The evolution of lncRNA repertoires and expression patterns in tetrapods

Anamaria Necsulea, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie Baker, Frank Grützner, Henrik Kaessmann

▶ **To cite this version:**

Anamaria Necsulea, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, et al.. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature, 2014, 505 (7485), pp.635-640. ⟨10.1038/nature12943⟩. ⟨hal-05451438⟩

HAL Id: hal-05451438

https://cnrs.hal.science/hal-05451438v1

Submitted on 9 Jan 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright - All rights reserved

The evolution of lncRNA repertoires and expression patterns in tetrapods

Anamaria Necsulea[1,2,*], Magali Soumillon[1,2,&], Maria Warnefors[1,2], Angélica Liechti[1,2], Tasman Daish[3], Ulrich Zeller[4], Julie C. Baker[5], Frank Grützner[3] and Henrik Kaessmann[1,2]

1) Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland.
2) Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.
3) The Robinson Institute, School of Molecular and Biomedical Science, University of Adelaide, Adelaide, South Australia 5005, Australia.
4) Department of Systematic Zoology, Faculty of Agriculture and Horticulture, Humboldt University Berlin, 10099 Berlin, Germany.
5) Department of Genetics, Stanford University School of Medicine, Stanford University, Stanford, California 94305, USA.
* Present address: Laboratory of Developmental Genomics, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland.
& Present address: Harvard Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA; Broad Institute, Cambridge, MA 02142, USA.

Correspondence and requests for materials should be addressed to Anamaria Necsulea (anamaria.necsulea@epfl.ch) and Henrik Kaessmann (henrik.kaessmann@unil.ch).

Abstract

Only a minuscule fraction of long non-coding RNAs (lncRNAs) are well characterized. The evolutionary history of lncRNAs can provide insights into their functionality, but the absence of lncRNA annotations in non-model organisms has precluded comparative analyses. Here, we present the first large-scale evolutionary study of lncRNA repertoires and expression patterns, in eleven tetrapod species. We identify ~11,000 primate-specific lncRNAs and ~2,500 highly conserved lncRNAs, including ~400 genes that likely originated more than 300 million years ago. We find that lncRNAs, in particular ancient ones, are generally actively regulated and may predominantly function in embryonic development. Most lncRNAs evolve rapidly in terms of sequence and expression levels, but tissue specificities are often conserved. We compared expression patterns of homologous lncRNA and protein-coding families across tetrapods to reconstruct an evolutionarily conserved co-expression network. This network suggests potential functions for lncRNAs in fundamental processes like spermatogenesis and synaptic transmission, but also in more specific mechanisms such as placenta development through miRNA production.

Evolutionary analyses of protein-coding gene sequences[1] and expression patterns[2] have provided important insights into the genetic basis of lineage-specific phenotypes and into individual gene functions. For long non-coding RNAs (lncRNAs), such analyses remain scarce, despite growing interest in these genes. Recent studies have identified thousands of lncRNAs in human[3–5], mouse[6–9], fruitfly[10], nematode[11] and zebrafish[12]. Although most lncRNAs have unknown functions, some are involved in fundamental processes like X chromosome dosage compensation[13], genomic imprinting[14], cellular pluripotency and differentiation[15]. As a class, lncRNAs appear to be versatile expression regulators, which recruit chromatin-modifying complexes to specific locations[16], enhance transcription in *cis*[17] or provide decoy targets for miRNAs[18]. Thus, lncRNA evolutionary studies can also be informative in the wider scope of regulatory networks evolution.

Although several highly conserved lncRNAs are known[19], lncRNAs generally have modest sequence conservation[6,20,21]. Furthermore, in mouse liver, lncRNA transcription undergoes rapid evolutionary turnover[22]. These observations suggest that many lncRNAs may have no biological relevance. Detailed evolutionary analyses can clarify lncRNA functionality, but such analyses have been hampered by lack of annotations in non-model organisms.

The evolutionary history of lncRNAs in 11 tetrapod species

We used RNA sequencing (RNA-seq) to determine lncRNA repertoires of 11 tetrapod species. We analyzed 185 samples and ~6 billion RNA-seq reads (Supplementary Table 1), representing the poly-adenylated transcriptomes of 8 organs (cortex/whole brain, cerebellum, heart, kidney, liver, placenta, ovary and testis) and 11 species (human, chimpanzee, bonobo, gorilla, orangutan, macaque, mouse, opossum, platypus, chicken and frog), which diverged ~370 million years (MY) ago[23]. We included 47 strand-specific samples (~2 billion reads), which allowed us to confirm gene orientation and to predict antisense transcripts (Methods).

Using this dataset, we recovered spliced transcripts for most known genes (Extended Data Table 1a, Supplementary Discussion). We evaluated the protein-coding potential of transcripts using genome-wide codon substitution frequency scores (CSF[24]) and the presence of sequence similarity with known proteins and protein domains (Methods), obtaining correct classifications for ~96% of protein-coding genes and ~97% of known non-coding RNAs, on average (Extended Data Table 1b). We thus identified between ~3,000 and ~15,000 multi-exonic lncRNAs in each species, including known lncRNAs for human[4,5] and mouse[6], as well as ~10,000 novel human and ~9,000 novel murine lncRNAs (Fig. 1a, Extended Data Table 2). Although part of the variability in lncRNA repertoire size may be biologically meaningful, much is likely explained by unequal sequencing depth and by variable genome sequence and assembly quality (Supplementary Discussion).

We reconstructed homologous families based on sequence similarity and we inferred a stringent minimum evolutionary age of lncRNAs, requiring transcription evidence as an additional criterion (Methods). We also estimated a "maximum" evolutionary age by explicitly accounting for between-species variations in RNA-seq coverage and annotation quality (Methods, Extended Data Table 3a). We thus identified 13,533 lncRNA families transcribed in at least 3 species. Most (81%) lncRNA families were primate-specific, but 2,508 (19%) families likely originated more than 90 MY ago and 425 (3%) more than 300 MY ago (Fig. 1a). Most homologous lncRNAs were found in conserved synteny, even for distantly related species (Extended Data Table 3b).

The large proportion of inferred young lncRNAs may be due to fast lncRNA evolution, which prevents detection of distant homologs. Furthermore, the phylogenetic distribution of the species in our dataset may contribute to the skewed distribution of estimated ages. To investigate these possibilities, we evaluated DNA sequence conservation across placental mammals[25] and variation within populations[26] for human lncRNAs (Fig. 1b,c, Methods). We found that young lncRNAs (inferred minimum age 25 MY or lower) have low levels of long-term exonic sequence conservation (median score ~0.02), significantly lower than random intergenic regions (median score ~0.05, Wilcoxon text, p-value < 10^{-10}). However, single nucleotide polymorphisms found in primate-specific (minimum evolutionary age 25 MY) lncRNA exons have significantly lower derived allele frequencies (mean 0.11) than those found in intergenic regions (mean 0.12, randomization test, p-value < 0.01), consistent with recent purifying selection[27]. The same conclusions were reached using maximum evolutionary age estimates (Extended Data Fig. 1a,b), and when controlling for GC-biased gene conversion[28] (Extended Data Fig. 1c) and for linkage to protein-coding genes (Extended Data Fig. 1d). The presence of selective constraint in recent evolution, but not on a broader timescale, is compatible with a recent origination or acquisition of novel functions for a fraction of primate-specific lncRNAs.

Overall, the two measures of selective constraint correlate with evolutionary age estimates (Fig. 1c,d). Remarkably, older lncRNAs (minimum age 90 MY or higher) have

higher levels of long-term exonic sequence conservation than UTRs, and the oldest age classes are comparable with coding exons (Fig. 1c, Wilcoxon test, p-value > 0.5). Furthermore, lncRNA promoters are as conserved as protein-coding gene promoters even for younger classes (Extended Data Fig. 1e,f), suggesting stronger selective constraints at the transcriptional level, as previously observed[8].

Active regulation and potential roles in development for ancient lncRNAs

We next asked if lncRNA expression patterns vary with evolutionary age. We found that lncRNAs are lowly transcribed, highly organ-specific and preferentially expressed in testes (Fig. 2a-c, Extended Data Fig. 2), consistent with previous observations[4,5]. However, the testes-specificity is stronger for young lncRNAs (55%) than for old lncRNAs (46%, Fig. 2a, Chi-square test, p-value $<10^{-10}$), in agreement with the hypothesis that the permissive testes chromatin favors new gene origination[29]. After testes, neural tissues generally express the largest numbers of lncRNAs (Fig. 2a, Extended Data Fig. 2), consistent with a previously reported enrichment of lncRNAs in mouse brain[9]. Surprisingly, for platypus, ovary appears to be the second most favorable tissue for lncRNA expression (Extended Data Fig. 2).

The low expression levels and the testes-specificity raise the question of whether lncRNAs are actively regulated, or if they result from non-specific transcription in open chromatin regions. To test these hypotheses, we analyzed the occurrence of transcription factor (TF) binding sites as an indicator of active regulation. Using a genome-wide set of evolutionarily conserved binding sites predicted *in silico*[30] and ChIP-seq TF binding data[31] (Methods), we found that lncRNA promoters were more frequently associated with TFs than random intergenic regions (Fig. 2d, Extended Data Fig. 3a,c). Moreover, binding site sequence conservation was stronger in lncRNA promoters than in random intergenic regions and even protein-coding gene promoters, in particular for ancient lncRNAs (Fig. 2e, Wilcoxon test p-value < 10^{-10}). Consistently, the evolutionary turnover of CEBPA and HNF4a binding[32] between human and mouse is significantly slower for lncRNAs than randomly expected (Extended Data

Fig. 3f,g, Fisher's exact test < 10^{-10}). Taken together, these results suggest that lncRNA transcription is overall actively regulated, in particular for ancient lncRNAs.

Using *in silico* binding site predictions, we also uncovered a remarkable difference between two TF classes: homeobox TFs, which function in embryonic development, bind preferentially in lncRNA promoters, while non-homeobox TFs bind more frequently in protein-coding promoters (Fig. 2d, Extended Data Fig. 3b). Strikingly, 31% of old lncRNA promoters have homeobox TF binding sites, more than twice the frequency observed for protein-coding genes (14%, Fisher's exact test, p-value < 10^{-10}). The ChIP-seq dataset consisted largely (95%) of non-homeobox TFs, 117 (98%) of which were associated significantly more often with protein-coding than with lncRNA promoters (Extended Data Fig. 3d). However, 2 factors bound more frequently in old lncRNA than in protein-coding promoters: Suz12, a member of the polycomb repressive complex 2 (PRC2) that functions in pluripotency and differentiation[33] (Fig. 2f) and Oct4/POU5F1, a homeobox TF that controls pluripotency[34] (Extended Data Fig. 3e). The association with homeobox TFs and PRC2 suggests that lncRNAs (especially ancient ones) may be important for embryonic development, pluripotency and differentiation[15].

Rapid evolution of lncRNA expression patterns

We next assessed the evolutionary conservation of lncRNA expression patterns. We first estimated the presence of shared transcription across species. To reduce the impact of weak lncRNA sequence conservation, we compared intergenic lncRNAs across closely related primate species (Fig. 3a) and we analyzed lncRNAs transcribed in antisense of protein-coding exons (Extended Data Fig. 4a). We found that lncRNA transcription evolves rapidly: only ~92% of human intergenic lncRNAs were also detected as expressed in chimpanzee or bonobo and only ~72% were expressed in macaque, while more than 98% of conservation was observed for protein-coding genes, for all primates (Fig. 3a). Likewise, the evolutionary turnover of antisense lncRNAs is rapid compared to protein-coding genes (Extended Data Fig. 4a). The discrepancy between lncRNAs and protein-coding genes remained considerable when controlling for low lncRNA

expression with a read resampling procedure (Fig. 3a, Extended Data Fig. 4a), indicating that rapid transcription evolution is a genuine feature of lncRNAs[22].

We also measured correlations of lncRNA expression levels between pairs of species (Fig. 3b). The difference between lncRNAs and protein-coding genes is striking (Fig. 3c): Spearman's correlation coefficient for lncRNA brain expression between human and chimpanzee (which diverged 6 MY ago) is ~0.55, lower than the correlation (0.66) observed for protein-coding genes between human and *Xenopus* (which diverged ~370 MY ago). However, low lncRNA expression levels explain much of this discrepancy, as differences between correlation coefficients for the two classes of genes were much lower after resampling controls (Fig. 3c). For both protein-coding genes and lncRNAs, the testes have the fastest rates of evolution (Extended Data Fig. 4b).

