, which are expected to overlap the same gene across species, can lead to an overestimation of the extent of synteny conservation, we repeated the analysis considering only intergenic lncRNAs.

To benchmark the sensitivity of our lncRNA family determinations, we compared them to a study that used sequence similarity between lncRNA exons to identify lncRNA orthology8. We extracted families that contained a human lncRNA and were termed “Mammalian-only” and “Amniote-only” and compared them to our 180 Mya and 300 Mya lncRNA families, respectively. As the number of mammalian and amniote species used for the lncRNA family reconstruction differs between the two studies, we calculated the fraction of available species that were found in each family and rounded to the first digit (e.g., 10%, 20%) to summarize the data into bins. We then compared the distribution of species fractions across matched lncRNA ages between the two studies (Extended Data Fig. 2d).

Comparison with Ensembl, genomic classification and integration with other datasets

We intersected the exons of our lncRNA annotations with all noncoding exons from the Ensembl annotation of the respective species in a strand specific manner using bedtools (2.25.0) intersect60. To estimate the number of newly identified transcripts we used a more recent Ensembl release, v9219. Since the genome assemblies for human, rhesus macaque, rat and chicken have been updated during the transition from v77 to v92, we used liftover chains61 to map the v92 Ensembl annotations to the old genome assemblies before intersecting. For all other analyses we used the Ensembl annotations matching our genome annotations, i.e., v75 for human and v77 for all other species.

For the genomic classification of our lncRNAs we used the sliding-window based classifier module of the tool FEELnc (1.0)62, classifying our lncRNA annotations against protein-coding genes from Ensembl (v75 for human and v77 for all other species). We used a maximum window extension of 100,000 bp and otherwise default settings. The results were filtered for the best hits according to the default criteria, which prioritize assignment to the closest genes and exonic over intronic interactions62. To simplify our analysis, we collapsed

Europe PMC Funders Author Manuscripts

the classification to the position (overlapping, upstream, downstream) and strand (sense or antisense, Extended Data Fig. 2a). LncRNAs transcribed in upstream antisense orientation and located up to 2 kb from their assigned coding gene were classified as divergent. Finally, lncRNAs located more than 100 Kb apart from their nearest coding gene were classified as "isolated intergenic".

We used the Reference Database for Functional LncRNAs (lncRNAdb, v2.0)27 to identify functionally validated human lncRNAs. To integrate with our annotation, we parsed the content of lncRNAdb for Ensembl IDs and then used the intersection to Ensembl as described above. For the CRISPRi screen library21, we used the primary transcription start site (TSS) provided by the authors to intersect with the first exon of our lncRNA annotations in a strand-specific manner. Since the precise TSS definition may differ between the two datasets, we extended the reported primary TSS (often provided at a single nucleotide resolution) by 500 bp in each direction (Extended Data Fig. 4k). To identify lncRNAs overlapping enhancers, we intersected our lncRNA exons with a set of human transcribed enhancers identified based on distinct bidirectional CAGE (Cap Analysis of Gene Expression) patterns from a total of 432 primary cell, 135 tissue and 241 cell line human samples25. For the positionally-conserved lncRNAs (pcRNAs)33, we downloaded transcript coordinates in bed12 format, lifted over from hg38 to hg19 and intersected the exonic regions with our annotation in a strand specific manner.

Controlling for maximum expression levels

Developmentally dynamic lncRNAs show significantly higher maximum expression compared to non-dynamic lncRNAs (Extended Data Fig. 3f; $P = 2.2 \times 10^{-16}$, two-sided Mann-Whitney U test). To control for the effect of maximum expression on the association of developmentally dynamic lncRNAs with conservation and functionally characterized transcripts, we generated sets of expression-matched human lncRNAs. First, we identified the non-dynamic lncRNAs that showed the closest maximum expression to each human dynamic lncRNA. Sampling without replacement failed to equalize the expression levels, so we sampled with replacement obtaining 3,098 non-dynamic lncRNAs. We then selected the dynamic lncRNAs that were closest in maximum expression to each of those non-dynamic lncRNAs (2,906 dynamic lncRNAs). Using this procedure we obtained similar numbers and almost identical distributions of maximum expression values for developmentally dynamic and non-dynamic lncRNAs (Extended Data Fig. 4c).

As this set of expression-matched lncRNAs was shifted towards expression levels more representative of the dynamic lncRNA population, we repeated the procedure with a second set of lncRNAs with maximum expression levels ranging from 0.25 to 0.75 RPKM to evaluate whether our observations also hold true for lowly expressed dynamic lncRNAs. 798 human dynamic lncRNAs fall within this range (as opposed to 7,100 non-dynamic lncRNAs). We then identified the 717 non-dynamic lncRNAs that showed the closest expression values to the dynamic lncRNAs (sampling with replacement), obtaining similar expression distributions (Extended Data Fig. 4g).

Spatiotemporal expression similarity of 1:1 orthologs

To estimate the expression similarity between human and mouse 1:1 orthologs we calculated the Spearman correlation for 1,663 lncRNA and 16,078 protein-coding gene pairs across the entire dataset, i.e., 67 organs/developmental stages between human and mouse (Supplementary Table 15). We then compared the distribution of Spearman correlation coefficients for lncRNA pairs that are non-dynamic, dynamic in only one species or dynamic in both species, and for protein-coding genes. As a control, we used the set of lncRNAs developmentally dynamic in both species, and calculated their expression correlation after shuffling their orthology relationships (sampling without replacement).

We used a set of 924 lncRNAs, identified as 1:1 orthologs between mouse and rat and developmentally dynamic in both species, to estimate the effect of evolutionary age constraint on lncRNA expression evolution. We divided our set of 1:1 orthologs based on the estimated age of the lncRNA family (families with 80 and 90 million years were combined). For each age group, we estimated expression similarity by calculating the Spearman correlation coefficient for the lncRNA pairs across 82 organs/developmental stages in mouse and rat (Supplementary Table 16)18.

Estimation of TF binding on promoters

Promoter regions were defined as regions 2,000 bp upstream to 1,000 bp downstream of a gene's TSS. For protein-coding genes, TSS coordinates were retrieved from Ensembl's BioMart19. For lncRNAs, the TSS was defined as the starting coordinate of the first exon of the longest isoform. We excluded antisense and divergently transcribed lncRNAs to avoid biases created by the overlap of lncRNA and protein-coding gene promoters. Randomly generated, non-repetitive, intergenic regions of matched length (3,000 bp) were generated as negative controls. We retrieved mouse TF binding sites from GTRD, a publicly available set of more than 5,000 uniformly processed ChIP-seq experiments for 432 mouse TFs29. The data have been summarized into meta-clusters corresponding to non-redundant binding positions of each TF to the mouse genome. We used bedtools (2.25.0) intersect60 to determine the overlap of TF binding sites with our regions of interest. Transcriptional regulation and complexity was calculated based on the number of distinct TFs bound to each region. As a complementary metric, we defined TF binding frequency for each TF as the fraction of promoters of each gene class that is bound by the respective TF.

The TF binding frequency was also used to determine tissue-specific transcriptional regulation. For each TF, we calculated the fraction of lncRNAs dynamic in each tissue with promoters bound by that TF. To identify the TFs with the highest binding variability, we normalized each binding frequency as a fraction of the maximum binding frequency of each TF and determined the standard deviation of the normalized frequencies. We removed TFs with a maximum frequency lower than 5% (less than 5% of the promoters of the lncRNAs dynamic in the organ with the highest frequency are bound by this TF), as these cases showed artificially high variability due to noise (Extended Data Fig. 5b). We then identified the 50 TFs with the highest normalized binding frequency variability across the organs. We used the normalized binding frequency to perform hierarchical clustering based on Euclidean distances in both dimensions (lncRNAs dynamic in each organ and TFs) using the

R package pheatmap (1.0.10)63. To examine the functional relevance of these TFs for the development of the organs where they show the maximum binding frequency on lncRNA promoters, we identified tissue-specific TFs as those with tissue-specificity greater than 0.6 and determined the tissue where they show maximum expression.

Classification and co-expression based on developmental trajectories

We identified the most common developmental trajectories in each organ using GPClust, a method to cluster time-series based on Gaussian process18,64–66. We combined lncRNAs and protein-coding genes dynamic in each organ and species and used the median variance-stabilized counts across replicates as input. We set the noise variance (k2.variance.fix) to 1.0 for mouse and 1.5 for human. We then classified clusters (and associated genes) as early, late or other based on their developmental trajectories (Fig. 3; Extended Data Fig. 8c). Representative functions were assigned to each cluster based on a gene ontology (GO) enrichment analysis for its coding genes with the R package WebGestaltR (0.1.1)67, using all dynamic coding genes in the respective organ as a background set.

Patterns of lncRNA developmental expression

We identified the protein-coding genes and lncRNAs that are differentially expressed between adjacent time-points in mouse using DESeq2 (with default settings)53. We required an adjusted *P*-value ≤ 0.05 and a $\log_2$ fold change ≥ 0.5. The sets of dynamic lncRNAs expressed in each organ and developmental stage were selected based on a median expression value across replicates of at least 1 RPKM. To estimate the degree of lncRNA conservation for each organ and developmental stage, we calculated the fraction of mouse lncRNAs with an inferred evolutionary age of at least 80 Mya (i.e., shared with at least one other species in our dataset besides rat). We estimated the degree of expression similarity between human and mouse, for each organ and developmental stage, by calculating the Spearman correlation coefficient of 1:1 orthologous lncRNAs dynamic in both species for matched developmental stages18. The differences in pleiotropy between different stages of organ development were estimated based on the tissue-specificity indexes for different classes of developmental trajectories, as described above. Similarly, we estimated the phenotypic impact of lncRNAs with different developmental trajectories based on the fraction of functionally validated lncRNAs (lncRNAdb)27 and growth phenotype-associated hits in the CRISPRi screen21. To test the enrichment for functionality of late-expressed developmentally dynamic lncRNAs compared to non-dynamic lncRNAs, we selected human dynamic lncRNAs that are classified as ‘late’ in all somatic organs in which they show dynamic expression profiles.

