Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression

Camille Berthelot[1,2*], Diego Villar[3*], Julie E. Horvath[4,5,6], Duncan T. Odom[3,7], Paul Flicek[1,7]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
[2]Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, 46 rue d'Ulm, 75230 Paris Cedex 05, France
[3]University of Cambridge, Cancer Research UK Cambridge Institute, Robinson Way, Cambridge, CB2 0RE, UK
[4]Biological and Biomedical Sciences, North Carolina Central University, Durham, NC 27707, USA.
[5]North Carolina Museum of Natural Sciences, Raleigh, NC 27601, USA.
[6]Evolutionary Anthropology Department, Duke University, Durham, NC 27707, USA.
[7]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
[*]These authors contributed equally to the work

[#]To whom correspondence should be addressed. Email addresses:

Duncan.Odom@cruk.cam.ac.uk and flicek@ebi.ac.uk

Abstract

To gain insight into how mammalian gene expression is controlled by rapidly evolving regulatory elements, we jointly analysed promoter and enhancer activity with downstream transcription levels in liver samples from fifteen species. Genes associated with complex regulatory landscapes generally exhibit high expression levels that remain evolutionarily stable. While the number of regulatory elements is the key driver of transcriptional output and resilience, regulatory conservation matters: elements active across mammals most effectively stabilise gene expression. In contrast, recently-evolved enhancers typically contribute weakly, consistent with their high evolutionary plasticity. These effects are observed across the entire mammalian clade and robust to potential confounders, such as gene expression level. Using liver as a representative somatic tissue, our results illuminate how the evolutionary stability of gene expression is profoundly entwined with both the number and conservation of surrounding promoters and enhancers.

INTRODUCTION

Mammalian gene expression is controlled by collections of non-coding promoter and enhancer regions, known to bind hundreds of transcription factors combinatorially[1-3]. Numerous studies have documented the rapid evolution of mammalian regulatory elements, especially enhancers[4-9], and the evolutionary turnover of tissue-specific transcription factor binding within [6,10,11].

In contrast, gene expression patterns are typically stable between species[12-14], with similar tissues across species being more correlated in expression than different tissues within a species. How stable tissue-specific gene expression is maintained by rapidly evolving collections of regulatory elements is a fundamental question in evolutionary genetics.

Previous work connecting gene expression and regulatory evolution has typically focused on how regulatory innovations direct lineage-specific phenotypes[4,8,15] (reviewed in[16]). Work across fruit flies, primates and mice has shown only limited correspondence between specific changes in gene expression and evolutionary changes in DNA methylation levels[17], transcription factor binding[18,19], or histone modifications[20].

Additionally, regulatory activities fall on a spectrum of conservation from fully conserved to lineage-specific. Reports across different species, tissues and developmental stages have suggested the greater functional relevance of conserved regulatory activity[6,7,11,19,21,22]. In contrast, lineage-specific elements appear partly compensatory of proximally lost events[11,22], and often arise in regions with pre-existing regulatory activity[4,21]. However, it remains unclear how much insight depth of conservation provides into regulatory function - partly because of a lack of datasets across divergent species[23,24] encompassing both regulatory and gene expression readouts.

Here, we evaluate the global relationship between regulatory evolution and gene expression divergence by jointly analyzing promoters, enhancers, and transcription levels measured in liver samples from fifteen mammalian species. Our results illuminate how the evolutionary resilience of gene expression is profoundly entwined with both the number and conservation of surrounding promoters and enhancers.

RESULTS

High conservation of gene expression levels across 25 mammals

We generated RNA sequencing (RNA-seq) data to quantify gene expression levels in liver tissue from a total 25 mammalian species (1-5 individuals each; Figure 1a; Methods; Supplementary Table 1). Promoters and enhancers active in liver have been reported for 20 of these species from largely the same samples[7]. Using gene annotations and orthology relationships from Ensembl[25], we compared the expression for 17,475 genes that are 1-to-1 orthologs between some or all of our study species (Figure S1; Methods).