We also observed that lncRNA tissue specificity is well conserved among primates, but not beyond. Indeed, a hierarchical clustering of samples based on pairwise correlations for eutherian lncRNA families revealed preferential grouping among related organs for primates, though all mouse samples clustered together (Fig. 3c, Extended Data Fig. 4f,g). Moreover, 47% of human tissue-specific lncRNAs had conserved specificity in all primates, while only 28% had conservation across all eutherians (Fig. 3d, Extended Data Fig. 4c-e). These proportions are significantly lower than for protein-coding genes, for which 81% are conserved across all primates and 72% across all eutherians (Fisher's exact test, p-value < 10^{-10}), but higher than randomly expected (randomization test, p-value < 0.01). The extent of conservation varies among tissues (Fig. 3d, Extended Data Fig. 4c-e), but is always significantly higher than expected by chance (randomization test, p-value < 0.01). These observations are illustrated by a lncRNA identified within a cluster of *GABA* receptors on human chromosome 5, expressed in neural tissues for primates, but detected only in liver in mouse (Fig. 3e).

Evolutionarily conserved co-expression network of lncRNAs and protein-coding genes

Finally, we evaluated the co-expression of lncRNAs and protein-coding genes, which can indicate functional relatedness[35] or regulatory relationships[36]. Since co-expression may also arise spuriously, we used evolutionary conservation as a criterion for significance[35]. We analyzed a set of 16,076 protein-coding gene families and 1,770 lncRNA families expressed in at least 3 species (Methods). We evaluated expression correlations for all gene pairs and tested if the combination of correlation coefficients across species was significantly higher (for positive associations) or lower (for negative associations) than expected by chance[35] (Methods). The conserved co-expression relationships formed a network with 9,388 nodes (8,971 protein-coding and 417 lncRNAs) and 97,556 edges (Supplementary Table 2). The same criteria applied on randomized gene families identified only ~160 co-expression relationships, proving the reconstruction specificity (Supplementary Discussion).

The co-expression network can predict functional relatedness, as illustrated by the high frequency of connections within gene ontology (GO) categories: out of 115 GO categories with at least 100 members, 101 (88%) had within-category connections more often than randomly expected (Fig. 4a). To verify if the direction of network connections may also predict regulatory associations, we analyzed 710 connections annotated as expression activation/inhibition relationships in the String[37] database. We found that ~70% of positive connections are annotated as activation relationships, significantly more than negative connections (30%, Fisher's exact test, p-value 0.01, Extended Data Fig. 4a). Accordingly, we found an overwhelming majority of negative connections for the *REST* and *HBP1* transcriptional repressors (Fig. 4b). Positive co-expression also often arises for genes that participate in RNA/protein complexes, such as the sodium channel subunit *SCNN1B* (Fig. 4b). Most (72%) network connections are positive co-expression cases. However, these occur more frequently between protein-coding genes, while lncRNAs have more negative connections (Fig. 4b). Interestingly, the imprinted lncRNA *H19*, which functions as a miRNA precursor[38], has a majority of negative connections (Fig. 4b).

The network connectivity depends on expression levels, as more connections were detected for highly expressed genes (Extended Data Fig. 5b,c). Expectedly, lncRNAs generally had lower connectivity (median degree 2) than protein-coding genes (median degree 5, Wilcoxon test p-value < 10^{-10}, Extended Data Fig. 5d), and transcription factors (TFs) were less well connected (median degree 4) than non-TF protein-coding genes. However, when resampling genes with similar expression levels, lncRNAs had higher degrees (median 3) than protein-coding genes (median 2, randomization test p-value < 0.01), and TFs had higher connectivity than other protein-coding genes (median 3, randomization test p-value = 0.02, Extended Data Fig. 5d), consistent with their central roles in regulatory networks. The highly connected lncRNAs may represent interesting candidates for further studies of gene expression regulation. Notably, lncRNAs had connections in *cis* more often than protein-coding genes (Extended Data Fig. 5e). An excess of connections in *cis* was also found for protein-coding genes acting in body plan development, in particular for *HOX* genes (Extended Data Fig. 5e, Supplementary Table 3).

Finally, we used the co-expression network to infer potential functions for lncRNAs. Using the Markov clustering algorithm (MCL[39]), we identified 1,326 groups of highly inter-connected genes, including 21 clusters with at least 50 genes (Fig. 4c, Supplementary Table 4). The proportion of lncRNAs in these clusters varied between 0 and 26% (Fig. 4d). The clusters were enriched for organ-specific functions, such as spermatogenesis (testis), synaptic transmission (neural tissues), catabolic processes (liver), muscle functions (heart) (Methods, Fig. 4c, Supplementary Table 4). We also recovered specific processes, such as anterior/posterior pattern formation in a cluster that includes *HOX* genes (Fig. 4c). The clusters with highest lncRNA proportions were enriched in spermatogenesis functions (Fig. 4c), in agreement with the predominant lncRNA testes-specificity. GO enrichment analyses for individual nodes suggested potential lncRNA involvement in nervous system development, cell adhesion, transcription *etc.* (Supplementary Table 5).

Potential miRNA precursors in the *H19* co-expression network

The only MCL cluster without significant GO enrichments (Fig. 4d) contains a high proportion (17.5%) of lncRNAs, including *H19*. As *H19* is a precursor for *miR-675*, which targets *Igf1r* and thus stalls placenta growth during late gestation[38], we scanned the network for other potential miRNA precursors (Methods). Surprisingly, genes positively connected with *H19* had the highest average density of embedded miRNAs (Extended Data Fig. 6a). These include one exceptional case: a lncRNA that could potentially promote the transcription of between 2 and 7 miRNAs in different species (Fig. 5a, Supplementary Table 6, Supplementary Discussion). This lncRNA (that we name *H19X,* for *H19* X-linked co-expressed lncRNA) is transcribed in all studied species and thus likely originated at least 370 MY ago, in the tetrapod ancestor. Strikingly, its expression pattern appears to have dramatically shifted during evolution, from an ancestral testis-predominant pattern to preferential expression in the chorioallantoic placenta of eutherians (Fig. 5a).

The miRNAs associated with *H19X* comprise 2 conserved tetrapod families, 4 placental mammal-specific families and 1 rodent-specific miRNA (Supplementary Discussion). Interestingly, the 2 oldest families (*miR-503/miR-16c* and *miR-424/miR322/mir-15c*) appear to have undergone accelerated sequence evolution in the eutherian ancestor (Extended Data Fig. 6b). In human and mouse, these miRNAs are generally highly expressed in the placenta (Extended Data Fig. 6c,d).

Finally, *H19X* is a neighbor of *Rsx*, the lncRNA that drives imprinted X-inactivation in marsupials[40] (Fig. 5b), suggesting that *H19X* may be itself imprinted. These results suggest that *H19X* may function like *H19*, by promoting miRNA transcription, preferentially in the placenta and in an imprinted manner. Although validation is needed, this illustrates how the reconstruction of a conserved co-expression network, enabled by the broad evolutionary perspective of our study, can predict lncRNA functions and stimulate further investigations.

METHODS SUMMARY

We sequenced poly-adenylated transcriptomes of 11 species and 8 tissues with Illumina GAII and HiSeq2000 technologies. We detected multi-exonic transcripts based on transcribed island and splice junction coordinates, using TopHat[41] and Cufflinks[42]. Protein-coding potential was inferred using codon substitution frequency scores (CSF[24]) and sequence similarity with known proteins[43] and protein domains[44]. We included published lncRNA annotations for human and mouse[45] and projected annotations across species. We reconstructed homologous families based on DNA sequence similarity, with single-link clustering. We inferred lncRNA evolutionary ages based on the phylogenetic distribution of species with transcription evidence, or for which its absence was due to low coverage or incomplete annotation. We computed RPKM levels using non-overlapping exonic regions and unambiguously mapped reads, and we normalized them through median-scaling[2]. We computed tissue specificity indexes as previously described[46]. To control for unequal coverage, we simulated read distributions by resampling identical numbers of reads *per* species and tissue, keeping proportions among genes unchanged. We reconstructed an evolutionarily conserved co-expression network by computing expression correlations between gene pairs and identifying cross-species combinations that are significantly higher/lower than randomly expected[35]. Network analysis was done with MCL[39] and Cytoscape[47]. All statistics and graphics were done in R[48].

ACKNOWLEDGEMENTS

We thank L. Froidevaux and D. Cortéz for help with genome sequencing, J. Meunier for help with preliminary miRNA analyses, K. Harshman and the Lausanne Genomics Technology Facility for high-throughput sequencing support, I. Xenarios for computational support, S. Bergmann and Z. Kutalik for advice on co-expression analyses. Human embryonic and fetal material was provided by the Joint MRC/Wellcome Trust (grant # 099175/Z/12/Z) Human Developmental Biology Resource (www.hdbr.org). The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. This research was supported by grants from the European Research Council (Starting Independent Researcher Grant 242597, SexGenTransEvolution) and the Swiss National Science Foundation (grant 31003A_130287) to H.K. A.N. was supported by a FEBS long-term postdoctoral fellowship.

AUTHOR CONTRIBUTIONS

A.N. conceived and performed all biological analyses and wrote the manuscript, with input from all authors. A.N. and M.W. processed RNA-seq data. M.S. and A.L. generated RNA-seq data. T.D. and F.G. collected platypus samples. U.Z. collected opossum samples. J.C.B. provided mouse placenta samples and contributed to *H19X* analyses. The project was supervised and originally designed by H.K.

FIGURE LEGENDS

Figure 1. Evolutionary age and genomic characteristics of lncRNA families.

a) Simplified phylogenetic tree. Internal branches and root: numbers of 1-1 orthologous lncRNA families, for each minimum evolutionary age. Tree tips: lncRNAs numbers for each species.

b) Exonic sequence conservation (placental PhastCons score), for random intergenic regions, lncRNA evolutionary age classes, coding and untranslated exons of protein-coding genes.

c) Mean derived allele frequency of autosomal non-CpG SNPs segregating in African populations (1000 genomes[26]). Intergenic SNPs were randomly drawn in regions matching lncRNAs recombination rates (Methods). Error bars: 95% confidence intervals based on 100 bootstrap resampling replicates.

Figure 2. LncRNA expression patterns and evidence for developmental regulation of old lncRNAs.

a) Distribution of the organ in which maximum expression is observed, for human protein-coding genes, old lncRNAs (minimum age 90-370 MY, 2,556 lncRNAs) and young lncRNAs (minimum age 0-25 MY, 12,126 lncRNAs).

b) Tissue specificity index. Values close to 1 represent high tissue specificity.

c) Distribution of the maximum expression level (log2-transformed RPKM).

d) Frequency of *in silico*-predicted binding sites for homeobox and non-homeobox transcription factors, in human gene promoters (2kb upstream) and in random intergenic regions.

e) Mean sequence conservation (PhastCons score) for TF binding sites. Error bars: 95% confidence intervals based on 100 bootstrap replicates.

f) Frequency of Suz12 binding (Encode ChIP-seq).

d),f) Error bars: 95% binomial proportion confidence intervals. We analyzed 793 'old' lncRNAs, 3,418 'young' lncRNAs and 16,566 protein-coding genes for which the predicted transcription start site was within 100bp of a CAGE tag.

Figure 3. Evolution of lncRNA expression patterns in tetrapods.

a) Percentage of human lncRNAs (4,430 intergenic primate lncRNA families) transcribed in other primates, in a pool of 5 somatic tissues (Methods).

b) Pairwise Spearman correlations between human and other species, for cortex/whole brain.

c) Hierarchical clustering of pairwise Spearman correlations, for 1,716 lncRNA families with 1-1 orthologs in all eutherians. Samples are color-coded according to the organ (horizontal) and species (vertical).

d) Proportion of human organ-specific lncRNAs (771 lncRNAs with minimum evolutionary age >90 MY, tissue specificity index >0.9, RPKM>0.1) for which organ-specificity is shared across primates. Red lines: random expectation. Dashed line: average conserved specificity across organs.

e) A lncRNA with conserved neural tissue specificity across primates. Error bars: range observed in biological replicates.

a),b) Red: estimates obtained with all reads, black: estimates obtained after resampling identical numbers of mapped reads *per* species and tissue. Error bars: 95% confidence intervals obtained with 100 bootstrap resampling replicates.

Figure 4. Evolutionary conserved co-expression network of protein-coding genes and lncRNAs.

a) Percentage of genes with connections within the same GO category, in real and randomized co-expression networks, for 115 biological process categories. Red: significant difference between real and randomized data.

b) Percentage of positive connections, for the entire network and for 6 genes with extreme positive/negative ratios.

c) Cytoscape[47] representation of the 21 largest MCL clusters in the co-expression network.

d) GO enrichment for the 21 largest MCL clusters; only the most significant GO category is displayed. Parentheses: numbers of coding and lncRNA genes in each cluster.