Co-expression with adjacent coding genes

Dynamic lncRNAs in human and mouse were assigned to their nearest protein-coding gene using bedtools (2.25.0) closest60 using the distance between gene bodies (similar results obtained using the distance between TSSs). Each protein-coding gene assigned to a lncRNA was then matched to its immediately neighboring protein-coding gene, which was used as a control. We estimated Pearson’s expression correlation between lncRNA-mRNA and mRNA-mRNA pairs using all samples in our dataset, except for sexually mature testis samples (P3 and later for mouse, young teenager and later for human). Median variance

stabilized counts across replicates were used as the input for these correlations. We observed that protein-coding genes annotated as paralogs in Ensembl showed significantly higher correlation coefficients compared to the other mRNA-mRNA pairs (Extended Data Fig. 9a; $P = 2.2 \times 10^{-16}$, two-sided Mann-Whitney U test). Paralogous genes most commonly arise through segmental DNA duplications, thus representing copies of the ancestral gene and sharing the same regulatory sequences68. Thus, although the functions and expression patterns of the two copies may diverge with time68, paralogous genes are on average expected to be more functionally related than protein-coding genes that only share a similar chromatin environment. Consequently, we removed triplets containing paralogous protein-coding genes from the comparison of correlation coefficients between lncRNA-mRNA and mRNA-mRNA pairs and from the identification of candidate co-expressed pairs. However, we still used paralogous genes to estimate the degree of correlation that implies functional relatedness, since the extent and significance of gene expression correlations vary depending on the size and nature of the dataset. Specifically, we compared the ratio of paralogous/non-paralogous protein-coding pairs identified as co-expressed using a range of Person's r correlation cutoffs (Extended Data Fig. 9b).

Based on this analysis, we identified candidate cis-coexpressed lncRNA-mRNA pairs as those with correlation coefficients greater than 0.75 and for which the correlation between the mRNA and the control was smaller than 0.75. To select only cases where the lncRNA-mRNA correlation was significantly higher than the mRNA-mRNA control, we additionally performed a Fisher Z-transformation and estimated the difference between the correlation coefficients for the lncRNA and the control using the function paired.r from the R package psych (1.8.4)69 to perform two-tailed tests for independent samples. We required our candidate lncRNA-mRNA pairs to have an adjusted $P \leq 0.05$. A gene ontology enrichment analysis was performed for the protein-coding genes of these pairs, using the R package WebGestaltR67.

To test the enrichment of co-expressed pairs shared between human and mouse with developmental functions, we used AmiGO (v2) to download all human protein-coding genes associated with the development of the organs in our dataset (brain development, GO:0007420; heart development, GO:0007507; kidney development, GO:0001822; liver development, GO:0001889; gonad development, GO:0008406) and performed a hypergeometric test to compare human protein-coding genes co-expressed with a lncRNA in both human and mouse to all human protein-coding genes co-expressed with a lncRNA.

We note that although we tried to control for the effect of a shared regulatory environment31,70 using mRNA-mRNA controls, lncRNAs are likely more susceptible to it due to their weaker regulatory complexity (Fig. 2f)28. Furthermore, as our data correspond to steady-states, positive correlations are used to identify functional relatedness between the lncRNA and its adjacent protein-coding gene but cannot be interpreted as mechanistic interactions. Even in cases when the lncRNA has a regulatory effect on the adjacent protein-coding gene, distinguishing between activating and repressive effects would require precise knowledge about the expression state of the target gene in the absence of the lncRNA, information that can only be obtained through perturbation approaches.

Europe PMC Funders Author Manuscripts

General statistics and plots

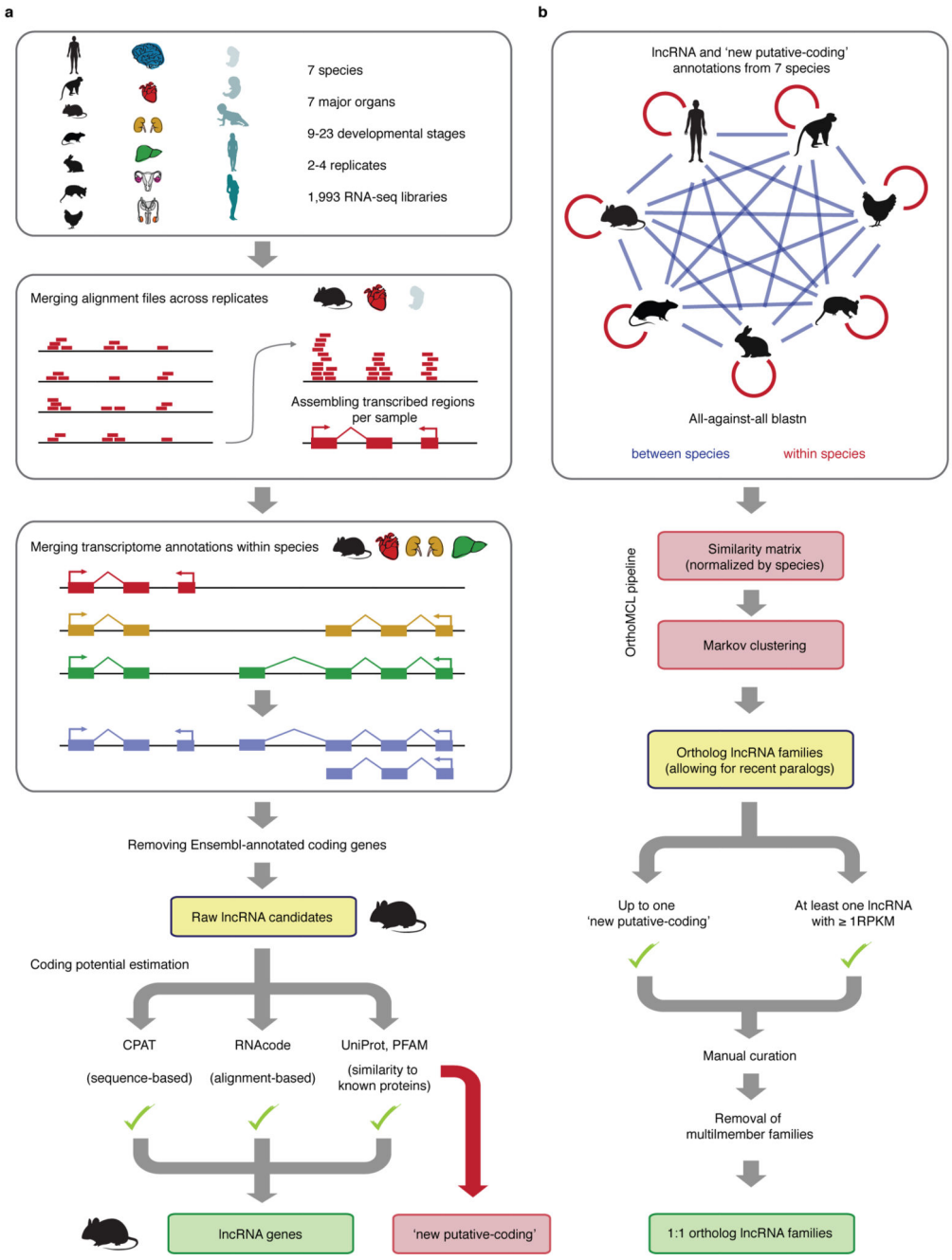
Unless otherwise stated, all statistical analyses were performed in R71, utilizing the packages dplyr (0.7.6)72, tidyr (0.8.1)73, stringr (1.3.1)74, data.table (1.11.4)75 and psych (1.8.4)69. All plots were generated in R71 using the packages ggplot2 (3.0.0)76, gridExtra (2.3)77, reshape2 (1.4.3)78, plyr (1.8.4)79, FactoMineR (1.41)80 and pheatmap (1.0.10)63. The R implementation of WebGestalt (0.1.1)67 was used for all GO enrichments.

Data availability

LncRNA annotations (Supplementary Data 1), expression tables (in RPKM; Supplementary Data 2) and homologous lncRNA families (Supplementary Table 8) are available as supplementary materials. We also provide a tabular summary of lncRNA genomic, evolutionary and expression features (Supplementary Tables 1-7). We created a public interactive tool that allows the visualization of lncRNA genomic coordinates and expression profiles (lncrnas.kaessmannlab.org).

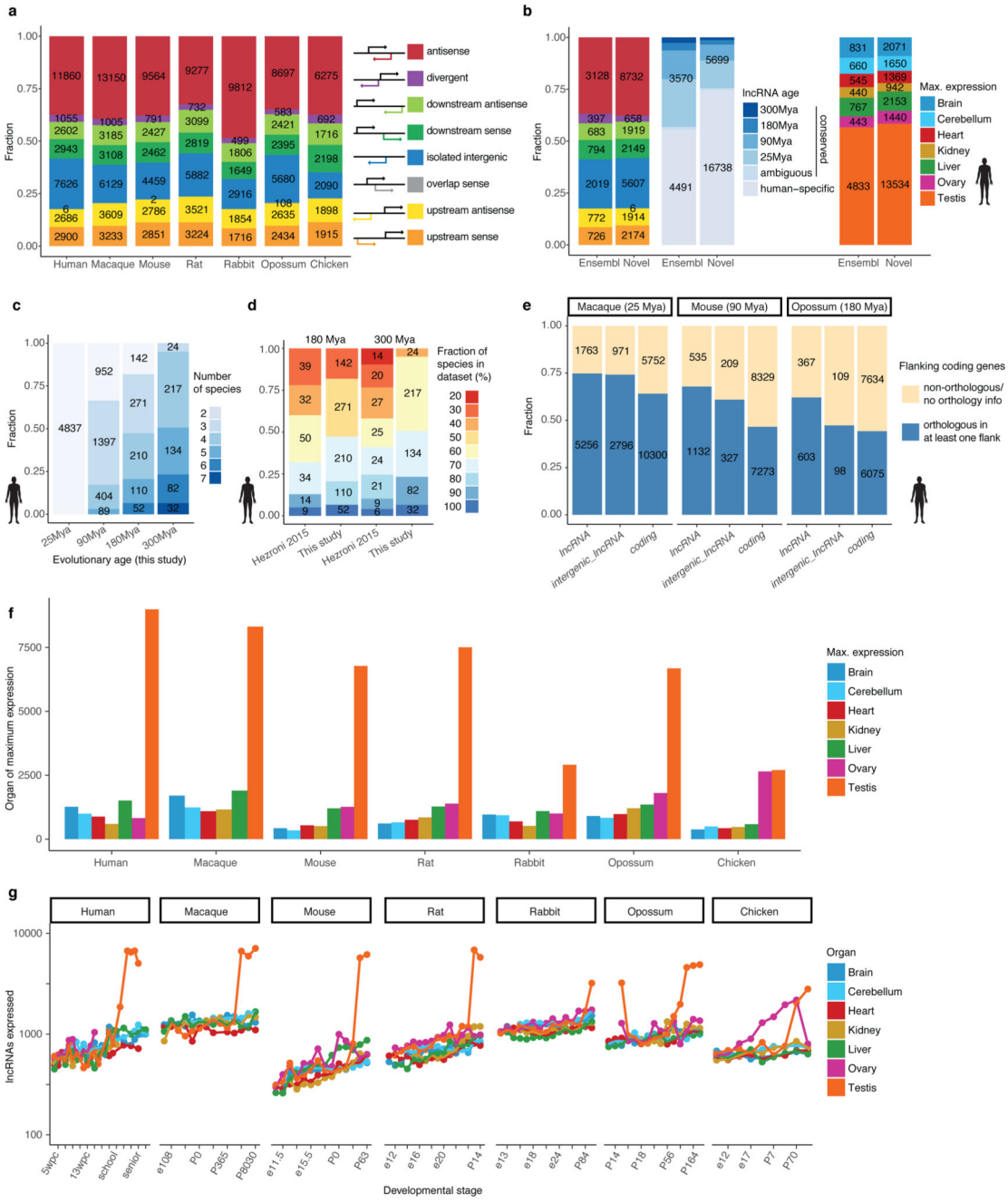
Extended Data

Europe PMC Funders Author Manuscripts



Extended Data Figure 1. Annotation and orthology assignment of lncRNAs.

a, Schematic representation of the lncRNA annotation pipeline. **b,** Schematic representation of the pipeline for the detection of 1:1 lncRNA families.