Our results closely agree with previous reports that tissue-specific gene expression levels are highly correlated among mammalian species[12-14,26,27]. For ten species, analysis of the RNA-seq data was negatively affected by the relatively low quality of reference genome assemblies (Figure 1, greyed italics; Figure S2). Although these species were excluded from additional analyses, we have released these datasets as a resource, to allow re-analysis by the community as reference genome assemblies improve. Therefore, the analyses herein use RNA-seq data from 15 species (Figure 1, blue font).

We asked whether conservation of expression levels is higher for groups of functionally-related genes, such as housekeeping genes[28] or genes with tissue-specific liver functions[29]. Because comparing the evolutionary stability of different subsets of genes is confounded by gene expression levels (Supplemental Text 1 and Figure S2), we matched each gene of interest with a control gene of similar expression, using the mean expression across species as the reference value (Methods). Confirming previous reports, housekeeping and core liver genes exhibited higher expression correlation across species than controls (Wilcoxon signed-rank test: both $p < 2*10^{-16}$; Figure 1b-c)[12,30]. In addition to gene expression correlation, we also used the coefficient of variation of each gene as a measure of divergence to classify genes as evolutionarily stable or variable (i.e. inter-species standard deviation normalised by mean expression across species; Methods, Figure S2). Both housekeeping and core liver genes were more likely to be classified as stable

(Chi-squared test: $p < 2*10^{-16}$ and $p = 2*10^{-8}$, respectively; Figure S2). Our results indicate that the expression levels of genes relevant to tissue function are under stabilizing evolutionary pressure, as proposed previously in other tissues[12,13] and developmental contexts[31]. Nevertheless, a substantial fraction of each set was classified as variable, suggesting that functionally relevant genes can exhibit large dynamic expression ranges across species. Thus, the coefficient of variation captures a different aspect of gene expression evolution than the expression correlation used in previous studies.

The number of promoters and enhancers correlates with gene expression stability across evolution

We sought to connect how gene expression evolution may be directed by the evolution of promoters and enhancers across mammals. To characterize regulatory landscapes, we used the profiles of two histone modifications (H3K4me3 and H3K27ac) previously obtained using ChIP-seq in twenty mammalian species, largely from the same liver samples we report here[7] (Table S1; Figure S3). Active promoters typically display high levels of both modifications[32,33], whereas H3K27ac marking alone is representative of active enhancers[34,35] (Figure S3).

We first asked how gene expression levels are affected by the overall complexity of regulatory landscapes. We defined complexity as the number of promoters and enhancers assigned to each gene in each species. As in our previous work[7], a regulatory association domain is defined for each gene as a genomic window up and downstream of the gene's TSS, following the strategy used by GREAT[36] (Figure 2a, Methods). In general, this approach associates a single regulatory element to no more than two genes. Nevertheless, some gene misassignments will occur for a fraction of regulatory elements, especially among enhancers[37-39].

We observed that regulatory complexity is moderately correlated across species (Figure 2b), reflecting the rapid evolution of mammalian regulatory elements[7]. To summarise the regulatory landscape at each gene, we took the median number of promoters and enhancers across species as a representative value in an average mammal (Figure 2c). Genes associated with larger numbers of transcriptional regulatory elements are more highly expressed (Figure 2d and S4), as observed in a single species[40-42]. This was especially true for enhancers, suggesting that the majority of the active enhancers identified have a measurable effect on gene expression (Figure 2d). Conversely, promoters showed more of a switch-like effect, where one active promoter is necessary to turn the gene on, but additional promoters are not associated with substantially higher gene expression levels. These associations were not due to biased ChIP-seq signal intensity or artefacts associated with highly expressed genes (Figure S4), or definition of regulatory association domains (Figure S5).

We next asked whether the number of promoters or enhancers associated to a gene also influence the evolutionary conservation of gene expression levels. To do this, genes associated to multiple promoters or enhancers across species were compared to control genes matched for expression level but with only a simple regulatory landscape (Figure 2e, grey insets). Genes with complex regulatory inputs showed significantly increased expression conservation (Wilcoxon signed-rank test: promoters and enhancers both $p < 2*10^{-16}$; Figure 2e).

We looked for sequence or experimental features differentiating the regulatory landscapes of genes with evolutionarily stable or variable expression. The elements associated with the two classes showed only marginal differences in reproducibility, signal coverage, sequence conservation, and information content (Figure S6). The absence of clearly discriminative features suggests that gene expression stability could be largely driven by sheer numbers of regulatory elements.