Figure 5. *H19* co-expression network and miRNA precursors

a) Expression pattern for an X-linked *H19* co-expressed lncRNA (*H19X*), in 5 tetrapod species. The error bars represent the range observed in biological replicates.
b) Genomic neighborhood of *H19X* in human and opossum.

EXTENDED DATA LEGENDS

Extended Data Table 1. Validation of the *de novo* detection and classification methods

a) Proportion of Ensembl-annotated (release 62) multi-exonic protein-coding genes, lncRNAs and processed transcripts recovered with our *de novo* detection methods.

b) Proportion of Ensembl-annotated protein-coding genes, lncRNAs, processed transcripts and other noncoding RNA genes (tRNA, rRNA) that were correctly classified as coding or noncoding with our approach.

Extended Data Table 2. LncRNA repertoires in 11 tetrapod species

a) LncRNA repertoires determined using all RNA-seq samples available for each species, including both strand-specific and non-strand-specific data.

b) LncRNA repertoires determined with strand-specific data.

Extended Data Table 3. LncRNA evolutionary age estimates and synteny conservation

a) Comparison between the minimum evolutionary age of lncRNA families (requiring transcription evidence in all species), and the maximum potential evolutionary age (Methods). The numbers represent the percentage of cases in which a given 'minimum age' estimate (rows) is associated with a given 'potential age' estimate (columns).

b) Synteny conservation for pairs of neighboring genes that contain at least 1 lncRNA. The neighboring gene pairs in the reference species were genes with 1-1 orthologs in the target species, separated by 5-100kb in the reference genome. The numbers represent the percentage of neighboring gene pairs in the reference species (rows) for which the 1-1 orthologs in the target species (columns) were found on the same chromosome, separated by at most 100kb.

Extended Data Figure 1.

a) Exonic sequence conservation (mean placental PhastCons score), for random intergenic regions, lncRNA maximum evolutionary age classes, coding and untranslated exons of protein-coding genes.

b) Mean DAF of autosomal non-CpG SNPs segregating in African populations (1000 genomes project[26]). Intergenic SNPs were randomly drawn in regions matching lncRNA recombination rates (Methods).

c) Mean DAF for the 4 classes of mutation orientation (W->S, or AT->GC; S->W, or GC->AT; W->W, or AT->AT and S->S, or GC->GC) for autosomal non-CpG SNPs found in primate-specific (age 25 MY) lncRNA exonic regions (blue) or in intergenic regions with matching recombination rates (grey). The W->S and S->W mutation classes are known to be affected by GC-biased gene conversion.

d) Same as (c), for lncRNAs that are found close to (left panel, maximum distance 10kb) or far from (right panel, minimum distance 50kb) Ensembl-annotated coding or noncoding genes.

e) Mean placental PhastCons score for promoter regions (1kb upstream) of lncRNA minimum evolutionary age classes (beige) and protein-coding genes (blue).

f) Mean placental PhastCons score for promoter regions (1kb upstream) of lncRNA maximum evolutionary age classes (beige) and protein-coding genes (blue).

(b,c,d) Error bars: 95% confidence intervals based on 100 bootstrap resampling replicates.

Extended Data Figure 2.

a) Distribution of the organ in which maximum expression is observed, for mouse protein-coding genes, old lncRNAs (shared across at least 2 species) and young lncRNAs (species-specific).

b) Tissue specificity index, for the same classes of mouse genes. Values close to 1 represent high tissue specificity.

c) Distribution of the maximum expression level (log2-transformed RPKM).

d),e),f) Same as a),b),c) for the opossum.

g),h),i) Same as a),b),c) for the platypus.

j),k),l) Same as a),b),c) for the chicken.

Extended Data Figure 3.

a) Comparison between the frequencies of *in silico* predicted transcription factor (TF) binding sites in lncRNA promoters (2kb upstream) and in random intergenic regions.
b) Comparison between the frequencies of *in silico* predicted TF binding sites in lncRNA and protein-coding gene promoters (2kb upstream). Homeobox TFs are depicted in blue.
c) Comparison between the frequencies of experimentally determined (ChIP-seq ENCODE) transcription factor (TF) binding sites in lncRNA promoters (2kb upstream) and in random intergenic regions.
d) Comparison between the frequencies of experimentally determined (ChIP-seq ENCODE) predicted TF binding sites in lncRNA and protein-coding gene promoters (2kb upstream).
e) Frequency of binding (Encode ChIP-seq data) for Oct4/POU5F1.
f),g) Proportion of HNF4a and CEBPA binding events shared between human and mouse, for random intergenic regions, lncRNA (321 lncRNAs with binding events and liver expression, supported by CAGE data) and protein-coding gene promoters (5kb upstream).

Extended Data Figure 4.
a) Percentage of human lncRNAs (found in antisense of protein-coding genes) that have transcription evidence in other species, as a function of the divergence time. Transcription evidence was assessed in a pool of brain and testis strand-specific RNA-seq data, for 2,535 human antisense lncRNAs that had 1-1 orthologs in at least one other species and transcription evidence in human (Methods).
b) Spearman correlation of human and mouse expression levels, in different tissues. The boxplots represent the variation observed in 100 bootstrap replicates.
c) Proportion of human organ-specific protein-coding genes (tissue specificity index >0.9, RPKM>0.1) for which the organ-specificity is shared across primates. Red lines: random expectation of shared organ-specificity. Horizontal black line: average conserved specificity for all organs.
d) Proportion of human organ-specific lncRNAs (minimum evolutionary age >90 MY, tissue specificity index >0.9, RPKM>0.1) for which the organ-specificity is shared across

eutherians. Red lines: random expectation of shared organ-specificity. Horizontal black line: average conserved specificity for all organs.

e) Same as (c), conservation across eutherian species.

f) Principal component analysis of lncRNA expression levels for families of eutherian 1-1 orthologs.

g) Principal component analysis of protein-coding gene expression levels for families of eutherian 1-1 orthologs.

Extended Data Figure 5.

a) Proportion of activation/inhibition relationships annotated in the String database, for positive and negative co-expression network connections.

b) Gene expression levels (maximum over all available sample and species for each co-expression network node) for different network connectivity classes.

c) Gene expression levels (maximum over all available sample and species for each co-expression network node) for connected lncRNAs, transcription factors (TFs) and non-TF protein-coding genes.

d) Network connectivity (node degree) for lncRNAs (black), transcription factors (medium gray) and for non-transcription factors protein-coding genes (light gray). Top: raw data; bottom: after correcting for expression level differences.

e) Difference between observed and expected proportions of connections in *cis,* for lncRNAs (red), protein-coding genes (blue) and for genes found in *HOX* clusters (black). The expected proportions were computed through randomizations (Methods).

Extended Data Figure 6.

a) Distribution of the average embedded miRNA density (miRNA hairpins / kb, in the gene body or 10 kb downstream), for genes that are positively connected with each network node. Red arrow: average miRNA density for genes that are positively connected with *H19*.

b) Maximum likelihood reconstruction of the phylogeny of the ancient *H19X*-associated miRNA family (*miR-503 /miR322/ miR-424/ miR-15c/ miR-16c*). MiRNAs associated with

H19X are displayed in red (*miR-503/miR-16c* sub-family) and blue (*miR-424/miR-322/miR-15c* sub-family).

c) Expression patterns of *mmu-miR-322*, associated with *H19X* in the mouse. The expression level was computed as the number of uniquely mapping reads *per* miRNA, after resampling the same number of reads per tissue.

d) Same as (c), for *mmu-miR-351*.

ONLINE METHODS

RNA sequencing and initial analysis

Our main dataset consists of 185 RNA-seq (135 previously published[2] and 50 new) samples, amounting to ~6 billion raw reads (Supplementary Table 1). The libraries were prepared using standard Illumina protocols and sequenced with Illumina GAII or HiSeq2000, as single-end reads, after polyA selection. After ensuring data comparability (Supplementary Discussion), we included 50 samples that we generated with a strand-specific RNA-seq protocol, for 6 species (human, mouse, opossum, platypus, chicken, *Xenopus*). To gain statistical power for co-expression network reconstruction, we incorporated 44 Illumina and 4 AB Solid RNA-seq samples published by other groups (Supplementary Table 1, Supplementary Discussion). We aligned the reads and detected splice junctions *de novo* using TopHat[41] 1.4.0 and Bowtie[49] 0.12.5. The genome sequences were retrieved from Ensembl[45] 62. Given the genetic similarity between chimpanzee and bonobo and the unavailability of the bonobo genome sequence when we started our project, we used the chimpanzee genome as a reference for all bonobo analyses.

LncRNA detection

To detect genes *de novo* with RNA-seq, we developed an algorithm that predicts multi-exonic transcribed loci based on transcribed island and splice junction coordinates and we used Cufflinks[42] to assemble transcripts from genomic read alignments (Supplementary Discussion). We combined multi-exonic transcripts detected with the two methods and Ensembl 62 annotations (including GENCODE lncRNAs[5]) into non-redundant datasets for each species. For human, we included ~8,000 lncRNAs predicted with RNA-seq[21]. To assess the evolution of sense-antisense transcripts, we repeated the detection procedure using only strand-specific samples. After the initial detection procedure, which used mainly in-house generated samples, we added to our analyses several previously published RNA-seq samples, mainly from the human ENCODE[5] and Illumina Human Body Map[4] projects, as well as several strand-specific samples that we

generated at a later stage to increase coverage for the placenta, ovary and testis for several species (Supplementary Table 1). We did not repeat the entire detection procedure with these new samples, but we used the additional splice junction information to join fragmented lncRNA loci. We also discarded *de novo* detected loci which thus appeared to be unannotated UTRs, as they were joined with protein-coding genes. We determined the coding potential of genes based on the codon substitution frequency (CSF[24]) score and on the presence of sequence similarity with known proteins (SwissProt[43] database) or protein domains (Pfam-A[44] database). Since *de novo* gene predictions can be incomplete or fragmented, we chose to assess the coding potential genome-wide rather than only for predicted exonic regions. We used the CSF score to define potential coding regions on a genome-wide scale, by scanning multiple species alignments (available through the UCSC Genome Browser[50]). Genes were said to be potentially noncoding if they were sufficiently distant (>2kb away) from a CSF-predicted coding region. Several distance thresholds were tested (Supplementary Discussion). We evaluated two additional methods (reading frame conservation[51] and presence of open reading frames), but these performed less well and were not used in our final analyses (Supplementary Discussion). After estimating the coding potential independently for each species, we verified that the classifications of the members of homologous families agreed, thus further reducing the possibility of misclassifications.

Cross-species annotation projection

To reduce the inequalities in annotation depth among species, we projected the annotations across species and included the projected gene models in each species' dataset. To do this, we searched for sequence similarity (blastn[52]) between the cDNAs of a reference species and the repeat-masked genomes of the target species. We accepted projections without rearrangements or internal repeats and with inferred intron sizes below 100kb. To avoid redundancy, the projections were added recursively, and only if they did not overlap with already annotated genes (Supplementary Methods).

We reduced the occurrence of fragmented gene predictions (a single gene is annotated as multiple neighbor loci), using a homology-directed defragmentation procedure that

takes advantage of the availability of multiple species. We searched for sequence similarity (blastn[52]) between the cDNA sequences of each species and classified as potentially "fragmented" those neighboring loci that could be reliably aligned with different regions of a single locus in another species (Supplementary Methods). For our final lncRNA dataset, we excluded candidates that clustered with protein-coding sequences (thus reducing the possibility of misclassifying UTRs as lncRNAs) and we used "de-fragmented" lncRNA annotations as controls for our analyses.

LncRNA filtering

We applied several filters to ensure reliability of the lncRNA dataset. For species-specific lncRNAs we required: minimum exonic length 200bp, at least 75% or 500bp of non-overlapping exonic sequence, minimum 5kb distance between lncRNA exons and Ensembl-annotated protein-coding gene exons, support by at least 5 non-strand-specific and 5 strand-specific reads (including splice junction reads), Ensembl gene biotypes (when available) "lincRNA" or "processed_transcript", no clustering (fragmentation) with protein-coding genes. For families of lncRNAs with *n* species, we required noncoding classification with both CSF and sequence similarity in at least *n-1* species and with at least one of the two criteria in all species, minimum exonic length 200bp (50bp for projected genes) in all species, support by at least 2 reads in at least 2 species, minimum distance 5kb to protein-coding gene exons for all species. For families that included Ensembl-annotated lncRNAs, we required the above criteria to be satisfied in at least *n-1* out of *n* species. For genes that overlapped on the antisense strand with other genes, we required support with strand-specific reads. We note that the list of lncRNAs provided for each species includes projected genes for which transcription evidence could not be found in the corresponding species, if these genes belonged to homologous families where at least 2 species had transcription evidence.