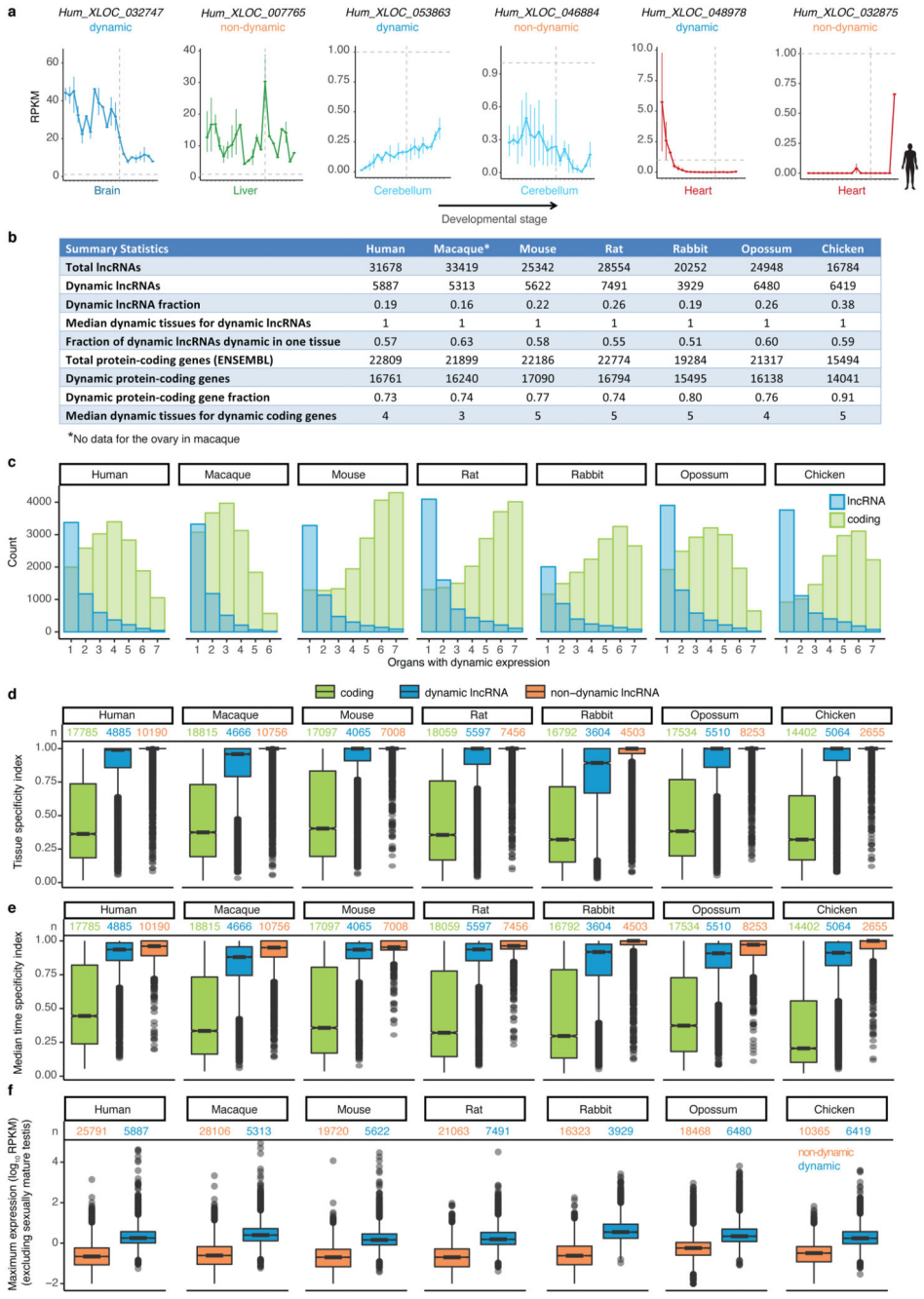
Extended Data Figure 2. Genomic classification and expression patterns of lncRNAs.

a, Distribution of lncRNAs among genomic classes in each species. **b,** Comparison of genomic classes (left), evolutionary age (middle) and organ of maximum expression (right) for known (Ensembl19) and newly annotated (novel) human lncRNAs. **c,** Number of species with a detected lncRNA member for human families of various evolutionary ages. **d,** Comparison of the fraction of species with a detected lncRNA member for human families conserved across mammals (180 Mya) and amniotes (300 Mya) with a previous study8. **e,** Fraction of lncRNAs and protein-coding gene orthologs found in conserved synteny with at

Europe PMC Funders Author Manuscripts

least one protein-coding gene neighbor for increasing evolutionary distances. **f,** Organ of maximum expression for expressed lncRNAs (≥ 1 RPKM) in each species. **g,** Number of lncRNAs expressed (≥ 1 RPKM) in each species during the development of each organ (in logarithmic scale).

Europe PMC Funders Author Manuscripts



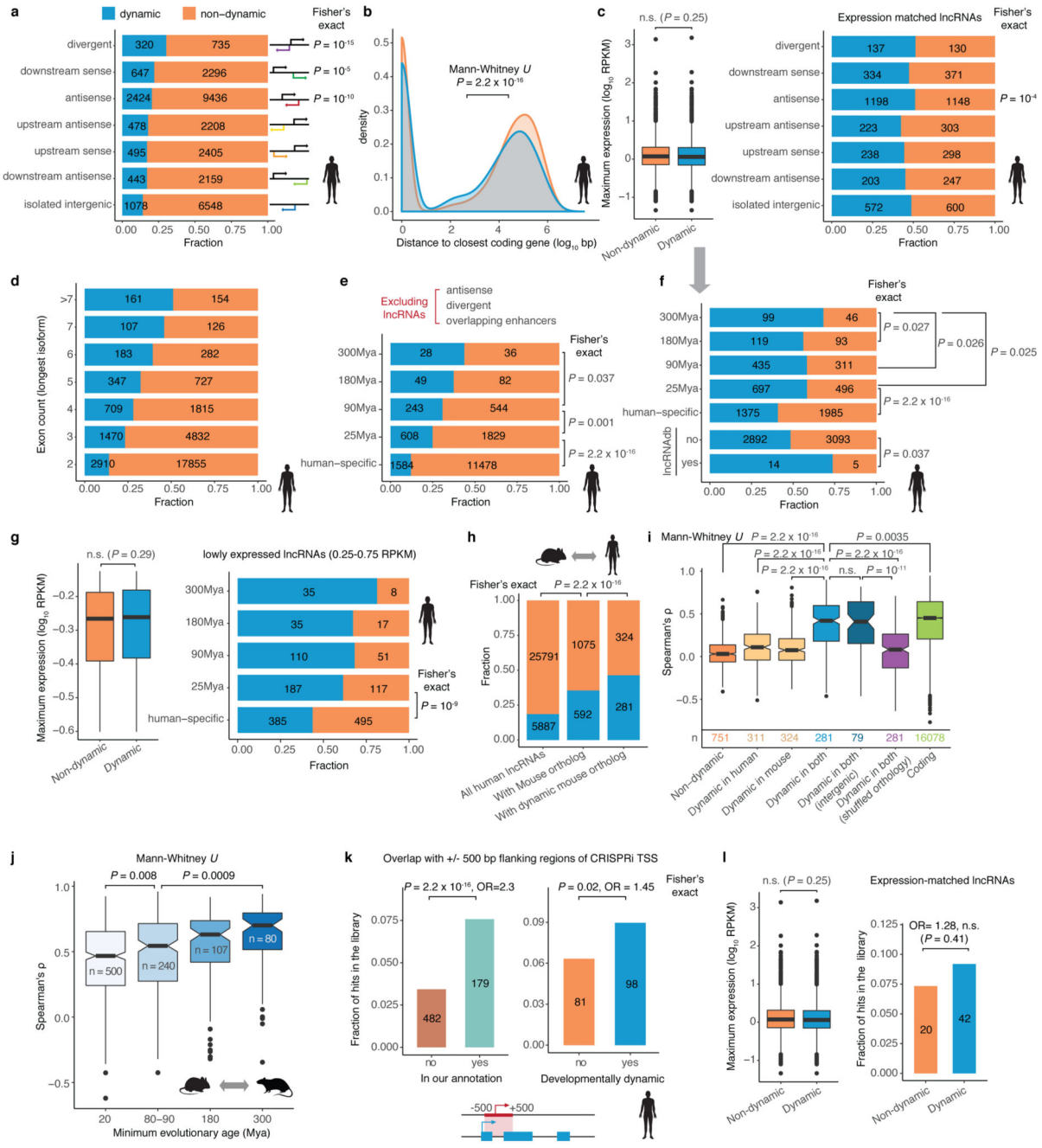
Extended Data Figure 3. Features of developmentally dynamic lncRNA expression.

a, Representative examples of human developmentally dynamic (n=5,887) and non-dynamic (n=25,791) lncRNAs' expression profiles (mean expression; vertical bars represent the minimum and maximum values across replicates) for varying levels of maximum expression, replicate reproducibility and expression windows. The vertical dashed line represents birth; the horizontal dashed line marks 1 RPKM. **b,** Summary statistics for the lncRNAs and protein-coding genes in this study. **c,** Number of organs with developmentally dynamic expression for dynamic lncRNAs and protein-coding genes in each species. **d, e,**

Europe PMC Funders Author Manuscripts

Tissue and median time-specificity of non-dynamic and dynamic lncRNAs, and protein-coding genes, across species. Tissue and time-specificity indexes range from 0 (broad expression) to 1 (specific expression). All comparisons between non-dynamic and dynamic lncRNAs, and protein-coding genes are significant ($P = 2.2 \times 10^{-16}$, two-sided Mann-Whitney *U* test). **f,** Maximum expression levels ($\log_{10}$ RPKM) for developmentally dynamic and non-dynamic lncRNAs across species (excluding samples from the sexually mature testis). Developmentally dynamic lncRNAs are more highly expressed in all species ($P = 2.2 \times 10^{-16}$, two-sided Mann-Whitney *U* test). In **d-f,** box plots represent median ± 25th and 75th percentiles, whiskers at 1.5 times the interquartile range.