Overall, these observations support a direct connection among the complexity of the regulatory landscape, gene expression, and gene expression conservation. In the remaining sections, we leverage our extensive phylogenetic scope to explore how regulatory conservation and regulatory complexity together influence gene expression evolution.

Conserved regulatory activity associates with high and evolutionarily stable gene expression

Active promoters are largely functionally conserved across mammalian species, while enhancer activity evolves rapidly[6,7,9,15,43]. Conserved regulatory regions are thought to be particularly important for gene expression control[21,44-46], but definitive evidence beyond individual cases is limited[23,47,48].

In previous work, we identified 1,872 promoters and 279 enhancers that exhibit conserved activity in the livers of most placental mammals ("placental-conserved" regulatory elements, Figure 3a; Figure S3)[7]. Placental-conserved elements typically are a minority within a gene's regulatory landscape, although these elements may disproportionally contribute to the levels and/or stability of gene expression. As previously reported[7], placental-conserved elements, and especially enhancers, showed longer length, more intense ChIP-seq signal, higher sequence constraint and information content, indirectly supporting their functional importance (Figure S7).

We first asked whether placental-conserved regulatory elements contribute more to gene expression levels than other elements. Genes associated with conserved elements exhibited higher transcription levels than controls associated with the same number of regulatory elements where none are placental-conserved (Figure 3b; Wilcoxon rank sum test: promoters $p = 2*10^{-8}$; enhancers $p = 0.001$). This result was consistent whether the expression was measured using the mean expression across all species or in a representative species (e.g. human; Figure S8). Thus, highly expressed genes appear to be associated with regulatory regions more likely to be maintained during evolution. Indeed, housekeeping and core liver genes are significantly more likely to be associated with placental-conserved promoters (Figure S8).

We next isolated the contribution of placental-conserved regulatory activity to gene expression stability. We compared sets of genes matched for expression levels and total number of associated regulatory regions, but differing by the presence or absence of placental-conserved elements. Genes associated with placental-conserved elements were more correlated in expression across species than those without (Figure 3c; Wilcoxon signed-rank test: promoters $p < 2*10^{-16}$; enhancers $p = 7*10^{-16}$). Analyzing expression stability based on the coefficient of variation further supports the enhanced importance of conserved regulatory elements (Figure S8).

Taken together, these results demonstrate that deeply conserved elements contribute disproportionately to maintaining both high and stable gene expression levels across species.

Recently-evolved regulatory activity has modest impact on gene expression levels

We also previously identified 794 promoters and 10,434 enhancers that were reproducibly active in human liver, but not in the livers of any other study species (Figure 4a; Figure S3 and Methods). Compared to the bulk of active enhancers, recently-evolved enhancers displayed only marginal differences in experimental and sequence features, including shorter length and lower sequence constraint, information content and experimental reproducibility (Figure S7). By combining these with our gene expression data, we asked whether recently-evolved regulatory elements influence gene expression (Figure 4b), and if so how much (Figure 4c).

Human genes putatively regulated by recently-evolved promoters are typically expressed well above background and show no difference in expression compared to control genes with more conserved promoters (Wilcoxon rank sum test: p = 0.64; Figure 4c). New promoters therefore seem as likely to be functional as those shared with at least one other species: indeed, 57% of genes targeted by a recently-evolved promoter in human apparently did not rely on any other promoter for expression in liver.

Whether recently-evolved enhancers have a measurable effect on gene expression is more problematic to establish, largely because identifying enough control genes was challenging. Specifically, 42% of human genes with 1-to-1 placental orthologs are associated with recently-evolved enhancers, and these genes were more likely to be associated with enhancers shared across species (mean: 3.3 vs. 0.8 shared enhancers;

Wilcoxon rank sum test: $p < 2*10^{-16}$). We therefore limited our analyses to the subsets of human genes that could be matched for expression level and/or landscape complexity (Methods).