Reconstruction of homologous lncRNA families and lncRNA evolutionary age

We reconstructed homologous lncRNA families based on DNA sequence similarity. We searched for similarity between the cDNA sequences of each species, using blastn[52]. As in Ensembl Compara[53], we extracted reciprocal best hits for each pair of species and

significant self-hits for each species and we clustered genes with single-linkage. As lncRNAs can overlap with protein-coding genes or with transposable elements, we repeated the procedure after masking these regions, with no significant change. For improved sensitivity, we searched for alignments of wider regions, including 5kb of flanking sequences, in whole genome alignments generated with blastz/multiz[54] (available through the UCSC Genome Browser). Potential homologs were called for alignments that mapped to a single target species gene. This homology inference was used as a control for our analyses. We inferred the minimum lncRNA evolutionary age with parsimony, based on the phylogenetic distribution of the species with transcription evidence in the homologous gene families. We note that this estimate represents a strict lower boundary, since transcription may be undetectable for lowly expressed genes, in particular for the species with lower overall read coverage.

In addition, we tested whether the absence of transcription in some species can be simply attributed to differences in RNA-seq read coverage, and we provide an additional estimate of the 'potential' evolutionary age of lncRNAs. We estimated the proportion of mapped reads assigned to a given lncRNA, separately for each species and tissue. For each lncRNA family and for each tissue, we then estimated the minimum such proportion (*p_min*), over all species in which the lncRNA was detected as transcribed. Given that for projected genes we often recover only a limited fraction of the original exonic length, the *p_min* probability was further adjusted to reflect the difference in exonic length between the species with no transcription evidence and the species in which *p_min* was observed (*p_min* was multiplied by the ratio of the two exonic lengths). We then assessed the probability of observing 0 reads out of the total N mapped reads, given a theoretical detection probability of *p_min* and assuming a binomial distribution, in the species for which transcription could not be detected in that tissue. If the tissue was not sampled for a given species (such as orangutan testis or non-human great ape placenta), the probability was set to 1. Finally, these probabilities were multiplied over all available tissues, to obtain a combined estimate of the likelihood that the absence of transcription in that species is simply due to differences in read coverage and/or annotated exonic length. We then re-estimated the

evolutionary age of the lncRNA family, taking into account the phylogenetic distribution of the species in which transcription was either detected, or for which the absence of transcription could be attributed to read coverage and/or exonic length issues. This additional age estimate is termed the 'maximum' evolutionary age.

Selective constraint on DNA sequences

We computed average PhastCons[25] scores for exons and promoter regions, using genome-wide nucleotide resolution scores from the UCSC Genome Browser[50]. We downloaded SNP data from the 1000 genomes project[26], we filtered the SNPs to exclude potential CpG sites and we computed the average derived allele frequency (DAF) for the African population. For DAF comparisons, we derived 95% confidence intervals from 100 bootstrap resampling replicates (parametric statistics cannot be applied due to non-normal distributions). We analyzed only autosomal SNPs, residing in regions of moderate recombination (<2 cM/Mb), as measured using the DECODE[55] sex-averaged recombination maps in 20kb windows centered on the SNP. As a neutral control, we resampled intergenic SNPs (>5kb away from coding or noncoding genes) found in regions of similar recombination rates as lncRNAs (Supplementary Discussion). For overlapping genes (e.g. sense/antisense transcripts), both measures of selective constraint were estimated using non-overlapping exonic regions.

Expression level estimation and normalization

We estimated RPKM (reads *per* kilobase of exon *per* million mapped reads) values from unambiguous read alignments obtained with TopHat[41]. To ensure an unbiased measurement, we considered only exonic regions that could be unambiguously assigned to a single gene. We also measured expression levels with Cufflinks 2.0.0, using all mapped reads, with the embedded multi-read and fragment bias correction methods (Supplementary Discussion). For projected genes, for which exon annotations are often incomplete, we included 1kb flanking sequences on each side in the expression computation, if this extended region did not overlap with other transcribed loci. We normalized expression levels among samples with a median scaling, using the 1000 least-varying genes as a reference, as described previously[2].

Transcription factor binding analysis

We used a genome-wide set of human transcription factor (TF) binding sites (~2.7 million sites, for 375 transcription factors), predicted *in silico*[30], as well as ChIP-seq peaks for 127 TFs (excluding those directly associated with PolII or PolIII) from the human Encode project[56]. We analyzed the occurrence of TF binding sites or peaks in promoter regions, exclusively for genes for which the predicted transcription start site was found within 100bp of a CAGE tag cluster (data from the FANTOM project[57]). Two promoter region sizes were tested (2kb and 5kb), reaching similar conclusions. We also used ChIP-seq data for HNF4a and CEBPA for human and mouse[32]. We aligned promoter regions for the two species and considered that TF binding was conserved if peaks were found in both species within 10kb of the aligned transcription start site. As a control, we analyzed TF binding and binding conservation for 20,000 randomly drawn intergenic regions.

Expression evolution analyses

For the qualitative assessment of transcription conservation, we analyzed 4,430 intergenic lncRNAs (>5kb away from protein-coding genes) that had 1-1 orthologs in all primate species and which had at least 2 mapped reads in human in a pool of brain, cerebellum, heart, kidney and liver samples, as well as 2,492 human lncRNAs that overlapped on the antisense strand with exons of protein-coding genes, which had orthologs in at least one of the other species with strand-specific data (mouse, opossum, platypus, chicken, *Xenopus*). These antisense lncRNAs were further filtered to extract genes that were expressed in human brain and testis. We evaluated Spearman's correlation coefficients between pairs of samples, on lncRNA or protein-coding gene RPKM values. All available 1-1 orthologs were used. As a control for our expression evolution analyses, we resampled the same average number of reads *per* gene for each species and tissue, keeping the proportions among genes identical to the original distribution.

Tissue-specific expression

We evaluated the tissue-specificity of the expression pattern with a previously proposed index[46], which varies between 0 for housekeeping genes and 1 for tissue-restricted genes:

$$\frac{\sum_{i=1}^{N}\left(1-\frac{\exp_i}{\exp_{max}}\right)}{N-1}$$

where N is the number of tissues, exp_i is the expression value in tissue i, and exp_{max} the maximum expression level over all tissues. We used RPKM and log2-transformed RPKM for expression values, reaching the same conclusions. The randomly expected proportion of conserved specificity across species was computed as the product of the observed proportions of tissue-specific genes in each species, for each tissue.

Reconstruction and analysis of the co-expression network

We reconstructed the evolutionarily conserved co-expression network for lncRNAs and protein-coding genes following a previously proposed method[35] (Supplementary Discussion). For each species and for each pair of genes (lncRNA or protein-coding), we computed the Pearson correlation coefficients of expression patterns. Given two homologous families, we tested if the combination of correlation coefficients measured in each species was significantly higher or lower than expected by chance. The statistical tests were done by comparing the observed ranks of the correlation coefficients with a random n-dimensional order statistics[35]. We computed correlations only for genes expressed in at least 3 samples for each species, and we computed p-values only if correlations were evaluated in at least 3 species. We allow negative connections, which have lower than expected rank combinations. We considered only lncRNAs estimated to have originated in the Eutherian ancestor or earlier, but without requiring representatives in all descendant species. As p-value computations are highly time-consuming with a large number of species, analyses were done on a representative subset of 7 species: human, macaque, mouse, opossum, platypus, chicken and *Xenopus*. For greater accuracy of the reconstruction we extended our in-house generated dataset to include previously published, comparable RNA-seq samples

(Supplementary Table 1). We visualized the network with Cytoscape[47] and we detected clusters of highly inter-connected genes with the Markov Cluster (MCL) algorithm[39].

Defining potential miRNA precursors

To search for lncRNAs that may promote transcription of miRNAs or are potentially processed into miRNAs, we extracted all miRNA hairpin sequences from miRBase[58] 18 and searched for sequence similarity (blastn[52]) against all annotated gene regions, including 10kb of flanking sequences. Genes with at least one miRNA hairpin alignment (95% identity, aligned on the entire length) on the same strand were considered potential miRNA precursors.

Statistical analyses

All statistical analyses and graphical representations (including gene expression clustering, principal component analysis, randomization tests for statistical significance) were done in R[48]. For statistical tests involving the co-expression network, we generated a set of 100 randomized networks by permuting the gene identifiers of the nodes for each edge. The randomized networks had the same distribution of edges types (positive, negative, coding-coding, coding-noncoding etc.), and the node degree was preserved. To test the significance of the network properties (e.g., *cis* connections), we derived a p-value by comparing the values observed in real and randomized networks. To compare the degrees of connectivity among gene types by controlling for unequal expression levels, we extracted lncRNAs with maximum expression levels (log2 RPKM) between 3 and 6, and divided them into 6 discrete expression classes ([3, 3.5], (3.5, 4], … ,(5.5,6] log2 RPKM). We then drew TF and non-TF protein-coding genes matching the relative proportions of lncRNAs in each expression class. The resampling was repeated 100 times.

Data availability

The sequencing data have been submitted to GEO (accession GSE43520) and SRA (PRJNA186438 and PRJNA202404). The lncRNA annotations and homologous families have been made available on the publisher's website (Supplementary Datasets 1 and 2),

as well as gene expression levels for lncRNAs and protein-coding genes (Supplementary Dataset 3) and miRNAs (Supplementary Dataset 4).

REFERENCES

1. Kosiol, C. *et al.* Patterns of positive selection in six mammalian genomes. *PLoS Genet* **4,** e1000144 (2008).

2. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).

3. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106,** 11667–11672 (2009).

4. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25,** 1915–1927 (2011).

5. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22,** 1775–1789 (2012).

6. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).

7. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28,** 503–10 (2010).

8. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science (80-.).* **309,** 1559–1563 (2005).

9. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* **105,** 716–21 (2008).

10. Young, R. S. *et al.* Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. *Genome Biol Evol* **4,** 427–42 (2012).

11. Nam, J. & Bartel, D. Long non-coding RNAs in C . elegans. *Genome Res* (2012). doi:10.1101/gr.140475.112

12. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147,** 1537–1550 (2011).

13. Chow, J. C., Yen, Z., Ziesche, S. M. & Brown, C. J. Silencing of the mammalian X chromosome. *Annu Rev Genomics Hum Genet* **6,** 69–92 (2005).

14. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415,** 810–813 (2002).

15. Dinger, M. E. *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18,** 1433–1445 (2008).

16. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–23 (2007).

17. Orom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143,** 46–58 (2010).

18. Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147,** 358–369 (2011).

19. Chodroff, R. A. *et al.* Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol* **11,** R72 (2010).

20. Marques, A. C. & Ponting, C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10,** R124 (2009).

21. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25,** 1915–27 (2011).

22. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet* **8,** e1002841 (2012).

23. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22,** 2971–2972 (2006).

24. Lin, M. F. *et al.* Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res* **17,** 1823–1836 (2007).

25. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–50 (2005).

26. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135,** 0–9 (2012).

27. Ward, L. D. & Kellis, M. Evidence of abundant purifying selection in humans for recently-acquired regulatory functions. *Science (80-.).* **337,** 1675–1678 (2012).

28. Galtier, N. & Duret, L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet* **23,** 273–7 (2007).

29. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20,** 1313–26 (2010).

30. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–82 (2011).

31. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

32. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (80-.).* **328,** 1036–40 (2010).

33. Walker, E., Manias, J. L., Chang, W. Y. & Stanford, W. L. PCL2 modulates gene regulatory networks controlling self-renewal and commitment in embryonic stem cells. *Cell Cycle* **10,** 45–51 (2011).

34. Chambers, I. & Tomlinson, S. R. The transcriptional foundation of pluripotency. *Development* **136,** 2311–22 (2009).

35. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science (80-.).* **302,** 249–255 (2003).

36. Shkumatava, A., Stark, A., Sive, H. & Bartel, D. P. Coherent but overlapping expression of microRNAs and their targets during vertebrate development. *Genes Dev* **23,** 466–81 (2009).

37. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41,** 808–815 (2012).

38. Keniry, A. *et al.* The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nat. Cell Biol* **14,** 659–65 (2012).

39. Van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol Biol* **804,** 281–95 (2012).

40. Grant, J. *et al.* Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature* **487,** 254–8 (2012).

41. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–11 (2009).

42. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28,** 511–515 (2010).

43. UniProt. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40,** D71–5 (2012).

44. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res* **40,** D290–301 (2012).

45. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* **40,** D84–90 (2012).

46. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21,** 650–9 (2005).

47. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27,** 431–2 (2011).

48. R Development Core Team. *R: A language and environment for statistical computing*. *R Found. Stat. Comput.* (2011). at <http://www.r-project.org>

49. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10,** R25 (2009).