Europe PMC Funders Author Manuscripts



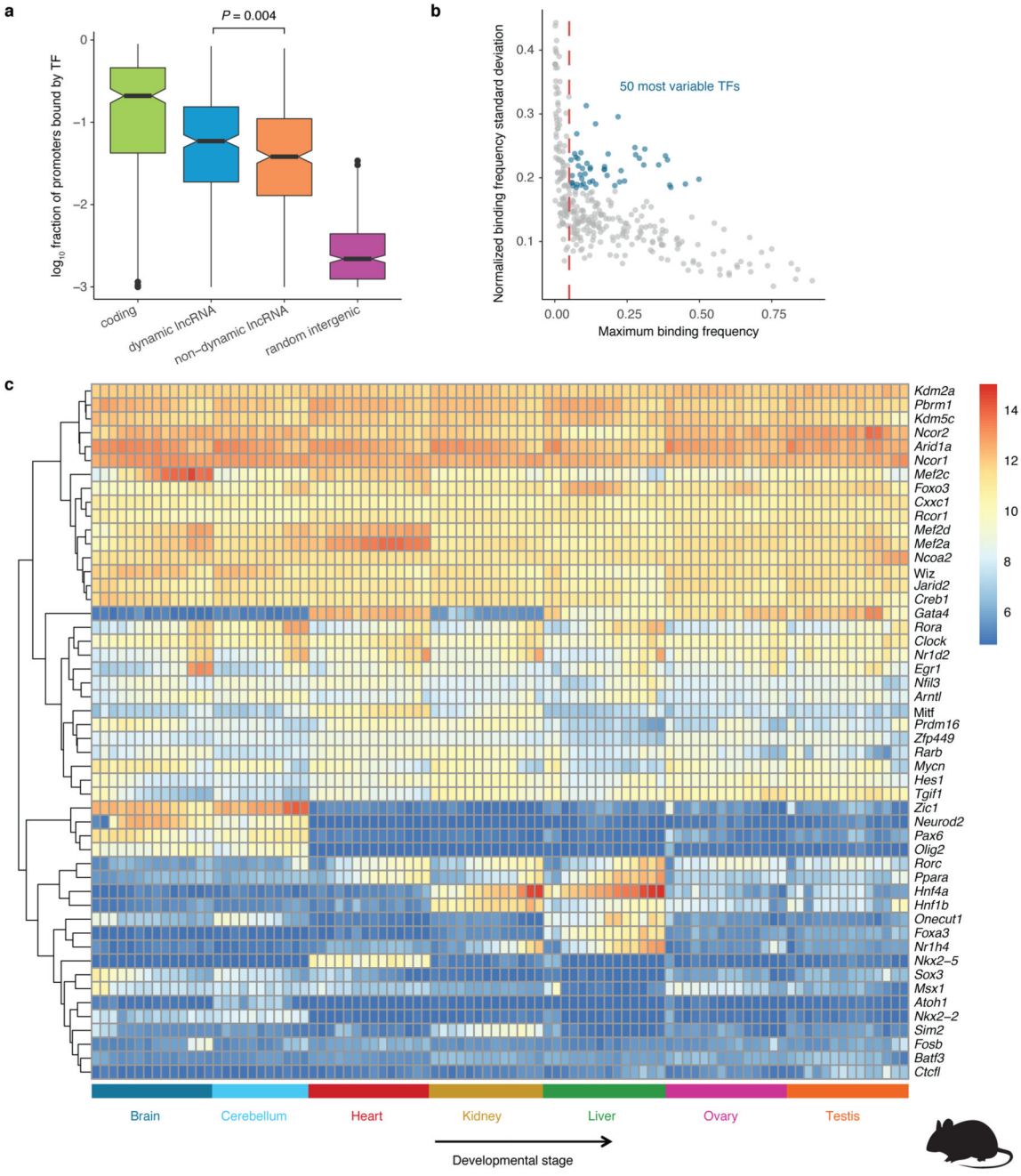
Extended Data Figure 4. Functionality signature enrichments of developmentally dynamic lncRNAs.

a, Fraction of developmentally dynamic human lncRNAs (n = 5,887) for different genomic classes. Overrepresented classes were determined by comparing the fraction of dynamic lncRNAs in each class against all other classes. **b,** Normalized density distribution of the distance to the nearest protein-coding gene for dynamic (n = 5,887) and non-dynamic (n = 25,791) human lncRNAs. **c,** Generation of expression-matched dynamic (n = 2,906) and non-dynamic lncRNAs (n = 3,098) and their distribution among genomic classes. **d,** Fraction

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

of developmentally dynamic human lncRNAs among isoforms with an increasing number of exons. The number of exons is significantly higher for developmentally dynamic lncRNAs ($P = 2.2 \times 10^{-16}$, two-sided Mann-Whitney *U* test). **e,** Fraction of human lncRNAs that are intergenic, developmentally dynamic and that do not overlap enhancers25 (n = 16,481) among different age groups. **f,** Fraction of developmentally dynamic genes across expression-matched (n = 6,004) human lncRNAs of different age groups (top) and functionally characterized lncRNAs27 (bottom). **g,** Generation of expression-matched, lowly expressed (0.25-0.75 RPKM) dynamic (n = 798) and non-dynamic (n = 717) human lncRNAs and their distribution across different age groups. **h,** Fraction of developmentally dynamic human lncRNAs (n = 5,887) with or without a mouse (dynamic or not) ortholog ($P = 2.2 \times 10^{-16}$, hypergeometric test). **i,** Similarity of spatiotemporal expression (Spearman's correlation coefficient between human and mouse organs/developmental stages) for 1:1 orthologs. **j,** Expression similarity across matched organs and developmental stages for mouse and rat 1:1 orthologous lncRNAs that are dynamic in both species, for different evolutionary ages. **k,** Fraction of lncRNAs present in the CRISPRi screen library21 resulting in a significant growth phenotype (hits) in at least one cell line for lncRNAs present (n = 2,364) or absent (n = 14,037) in our annotation and dynamic (n = 1,093) or non-dynamic (n = 1,277). **l,** Fraction of lncRNAs present in the CRISPRi screen library21 resulting in a significant growth phenotype (hits) in expression-matched dynamic (n = 2,906) and non-dynamic lncRNAs (n = 3,098). In **c, g, h-j** and **l**, box plots represent median ± 25th and 75th percentiles, whiskers at 1.5 times the interquartile range. In **a-l,** statistical tests are two-sided.



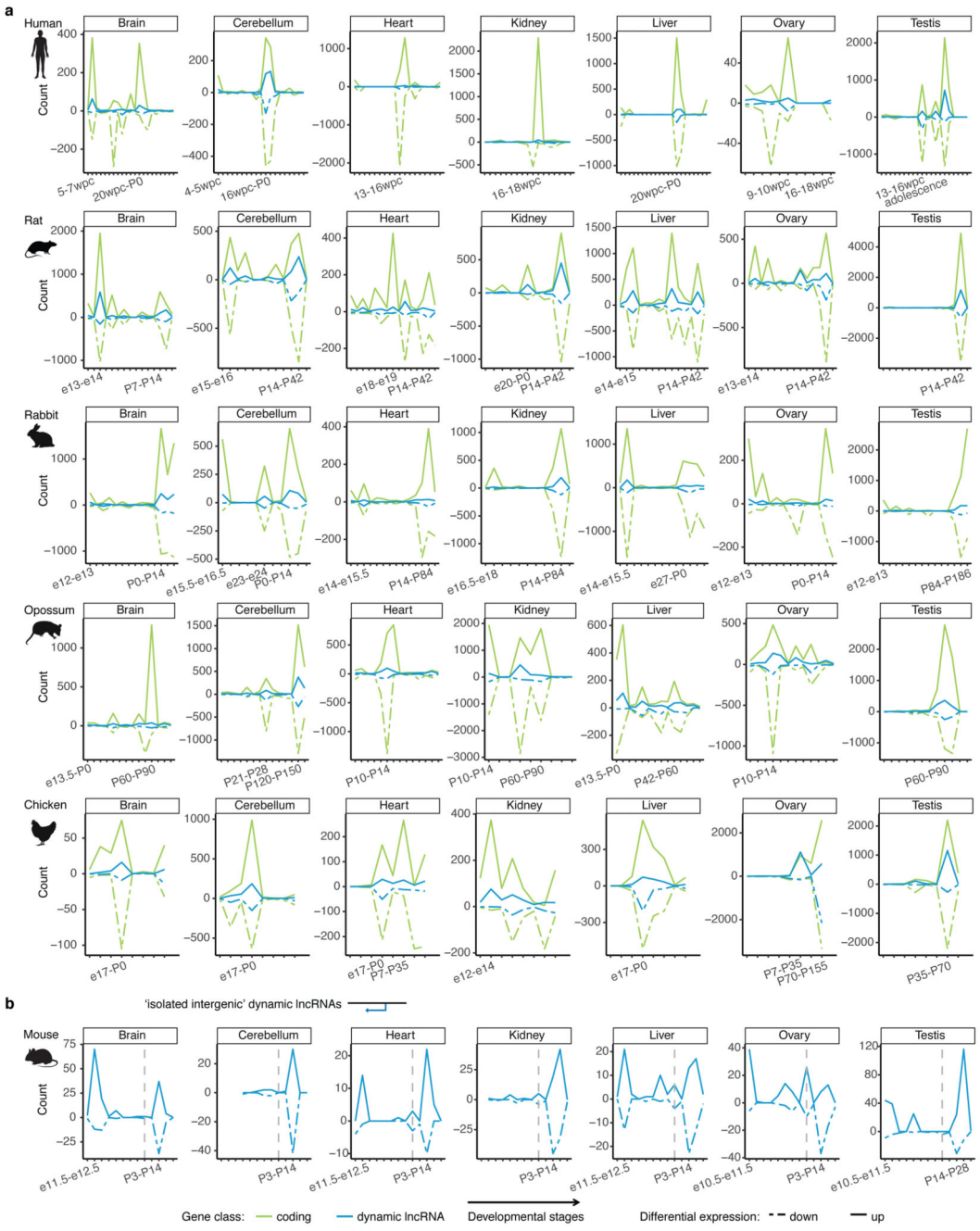
Extended Data Figure 5. Transcriptional regulation of dynamic lncRNAs in mouse.

a, Fraction of promoters of protein-coding genes, dynamic and non-dynamic lncRNAs, and size-matched random intergenic regions that overlap with binding sites for TFs. Each data point corresponds to a TF (n = 355). Box plots represent median ± 25th and 75th percentiles, whiskers at 1.5 times the interquartile range. **b,** Selection of the 50 TFs with the highest binding variability across promoters of lncRNAs dynamic in different organs (in blue). TFs with maximum binding frequency ≤ 0.05 (red line) were not considered, as their high variability is likely associated with a low binding frequency. **c,** Spatiotemporal expression

Europe PMC Funders Author Manuscripts

patterns of the 50 most variable TFs in mouse. The heatmap is clustered by rows and shows expression levels in counts (after variance-stabilizing transformation).

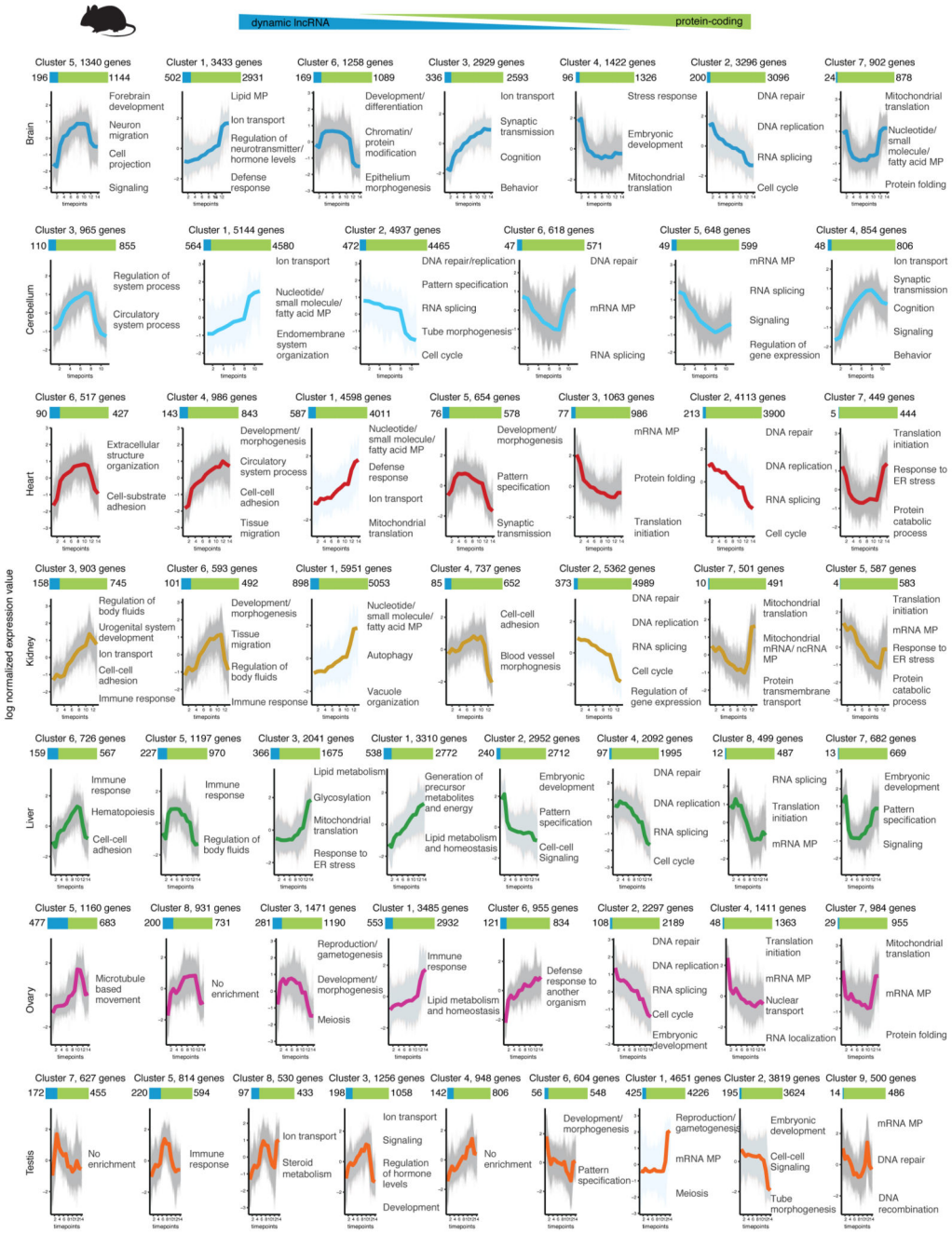
Europe PMC Funders Author Manuscripts



Extended Data Figure 6. Patterns of lncRNA expression in mammalian development.

a, Number of differentially expressed protein-coding genes and dynamic lncRNAs between adjacent stages of organ development in human, rat, rabbit, opossum and chicken. **b,** Number of differentially expressed 'isolated intergenic' (> 100 kb from the closest protein-coding-gene) dynamic lncRNAs between adjacent stages during mouse development.

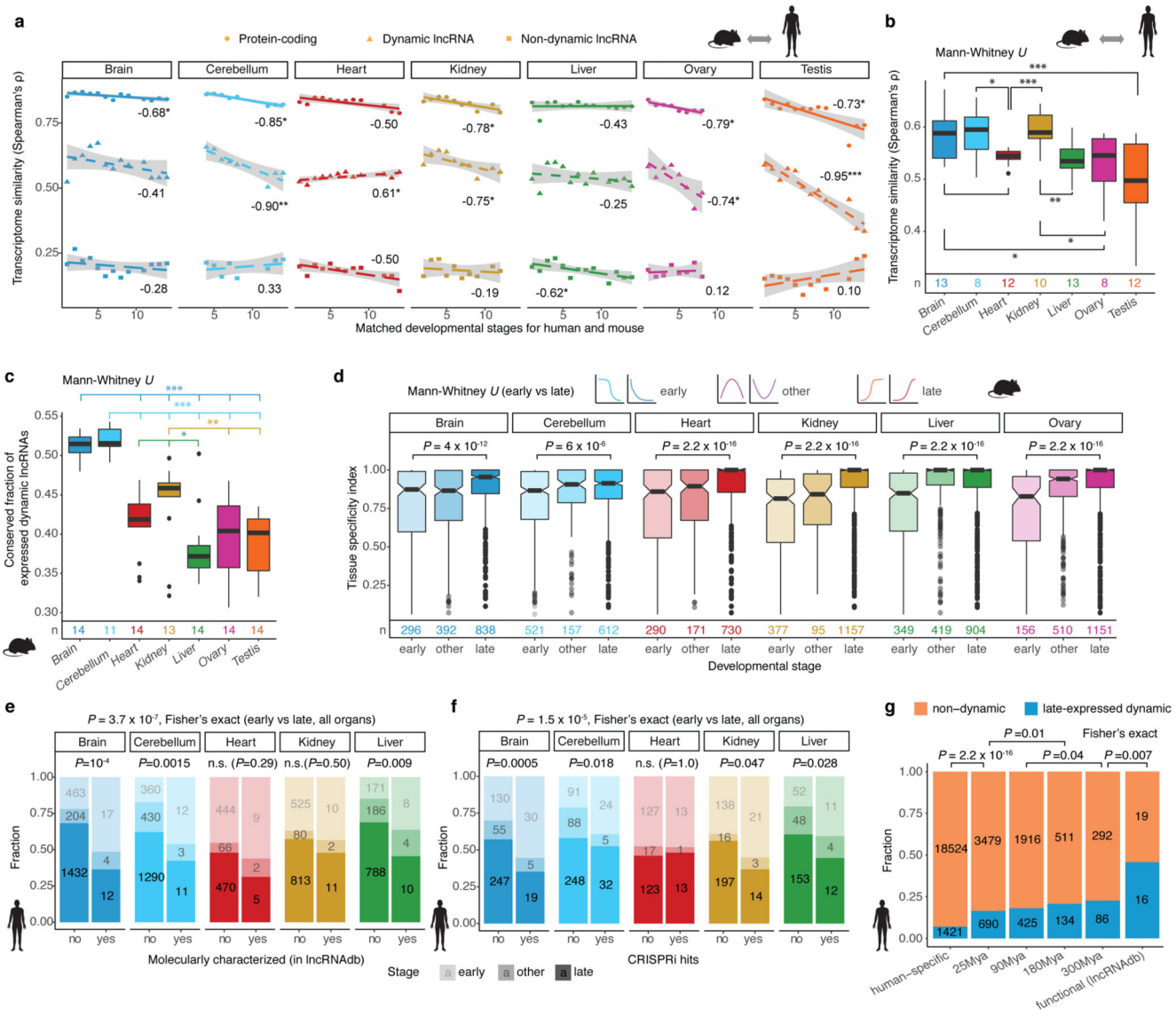
Europe PMC Funders Author Manuscripts



Extended Data Figure 7. Clustering of dynamic lncRNAs based on developmental trajectories.

Clusters of developmentally dynamic lncRNAs and protein-coding genes across mouse organs (brain = 14,629 genes; cerebellum = 13,166; heart = 12,382; kidney = 14,634; liver = 13,888; ovary = 12,694; testis = 13,749). Gray lines represent individual gene trajectories and solid lines posterior mean trajectories for each cluster. Clusters are arranged by decreasing fraction of lncRNAs. Enriched representative biological processes (Benjamini-Hochberg adjusted $P < 0.05$, hypergeometric test) are shown for each cluster.