50. Rhead, B. *et al.* The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* **38,** D613–9 (2010).

51. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comp Biol* **11,** 319–355 (2004).

52. Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. *J Mol Biol* **215,** 403–410 (1990).

53. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19,** 327–35 (2009).

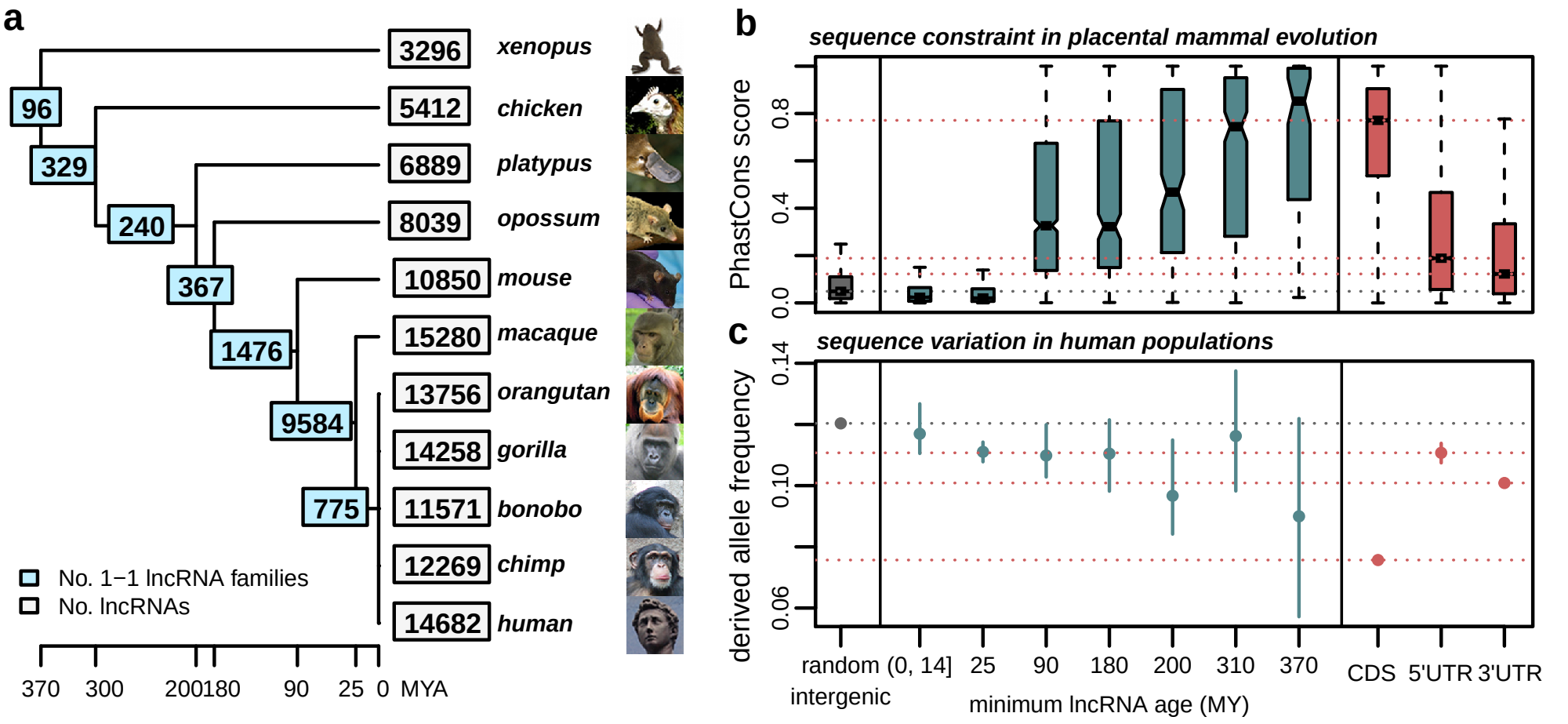
54. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14,** 708–715 (2004).

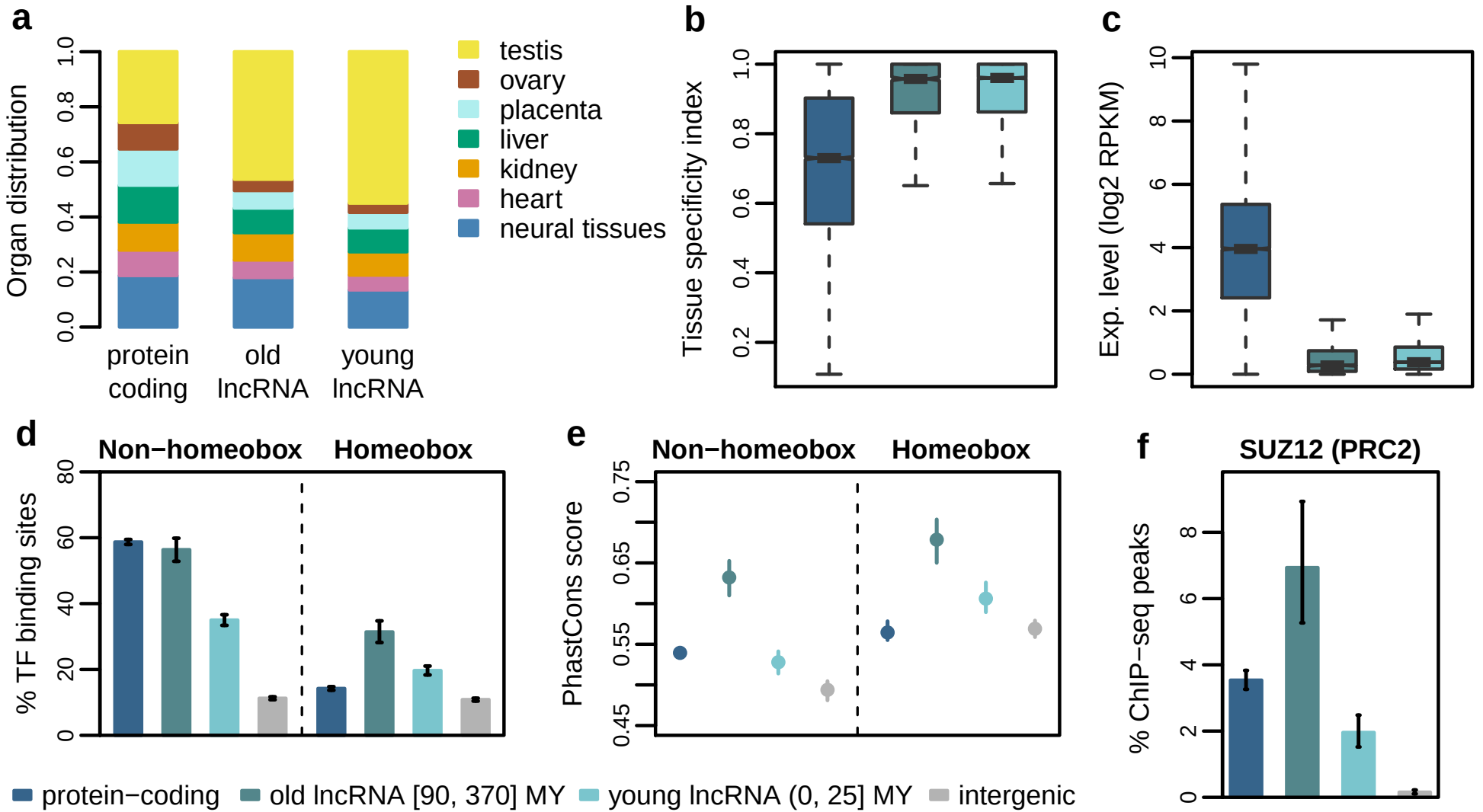
55. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–103 (2010).

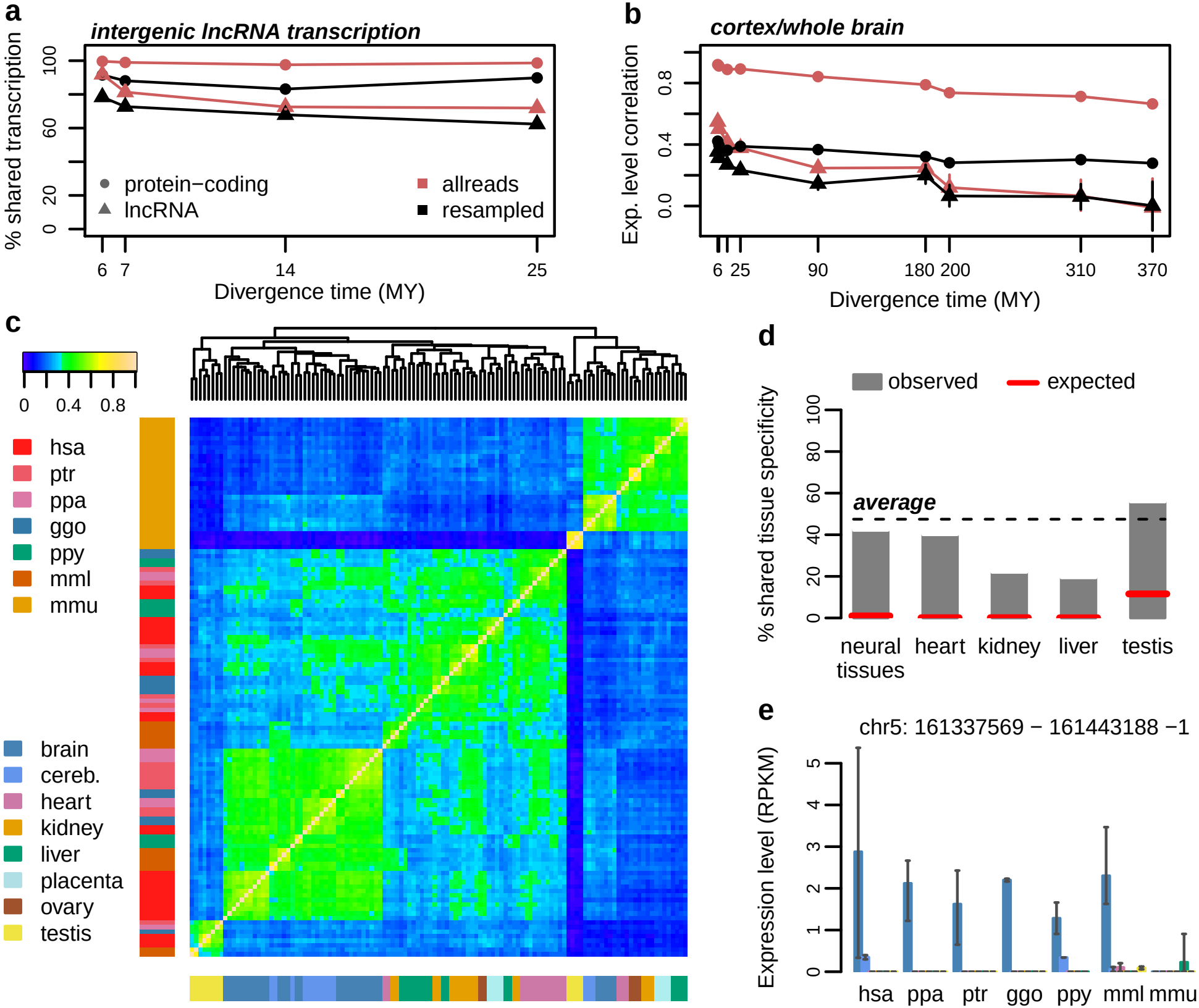
56. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100 (2012).