Europe PMC Funders Author Manuscripts



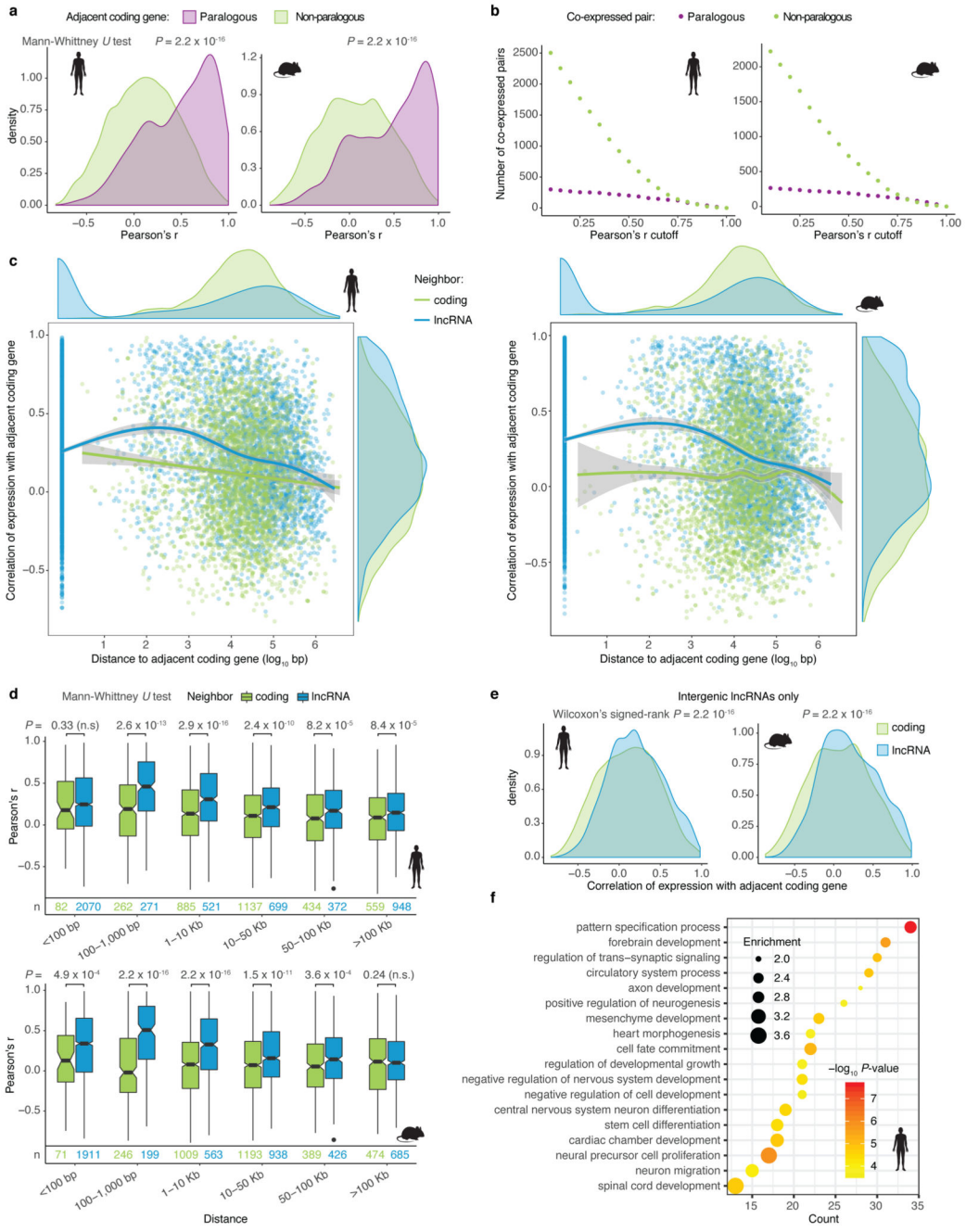
Extended Data Figure 8. Characteristics of dynamic lncRNAs expressed in different developmental stages.

a, Expression similarity between human and mouse 1:1 orthologous protein-coding genes (n = 16,078), developmentally dynamic (n = 281) and non-dynamic (n = 1,386) lncRNAs across organs/developmental stages. Each point corresponds to the Spearman's correlation coefficient of expression between human and mouse orthologs for matching samples. Lines and the 95% confidence interval (shaded regions) correspond to linear model predictions. Spearman's correlation coefficients between expression similarity and developmental stage are given for each comparison (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). **b,** Expression similarity between dynamic human and mouse orthologous lncRNAs from **a,** summarized by organ (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Mann-Whitney U test). **c,** Fraction of conserved (≥ 80 Mya) dynamic lncRNAs expressed in each mouse organ during development (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Mann-Whitney U test; the

Europe PMC Funders Author Manuscripts

color signifies the focal organ for each comparison). **d,** Tissue-specificity for mouse lncRNAs with different developmental trajectories. **e,** Fraction of human lncRNAs with different developmental trajectories among functionally characterized lncRNAs27 (n = 59) and **f,** CRISPRi growth screen hits21 (n = 98). **g,** Fraction of late-expressed dynamic (n = 2,956) and non-dynamic lncRNAs (n = 25,791) for different age groups and functionally characterized27 human lncRNAs. In **b-d**, box plots represent median ± 25th and 75th percentiles, whiskers at 1.5 times the interquartile range. In **a-g**, the statistical tests are two-sided.

Europe PMC Funders Author Manuscripts



Extended Data Figure 9. Co-expression of dynamic lncRNAs with adjacent protein-coding genes. **a,** Normalized density distribution of Pearson's correlation coefficients (*r*) of spatiotemporal gene expression between adjacent paralogous (human = 267; mouse = 263) and non-paralogous (human = 3,359; mouse = 3,382) mRNA-mRNA pairs. **b,** Number of paralogous (human = 267; mouse = 263) and non-paralogous (human = 3,359; mouse = 3,382) adjacent mRNA-mRNA pairs detected as co-expressed above a range of Pearson's *r* cutoffs. **c,** Relationship between distance and Pearson's correlation of expression for lncRNA-mRNA (human = 4,881; mouse = 4,722) and mRNA-mRNA (human = 3,359; mouse = 3,382) pairs.

Europe PMC Funders Author Manuscripts

Lines were estimated through loess regression and the 95% confidence interval is shown in gray. **d,** Distribution of Pearson's *r* for lncRNA-mRNA and mRNA-mRNA pairs across different distance intervals. Box plots represent median ± 25th and 75th percentiles, whiskers at 1.5 times the interquartile range. **e,** Density distributions of Pearson's *r* between a protein-coding gene and its nearest dynamic lncRNA (human=2,440; mouse=2,549) and protein-coding gene (human=1,606; mouse=1,777) after excluding antisense and divergently transcribed lncRNAs. **f,** Enriched biological processes among human protein-coding genes with significantly higher expression correlations with their adjacent dynamic lncRNA than with the control protein-coding gene (n=358; Benjamini-Hochberg adjusted $P < 0.01$, hypergeometric test; data for mouse in Fig. 4b). In **a-f**, statistical tests are two-sided.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank S. Anders, M. Sepp, E. Leushkin and members of the Kaessmann group for discussions, M. Sanchez-Delgado and N. Trost for assistance in figure design and I. Moreira for help in the development of the interactive tool. We acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1134-1 FUGG. This research was supported by grants from the European Research Council (615253, OntoTransEvol) and Swiss National Science Foundation (146474) to H.K., by the Marie Curie FP7-PEOPLE-2012-IIF to M.C.M. (329902) and by a scholarship for MSc studies by the Alexander S. Onassis Public Benefit Foundation (F ZL 084-1/2015-2016) to I.S.

References

1. Cabili M, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25:1915–1927. [PubMed: 21890647]
2. Derrien T, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012; 22:1775–1789. [PubMed: 22955988]
3. Iyer MK, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015; 47:199–208. [PubMed: 25599403]
4. Hon CC, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature. 2017; 543:199–204. [PubMed: 28241135]
5. Carninci P, et al. The Transcriptional Landscape of the Mammalian Genome. Science (80-.). 2005; 309:1559–1563.
6. Necsulea A, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014; 505:635–640. [PubMed: 24463510]
7. Washietl S, Kellis M, Garber M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Res. 2014; 24:616–628. [PubMed: 24429298]
8. Hezroni H, et al. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. Cell Rep. 2015; 11:1110–1122. [PubMed: 25959816]
9. Kopp F, Mendell JT. Functional Classification and Experimental Dissection of Long Noncoding RNAs. Cell. 2018; 172:393–407. [PubMed: 29373828]
10. Ponting CP, Oliver PL, Reik W. Evolution and Functions of Long Noncoding RNAs. Cell. 2009; 136:629–641. [PubMed: 19239885]
11. Ulitsky I. Evolution to the rescue: Using comparative genomics to understand long non-coding RNAs. Nat Rev Genet. 2016; 17:601–614. [PubMed: 27573374]
12. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. Nat Rev Genet. 2014; 15:734–748. [PubMed: 25297727]

13. Guttman M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458:223–227. [PubMed: 19182780]
14. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell. 2011; 147:1537–1550. [PubMed: 22196729]
15. Sauvageau M, et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife. 2013; 2013:1–24.
16. Grote P, Herrmann BG. Long noncoding RNAs in organogenesis: Making the difference. Trends Genet. 2015; 31:329–335. [PubMed: 25743487]
17. Goff LA, et al. Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain. Proc Natl Acad Sci U S A. 2015; 112:6855–62. [PubMed: 26034286]
18. Cardoso-Moreira M, et al. Gene expression across mammalian organ development. Nature. 2019
19. Zerbino DR, et al. Ensembl 2018. Nucleic Acids Res. 2018; 46
20. Conesa A, Nueda MJ, Ferrer A, Talón M. maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics. 2006; 22:1096–1102. [PubMed: 16481333]
21. Liu SJ, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science (80-.). 2017; 355
22. Mukherjee N, et al. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. Nat Struct Mol Biol. 2017; 24:86–96. [PubMed: 27870833]
23. Soumillon M, et al. Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Rep. 2013; 3:2179–2190. [PubMed: 23791531]
24. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012; 482:339–346. [PubMed: 22337053]
25. Andersson R, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014; 507:455–461. [PubMed: 24670763]
26. Kutter C, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. 2012; 8
27. Quek XC, et al. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res. 2015; 43:D168–D173. [PubMed: 25332394]
28. Melé M, et al. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. Genome Res. 2017; 27:27–37. [PubMed: 27927715]
29. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F. GTRD: A database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 2017; 45:D61–D67. [PubMed: 27924024]
30. Olson EN. Gene regulatory networks in the evolution and development of the heart. Science. 2006; 313:1922–7. [PubMed: 17008524]
31. Ruf S, et al. Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. Nat Genet. 2011; 43:379–86. [PubMed: 21423180]
32. Engreitz JM, et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature. 2016; 539:452–455. [PubMed: 27783602]
33. Amaral PP, et al. Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. Genome Biol. 2018; 19:1–21. [PubMed: 29301551]
34. Luo S, et al. Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. Cell Stem Cell. 2016; 18:637–652. [PubMed: 26996597]
35. Bester AC, et al. An Integrated Genome-wide CRISPRa Approach to Functionalize lncRNAs in Drug Resistance. Cell. 2018; 173:649–652.e20. [PubMed: 29677511]
36. Jiang W, Liu Y, Liu R, Zhang K, Zhang Y. The lncRNA DEANR1 facilitates human endoderm differentiation by activating FOXA2 expression. Cell Rep. 2015; 11:137–48. [PubMed: 25843708]
37. Jian X, Felsenfeld G. Insulin promoter in human pancreatic β cells contacts diabetes susceptibility loci and regulates genes affecting insulin metabolism. Proc Natl Acad Sci U S A. 2018; 115:E4633–E4641. [PubMed: 29712868]