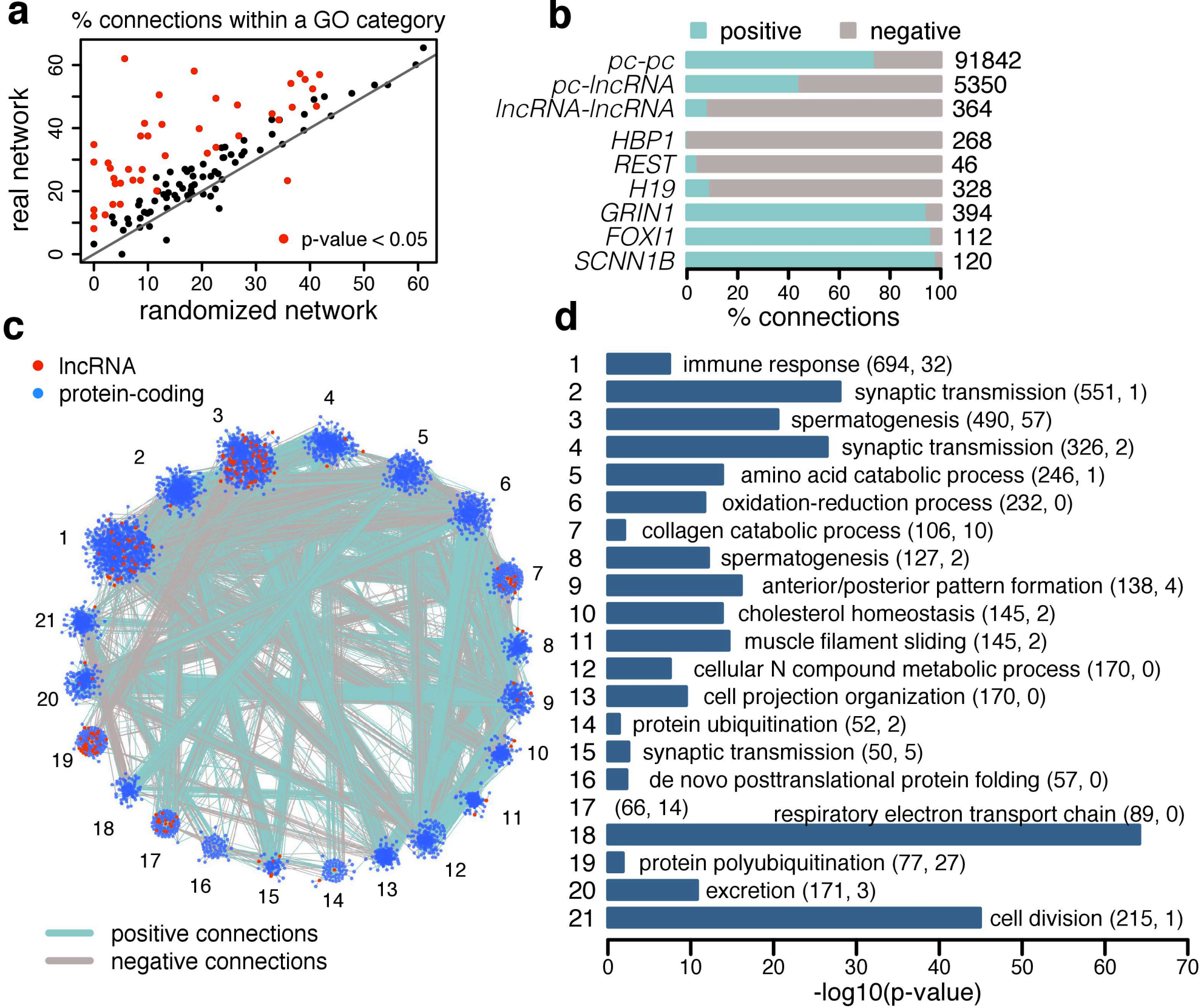
57. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38,** 626–35 (2006).

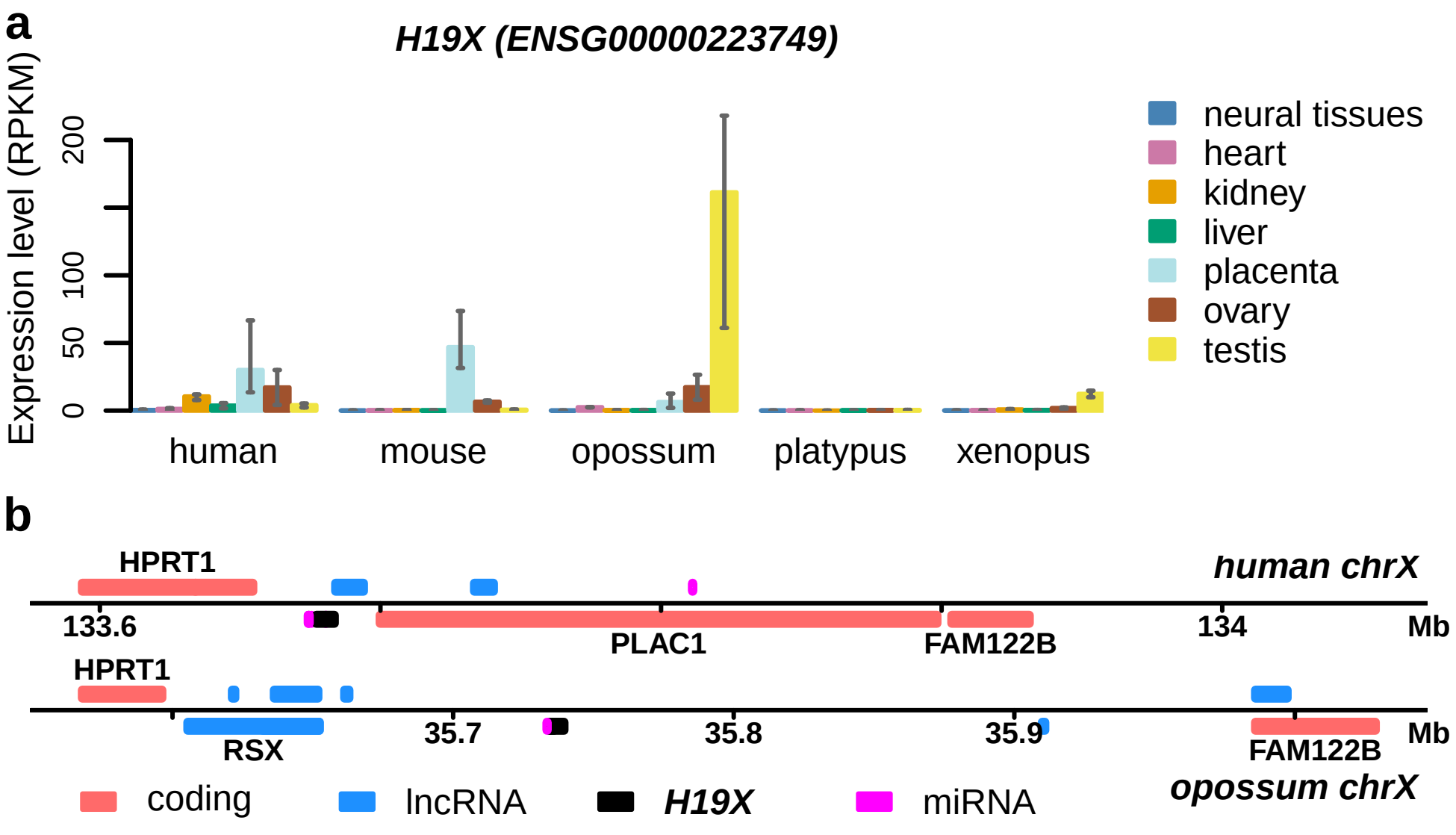
58. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39,** D152–7 (2011).

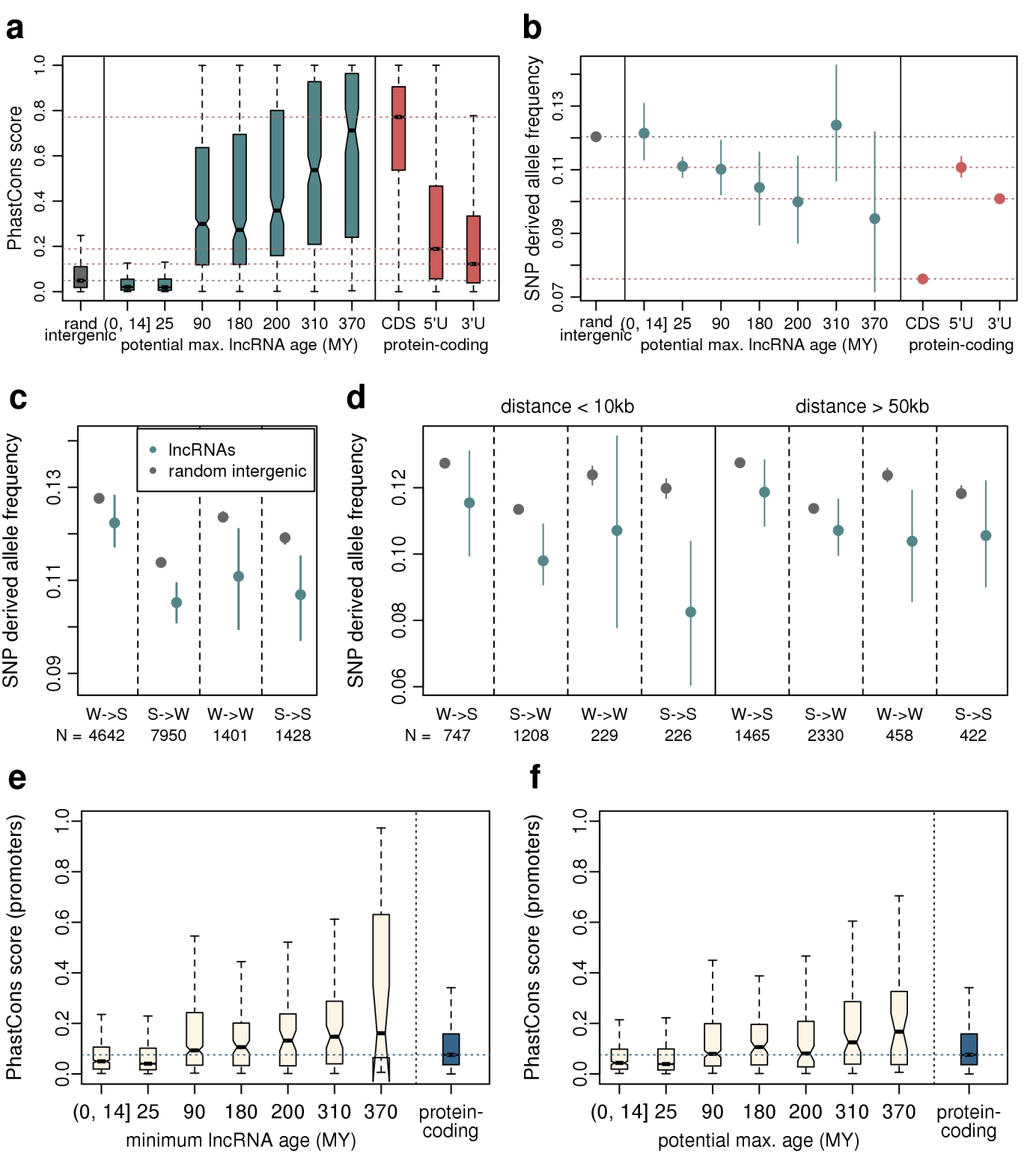

a
3296
xenopus
96
5412
chicken
329
6889
platypus
240
8039
opossum
367
10850
mouse
1476
15280
macaque
9584
13756
orangutan
14258
gorilla
775
11571
bonobo
12269
chimp
14682
human
No. 1−1 lncRNA families
No. lncRNAs
370 300 200 180 90 25 0 MYA
b
sequence constraint in placental mammal evolution
PhastCons score
0.0
0.4
0.8
c
sequence variation in human populations
derived allele frequency
0.06
0.10
0.14
random intergenic
(0, 14]
25
90
180
200
310
370
CDS
5'UTR
3'UTR
minimum lncRNA age (MY)

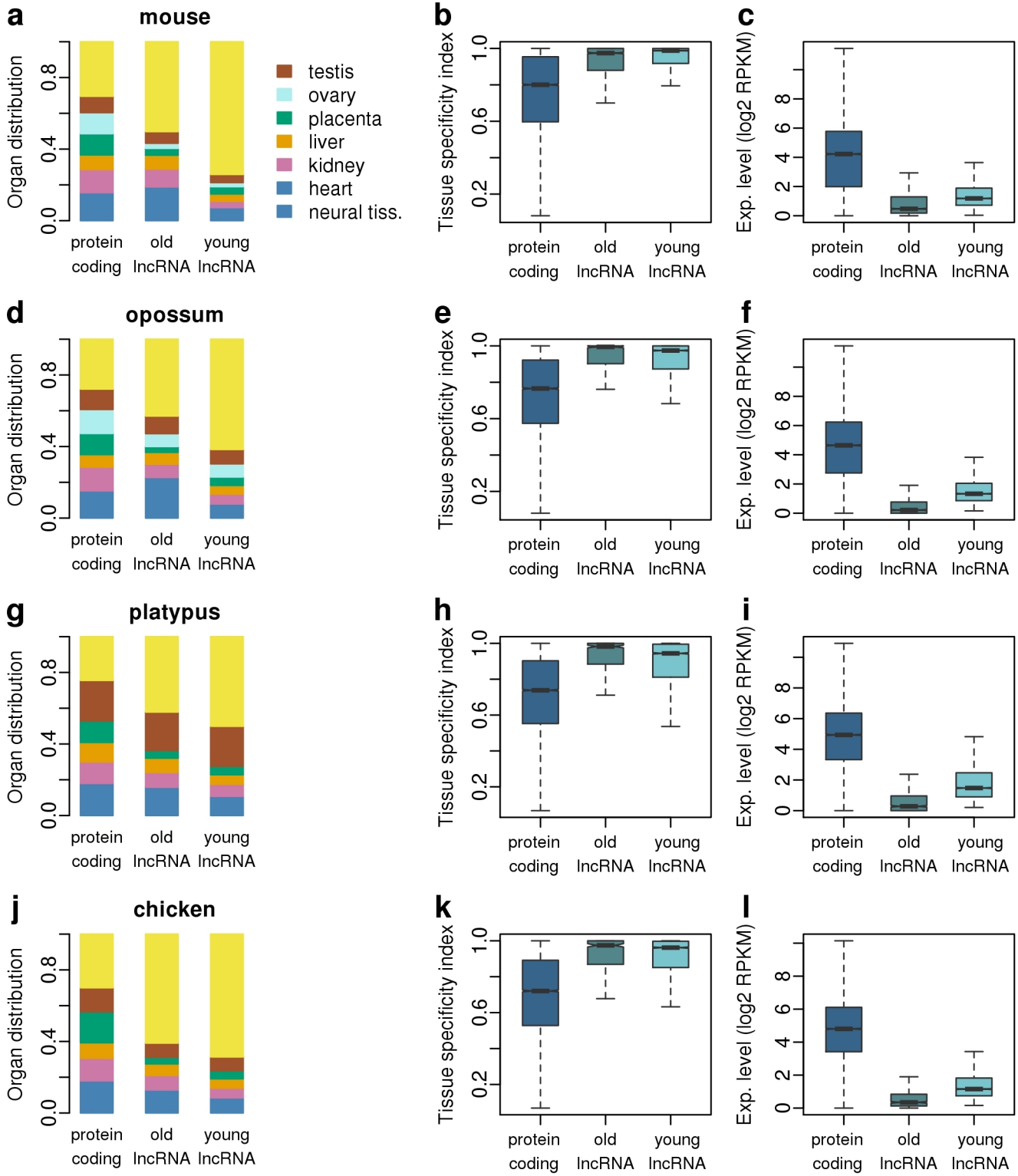

a
Organ distribution
0.0
0.2
0.4
0.6
0.8
1.0
testis
ovary
placenta
liver
kidney
heart
neural tissues
protein coding
old lncRNA
young lncRNA
b
Tissue specificity index
0.2
0.4
0.6
0.8
1.0
c
Exp. level (log2 RPKM)
0
2
4
6
8
10
d
Non−homeobox
Homeobox
% TF binding sites
0
20
40
60
80
e
Non−homeobox
Homeobox
PhastCons score
0.45
0.55
0.65
0.75
f
SUZ12 (PRC2)
% ChIP−seq peaks
0
2
4
6
8
protein−coding
old lncRNA [90, 370] MY
young lncRNA (0, 25] MY
intergenic

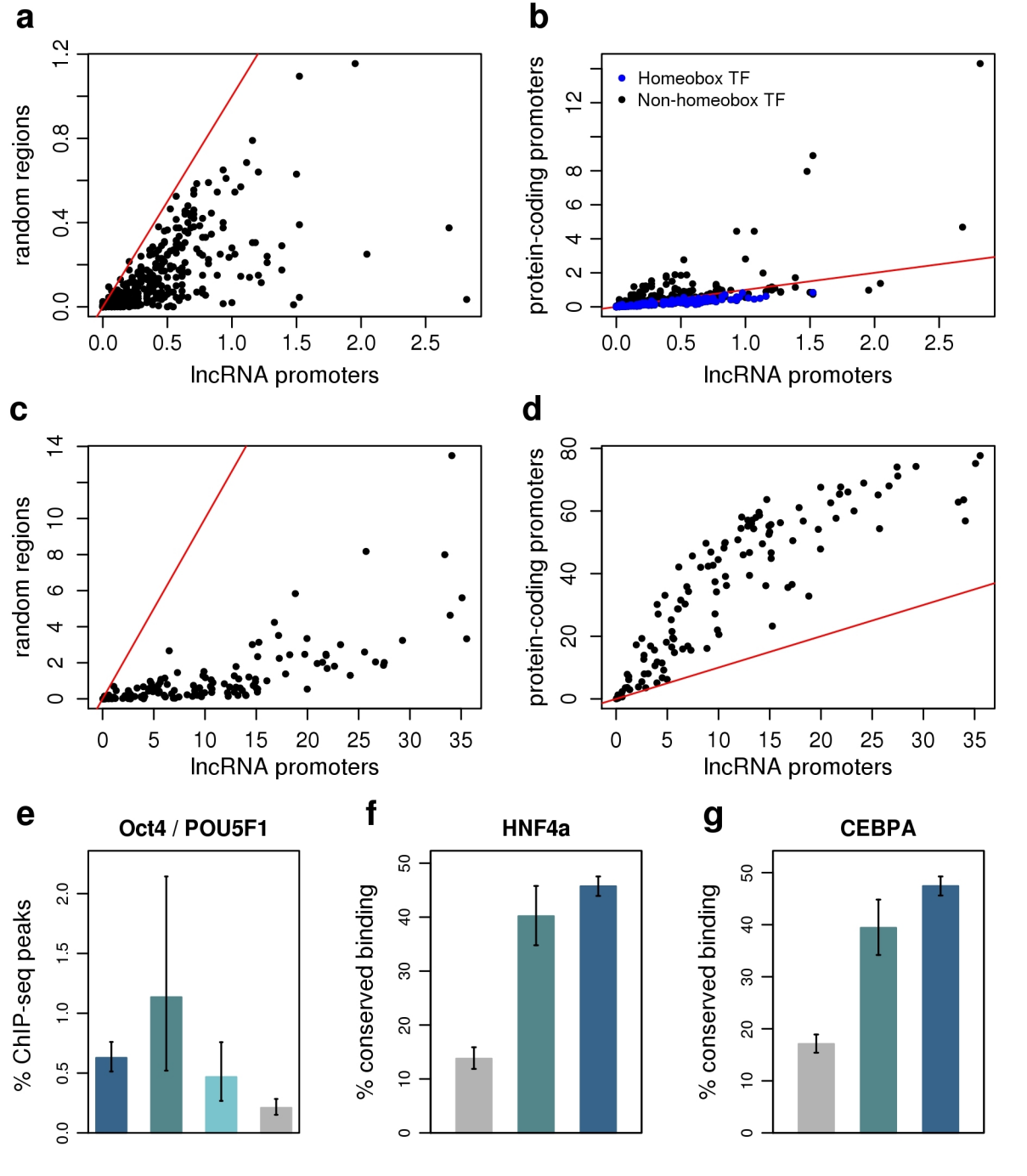

a
intergenic lncRNA transcription
% shared transcription
0
20
60
100
protein−coding
lncRNA
allreads
resampled
6
7
14
25
Divergence time (MY)
b
cortex/whole brain
Exp. level correlation
0.0
0.4
0.8
6
25
90
180
200
310
370
Divergence time (MY)
c
0
0.4
0.8
hsa
ptr
ppa
ggo
ppy
mml
mmu
brain
cereb.
heart
kidney
liver
placenta
ovary
testis
d
observed
expected
% shared tissue specificity
0
20
40
60
80
100
average
neural tissues
heart
kidney
liver
testis
e
chr5: 161337569 − 161443188 −1
Expression level (RPKM)
0
1
2
3
4
5
hsa
ppa
ptr
ggo
ppy
mml
mmu

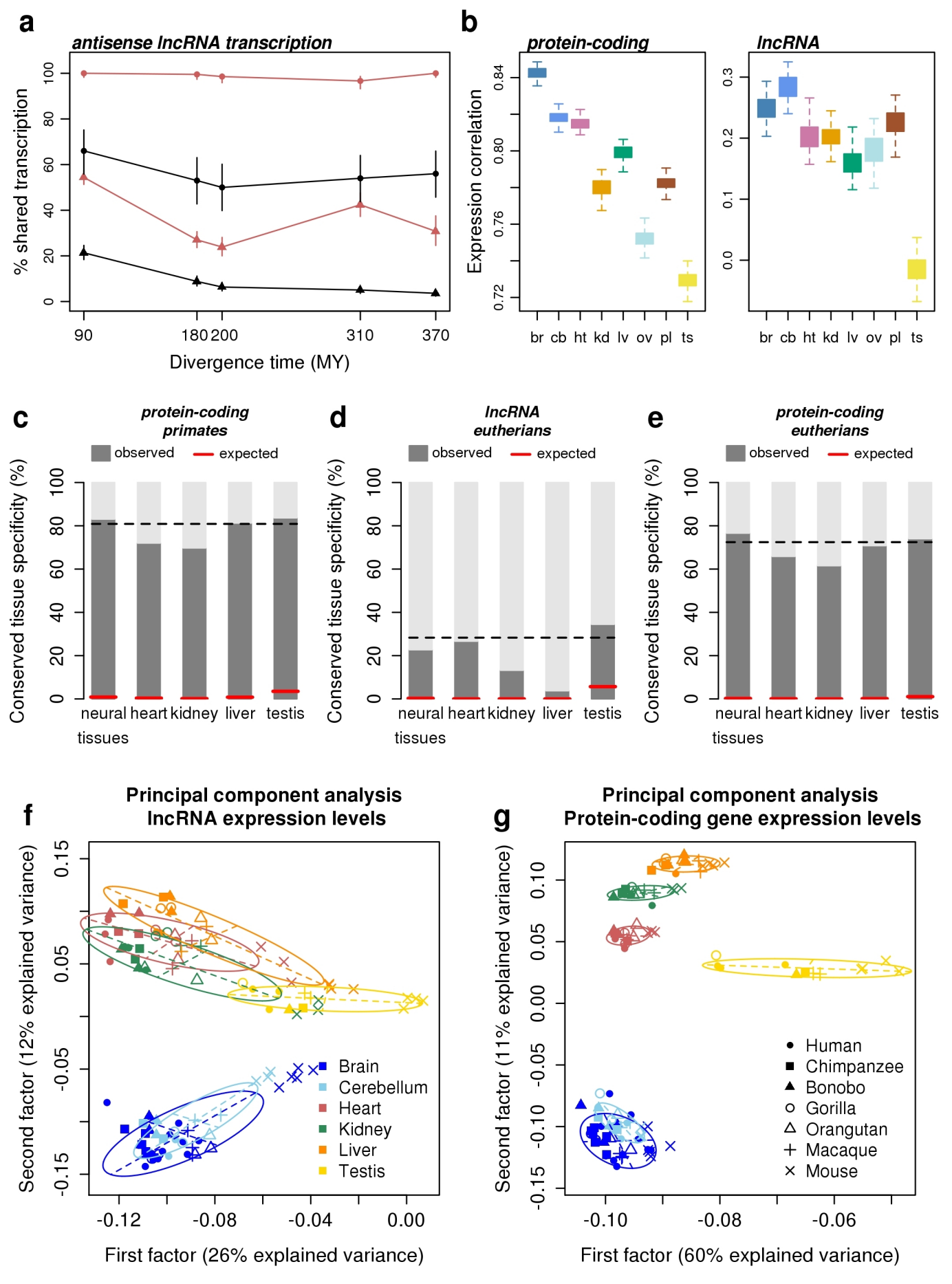

a
% connections within a GO category
real network
randomized network
p-value < 0.05
b
positive
negative
pc-pc
pc-lncRNA
lncRNA-lncRNA
HBP1
REST
H19
GRIN1
FOXI1
SCNN1B
91842
5350
364
268
46
328
394
112
120
% connections
c
lncRNA
protein-coding
positive connections
negative connections
d
1 immune response (694, 32)
2 synaptic transmission (551, 1)
3 spermatogenesis (490, 57)
4 synaptic transmission (326, 2)
5 amino acid catabolic process (246, 1)
6 oxidation-reduction process (232, 0)
7 collagen catabolic process (106, 10)
8 spermatogenesis (127, 2)
9 anterior/posterior pattern formation (138, 4)
10 cholesterol homeostasis (145, 2)
11 muscle filament sliding (145, 2)
12 cellular N compound metabolic process (170, 0)
13 cell projection organization (170, 0)
14 protein ubiquitination (52, 2)
15 synaptic transmission (50, 5)
16 de novo posttranslational protein folding (57, 0)
17 (66, 14) respiratory electron transport chain (89, 0)
18
19 protein polyubiquitination (77, 27)
20 excretion (171, 3)
21 cell division (215, 1)
-log10(p-value)