38. Spigoni G, Gedressi C, Mallamaci A. Regulation of Emx2 Expression by Antisense Transcripts in Murine Cortico-Cerebral Precursors. PLoS One. 2010; 5:e8658. [PubMed: 20066053]
39. Ramos AD, et al. Integration of Genome-wide Approaches Identifies lncRNAs of Adult Neural Stem Cells and Their Progeny In Vivo. Cell Stem Cell. 2013; 12:616–628. [PubMed: 23583100]
40. Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet. 2016; 17:207–223. [PubMed: 26948815]
41. Liu SJ, et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol. 2016; 17:1–17. [PubMed: 26753840]
42. Lagarde J, et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat Genet. 2017; 49:1731–1740. [PubMed: 29106417]
43. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]
44. Pertea M, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015; 33:290–295. [PubMed: 25690850]
45. Trapnell C, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7:562–578. [PubMed: 22383036]
46. Wang L, et al. CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013; 41:1–7. [PubMed: 23143271]
47. Washietl S, et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA. 2011; 17:578–94. [PubMed: 21357752]
48. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]
49. Bateman A, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017; 45:D158–D169. [PubMed: 27899622]
50. Finn RD, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44:D279–D285. [PubMed: 26673716]
51. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31:166–169. [PubMed: 25260700]
52. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40. [PubMed: 19910308]
53. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. [PubMed: 25516281]
54. Yanai I, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005; 21:650–659. [PubMed: 15388519]
55. Li L, Stoeckert CJJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 2003; 13:2178–2189. [PubMed: 12952885]
56. Duret L, Chureau C, Samain S, Weissenbach J, Avner P. The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene. Science (80-.). 2006; 312:1653–1655.
57. Hezroni H, et al. A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. Genome Biol. 2017; 18:1–15. [PubMed: 28077169]
58. Smit, A; Hubley, R; Green, P. RepeatMasker Open-4.0. URL http://www.repeatmasker.org
59. Chen J, et al. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol. 2016; 17:1–17. [PubMed: 26753840]
60. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]
61. Hinrichs AS, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006; 34:D590–D598. [PubMed: 16381938]
62. Wucher V, et al. FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. 2017; 45:1–12. [PubMed: 27899559]
63. Kolde R. pheatmap: Pretty Heatmaps. 2015
64. Hensman J, Rattray M, Lawrence ND. Fast Variational Inference in the Conjugate Exponential Family. Adv Neural Inf Process Syst. 2012:2888–2896.

65. Hensman J, Lawrence ND, Rattray M. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. BMC Bioinformatics. 2013; 14:252. [PubMed: 23962281]
66. Hensman J, Rattray M, Lawrence ND. Fast Nonparametric Clustering of Structured Time-Series. IEEE Trans Pattern Anal Mach Intell. 2015; 37:383–393. [PubMed: 26353249]
67. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 2017; 45:W130–W137. [PubMed: 28472511]
68. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010; 20:1313–26. [PubMed: 20651121]
69. Revelle W. psych: Procedures for Personality and Psychological Research. 2017
70. Ebisuya M, Yamamoto T, Nakajima M, Nishida E. Ripples from neighbouring transcription. Nat Cell Biol. 2008; 10:1106–1113. [PubMed: 19160492]
71. R: A language and environment for statistical computing. 2008
72. Wickham H, Romain F, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. 2017
73. Wickham H. tidyr: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions. 2018
74. Wickham H. stringr: Simple, Consistent Wrappers for Common String Operations. 2018
75. Dowle M, Srinivasan A. data.table: Extension of `data.frame`. 2017
76. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. 2016.
77. Auguie B. gridExtra: Miscellaneous Functions for ‘Grid’ Graphics. 2017
78. Wickham H. Reshaping Data with the reshape Package. J Stat Softw. 2007; 21:1–20.
79. Wickham H. The Split-Apply-Combine Strategy for Data Analysis. J Stat Softw. 2011; 40:1–29.
80. Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. J Stat Softw. 2008; 25:1–18.

Europe PMC Funders Author Manuscripts

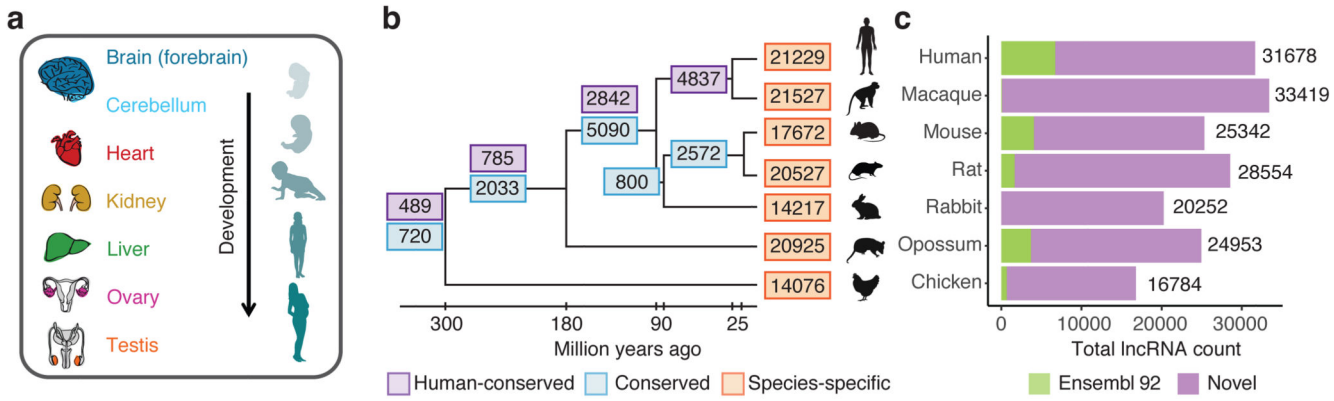


Figure 1. lncRNAs expressed during mammalian organ development.

a, Schematic representation of the dataset. **b,** Phylogenetic distribution of 1:1 orthologous lncRNA families (branches) and species-specific lncRNAs (leaves). **c,** Overlap with Ensembl v92 annotations.

Europe PMC Funders Author Manuscripts

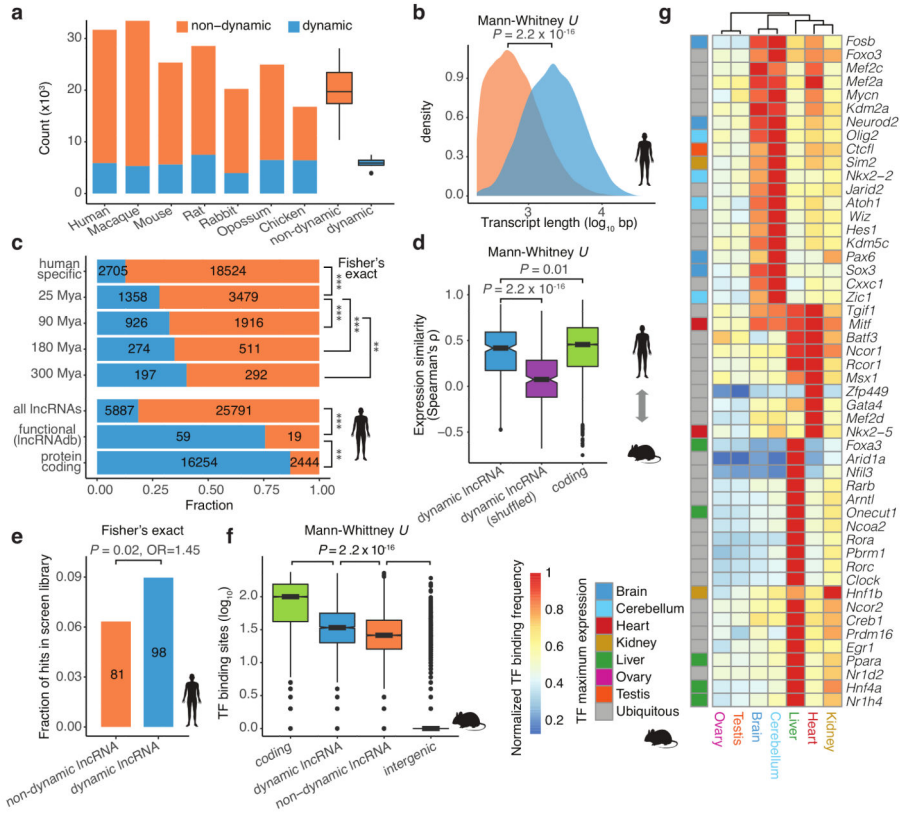


Figure 2. Developmentally dynamic lncRNAs are enriched for functional loci.

a, Number of non-dynamic and dynamic lncRNAs identified in each species. The box plots summarize the variability in the size of the repertoires across species (n = 7). **b,** Density distribution of transcript length for non-dynamic (n = 25,791) and dynamic human lncRNAs (n = 5,887). **c,** Fraction of dynamic loci for human lncRNAs of different evolutionary ages (top), functionally characterized lncRNAs27 and protein-coding genes (bottom; $^{**}P < 0.01$, $^{***}P < 0.001$). **d,** Similarity of spatiotemporal expression (Spearman's correlation coefficient between human and mouse organs/developmental stages) for 1:1 orthologs (dynamic lncRNAs = 281, protein-coding genes = 16,078). **e,** Fraction of a CRISPRi screen library21 resulting to a significant growth phenotype ("hit") for non-dynamic (n = 1,277) and dynamic human lncRNAs (n = 1,093). **f,** Number of TF binding sites29 overlapping the promoters of protein-coding genes (n = 20,202), dynamic (n = 3,169) and non-dynamic lncRNAs (n = 11,818), and size-matched random intergenic regions (n = 20,202). **g,** Normalized TF binding frequency (heatmap) of the 50 TFs with the highest binding variability across organs. Rows and columns are hierarchically clustered. The row annotation depicts the organ of maximum expression for organ-specific TFs. In **a, d** and **f,** box plots represent median ± 25th and 75th percentiles, whiskers at 1.5 times the interquartile range. In **a-f,** statistical tests are two-sided.

Europe PMC Funders Author Manuscripts

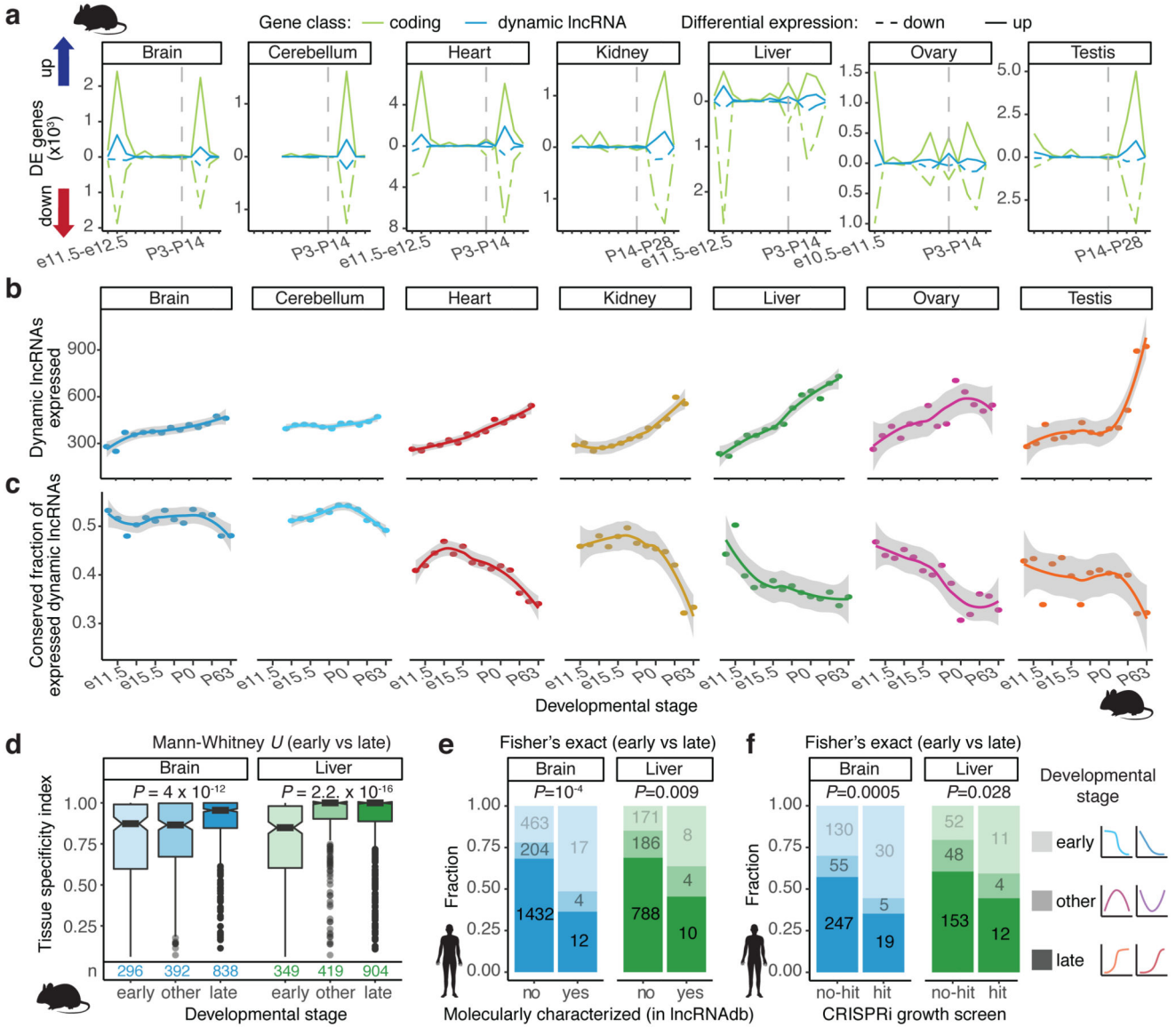


Figure 3. Patterns of dynamic lncRNA expression.

a, Number of differentially expressed **(DE)** protein-coding genes and dynamic lncRNAs between adjacent developmental stages (additional species in Extended Data Fig. 6a). **b,** Number of dynamic lncRNAs (n = 5,622) expressed and **c,** fraction of those conserved (evolutionary age ≥ 80 million years), during mouse organ development. Lines estimated through loess regression; 95% confidence interval shown in gray. **d,** Tissue-specificity of lncRNAs with different developmental trajectories. Box plots represent median ± 25th and 75th percentiles, whiskers at 1.5 times the interquartile range. **e,** Proportions of lncRNAs with different developmental trajectories among functionally characterized lncRNAs27 (n = 59) and **f,** CRISPRi growth screen hits21 (n = 98). Data for the remaining organs in Extended Data Fig. 8. In **c-e,** statistical tests are two-sided.

Europe PMC Funders Author Manuscripts

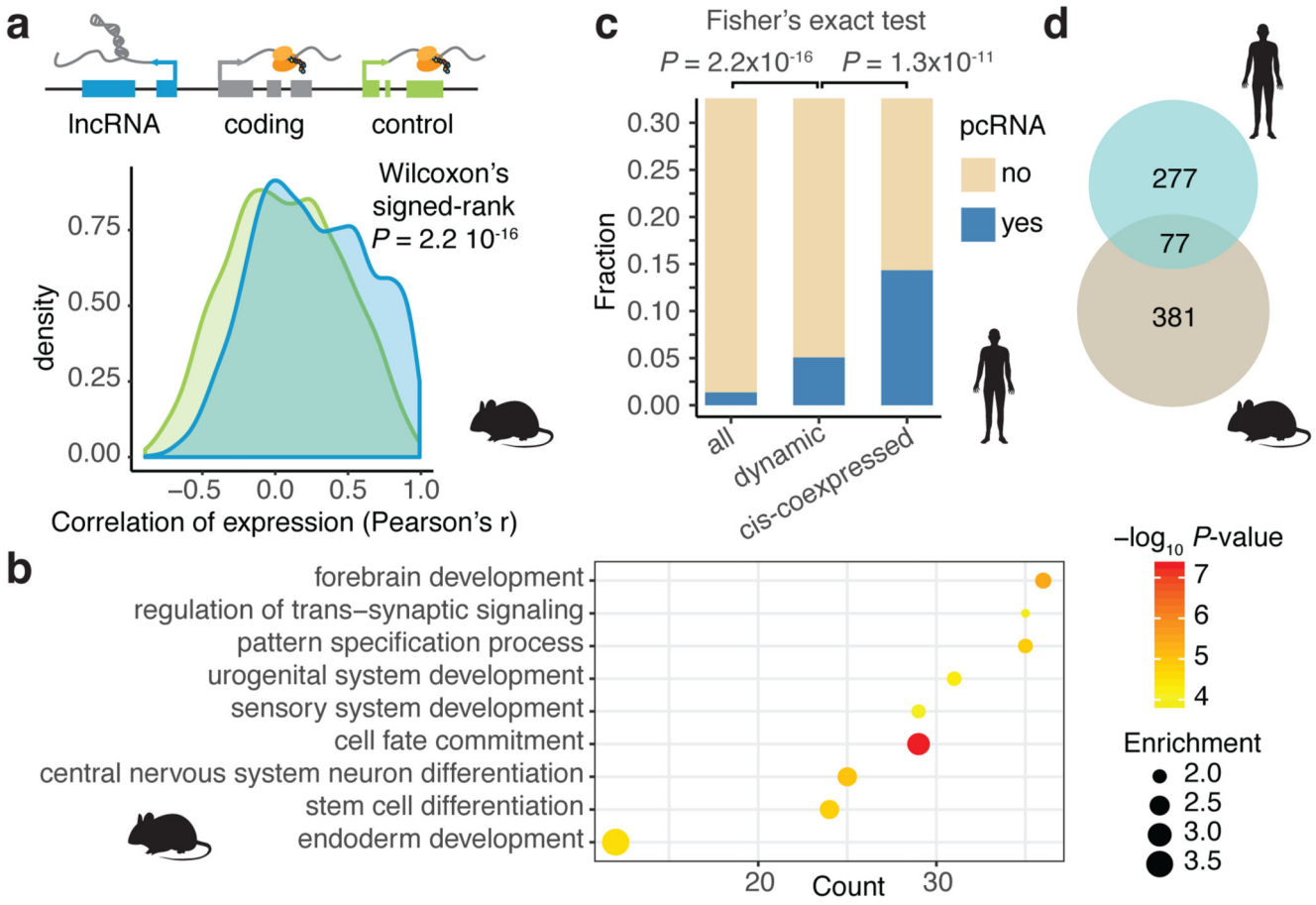


Figure 4. Co-expression with adjacent protein-coding genes.

a, Density distributions of the Pearson correlation coefficients between a protein-coding gene and its nearest dynamic lncRNA (n = 4,722) and protein-coding gene (control; n = 3,382). **b,** Enriched biological processes among protein-coding genes with significantly higher expression correlation with their adjacent dynamic lncRNA than with the control protein-coding gene (n = 449; Benjamini-Hochberg adjusted $P < 0.01$, hypergeometric test). **c,** Fraction of positionally-conserved lncRNAs (pcRNAs)33 among all lncRNAs (n = 31,678), developmentally dynamic lncRNAs (n = 5,887) and lncRNAs co-expressed with their adjacent protein-coding genes (n = 411). **d,** Overlap between human and mouse protein-coding genes that have a significantly higher expression correlation (Pearson's *r*) with their adjacent dynamic lncRNA than with the control protein-coding gene. In **a-c,** statistical tests are two-sided.

Europe PMC Funders Author Manuscripts