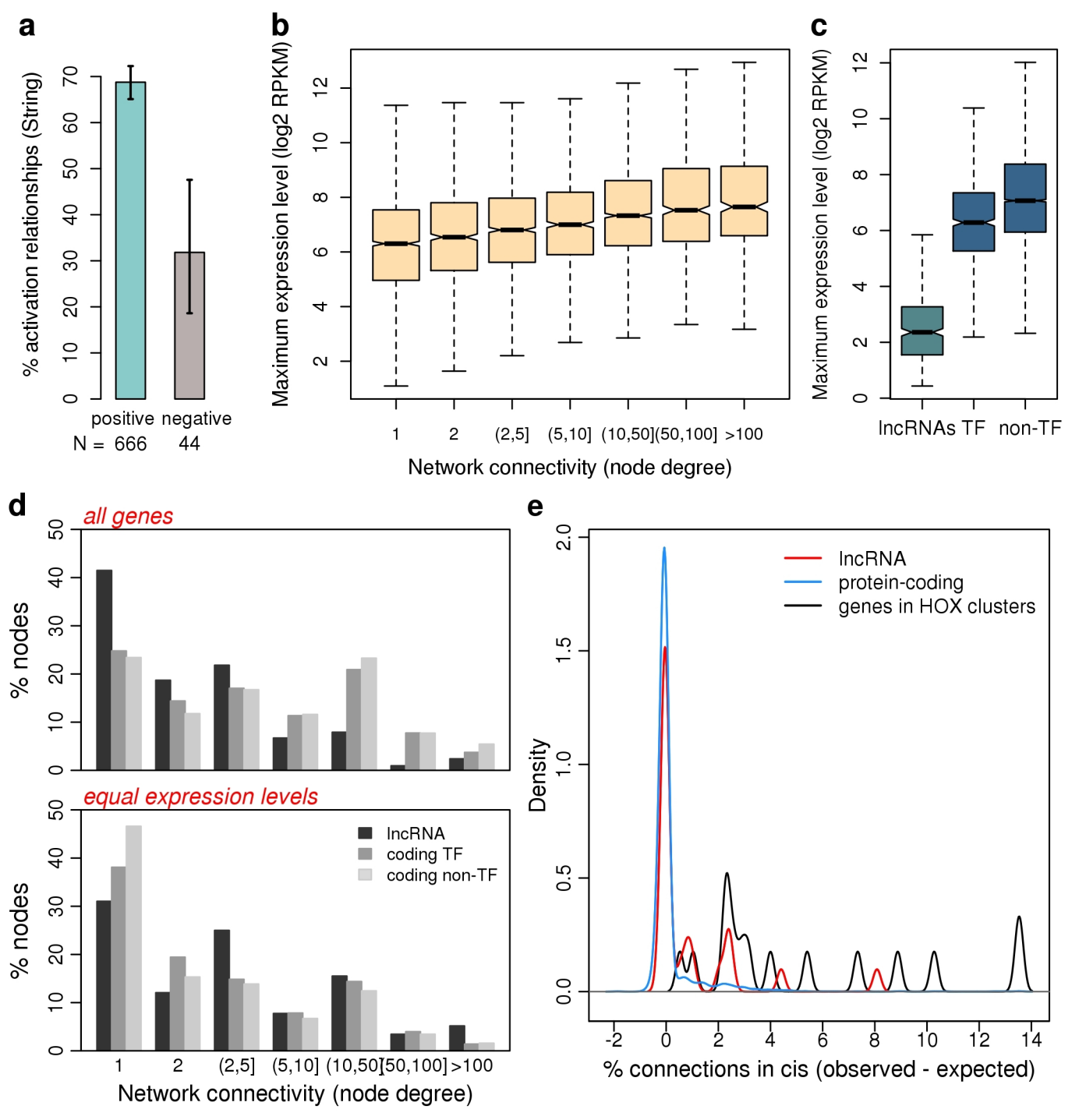

a
H19X (ENSG00000223749)
Expression level (RPKM)
0
50
100
200
human
mouse
opossum
platypus
xenopus
neural tissues
heart
kidney
liver
placenta
ovary
testis
b
HPRT1
human chrX
133.6
PLAC1
FAM122B
134
Mb
HPRT1
RSX
35.7
35.8
35.9
FAM122B
Mb
opossum chrX
coding
lncRNA
H19X
miRNA

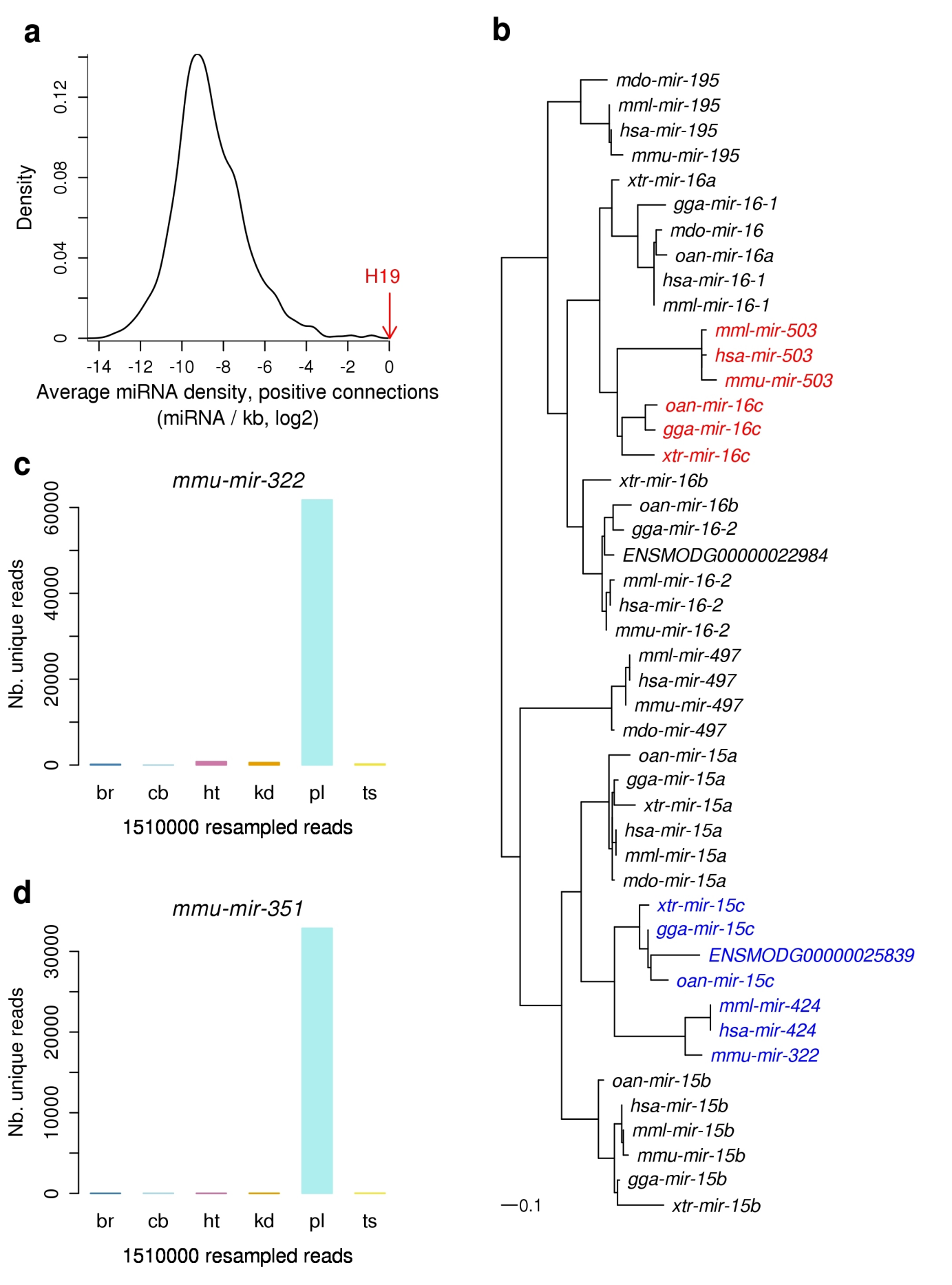

a
PhastCons score
0.0
0.2
0.4
0.6
0.8
1.0
rand intergenic
(0, 14]
25
90
180
200
310
370
CDS
5'U
3'U
potential max. lncRNA age (MY)
protein-coding
b
SNP derived allele frequency
0.07
0.09
0.11
0.13
rand intergenic
(0, 14]
25
90
180
200
310
370
CDS
5'U
3'U
potential max. lncRNA age (MY)
protein-coding
c
SNP derived allele frequency
0.09
0.11
0.13
lncRNAs
random intergenic
W->S
S->W
W->W
S->S
N = 4642
7950
1401
1428
d
distance < 10kb
distance > 50kb
SNP derived allele frequency
0.06
0.08
0.10
0.12
W->S
S->W
W->W
S->S
N = 747
1208
229
226
1465
2330
458
422
e
PhastCons score (promoters)
0.0
0.2
0.4
0.6
0.8
1.0
(0, 14]
25
90
180
200
310
370
protein-coding
minimum lncRNA age (MY)
f
PhastCons score (promoters)
0.0
0.2
0.4
0.6
0.8
1.0
(0, 14]
25
90
180
200
310
370
protein-coding
potential max. age (MY)


a
mouse
Organ distribution
0.0
0.4
0.8
testis
ovary
placenta
liver
kidney
heart
neural tiss.
protein coding
old lncRNA
young lncRNA
b
Tissue specificity index
0.2
0.6
1.0
c
Exp. level (log2 RPKM)
0
2
4
6
8
d
opossum
e
f
g
platypus
h
i
j
chicken
k
l


a
random regions
lncRNA promoters
b
Homeobox TF
Non-homeobox TF
protein-coding promoters
lncRNA promoters
c
random regions
lncRNA promoters
d
protein-coding promoters
lncRNA promoters
e
Oct4 / POU5F1
% ChIP-seq peaks
f
HNF4a
% conserved binding
g
CEBPA
% conserved binding


a
antisense lncRNA transcription
% shared transcription
Divergence time (MY)
b
protein-coding
lncRNA
Expression correlation
br cb ht kd lv ov pl ts
c
protein-coding primates
d
lncRNA eutherians
e
protein-coding eutherians
observed
expected
Conserved tissue specificity (%)
neural heart kidney liver testis
tissues
f
Principal component analysis
lncRNA expression levels
Second factor (12% explained variance)
First factor (26% explained variance)
Brain
Cerebellum
Heart
Kidney
Liver
Testis
g
Principal component analysis
Protein-coding gene expression levels
Second factor (11% explained variance)
First factor (60% explained variance)
Human
Chimpanzee
Bonobo
Gorilla
Orangutan
Macaque
Mouse


a
% activation relationships (String)
positive
negative
N = 666 44
b
Maximum expression level (log2 RPKM)
1
2
(2,5]
(5,10]
(10,50]
(50,100]
>100
Network connectivity (node degree)
c
Maximum expression level (log2 RPKM)
lncRNAs
TF
non-TF
d
all genes
% nodes
equal expression levels
lncRNA
coding TF
coding non-TF
Network connectivity (node degree)
e
lncRNA
protein-coding
genes in HOX clusters
Density
% connections in cis (observed - expected)


a
Density
0
0.04
0.08
0.12
H19
-14
-12
-10
-8
-6
-4
-2
0
Average miRNA density, positive connections
(miRNA / kb, log2)
b
mdo-mir-195
mml-mir-195
hsa-mir-195
mmu-mir-195
xtr-mir-16a
gga-mir-16-1
mdo-mir-16
oan-mir-16a
hsa-mir-16-1
mml-mir-16-1
mml-mir-503
hsa-mir-503
mmu-mir-503
oan-mir-16c
gga-mir-16c
xtr-mir-16c
xtr-mir-16b
oan-mir-16b
gga-mir-16-2
ENSMODG00000022984
mml-mir-16-2
hsa-mir-16-2
mmu-mir-16-2
mml-mir-497
hsa-mir-497
mmu-mir-497
mdo-mir-497
oan-mir-15a
gga-mir-15a
xtr-mir-15a
hsa-mir-15a
mml-mir-15a
mdo-mir-15a
xtr-mir-15c
gga-mir-15c
ENSMODG00000025839
oan-mir-15c
mml-mir-424
hsa-mir-424
mmu-mir-322
oan-mir-15b
hsa-mir-15b
mml-mir-15b
mmu-mir-15b
gga-mir-15b
xtr-mir-15b
0.1
c
mmu-mir-322
Nb. unique reads
0
20000
40000
60000
br
cb
ht
kd
pl
ts
1510000 resampled reads
d
mmu-mir-351
Nb. unique reads
0
10000
20000
30000
br
cb
ht
kd
pl
ts
1510000 resampled reads