Overall, our results revealed that recently-evolved enhancers typically increase gene expression slightly less than do shared enhancers. First, the presence of recently-evolved enhancer(s) is associated with modestly higher expression compared to control genes with the same background of evolutionarily shared enhancers (Wilcoxon rank sum test: $p = 4*10^{-5}$; Figure 4b). Second, compared to genes with the same total number of enhancers, genes with recently-evolved enhancer(s) exhibit slightly lower expression (Wilcoxon rank sum test: $p = 2*10^{-6}$; Figure 4c). These results were confirmed in other species (Figure S9), where recently-evolved elements span different age depths.

Together, these observations indicate that recently-evolved regulatory elements have a measurable effect on gene expression. Our results depict recently-evolved enhancers as functionally weaker than those active in several species, consistent with previous observations regarding the age of conserved DNA sequences active during mammalian cortical development[21]. Nevertheless, recently-evolved regulatory activity appears at least partly functional and pervasively modulates gene expression across species.

Recently-evolved elements consistently contribute to increased expression stability

We asked if recently-evolved regulatory activities have weaker stabilising impact on gene expression than do shared regulatory elements. At the scale of a single species (human), we observed that genes with and without recently-evolved regulatory elements showed no difference in expression conservation when controlling for total number of regulatory elements and expression level (Wilcoxon signed-rank test: promoters p = 0.43, enhancers p = 0.24; Figure S10). Moreover, recently-evolved human enhancers were equally likely to be associated to genes with either evolutionarily stable or variable expression (Chi-squared test: p = 0.11); if anything, recently-evolved promoters were weakly associated with stable gene expression (Chi-squared test: $p = 0.03$). Thus, recently-evolved regulatory activity in a single species has no obvious relationship with expression divergence.

Interestingly, recently-evolved promoters and enhancers from different species concentrated more often than expected in the vicinity of the same genes (Figure 5a-c). This effect remained significant regardless of the size of the regulatory association domain (Figure S10). We delineated a set of genes recurrently associated with recently-evolved elements across different mammals (Figure 5a; Methods). Surprisingly, these genes were significantly more correlated in expression than expected based on their expression levels (Wilcoxon signed-rank test: promoters and enhancers both $p < 2*10^{-16}$; Figure 5d). However, these genes also exhibited particularly complex regulatory landscapes (1.3 promoters and 8.6 enhancers on average), which associate with stable gene expression. When controlling for both expression level and the total number of enhancers, genes with a recurrent accumulation of recently-evolved enhancers exhibited faster expression divergence than those without (Wilcoxon signed-rank test: promoters and enhancers both $p < 2*10^{-16}$; Figure 5e). In contrast, the accumulation of recently-evolved promoters associated with increased gene expression stability.

These results suggest that recently-evolved elements contribute to gene expression stability across species by maintaining the complexity of the regulatory landscape. Our findings extend previous observations on the evolutionary turnover of liver-specific transcription factor binding, for which newly acquired binding locations were often proximal to lost binding events - thus likely compensatory[11,22]. Nevertheless, recently-evolved enhancers appear weaker at buffering expression changes than elements conserved across species.

The composite liver regulatory landscape across mammals

Previous sections have dissected the impact of regulatory elements that are either conserved across species (Figure 3), or singular to one species in the dataset (Figures 4 and 5); however, these elements make up less than half of the regulatory regions identified in every species (Figure S3). Here, we exploit the full extent of our genome-wide datasets to characterize the continuous relationship between regulatory evolution and gene expression.

We built a reference-free map of the regulatory landscape across mammalian livers (Methods; Figure 6a) by projecting all twenty regulatory landscapes onto a single summary landscape for each gene, to create 17,475 meta-genes. These meta-genes collect all the independent regulatory elements associated to a gene, regardless of the number or subset of species in which each element is active (Figure 6a). Therefore each meta-gene's summary landscape explicitly integrates both regulatory complexity and regulatory conservation. The reference-free map treats each meta-promoter and meta-enhancer as a single evolutionary acquisition and describes regulatory evolution with simple metrics (total accumulated elements across lineages and number of species where activity is present; Methods). On average, meta-genes were associated with 2.3 meta-promoters and 11 meta-enhancers (sd: 2.2 and 13.0, respectively).

To investigate the overall impact of regulatory complexity on gene expression evolution, we stratified meta-genes by the number of associated meta-promoters (Figure 6b) and meta-enhancers (Figure 6c). We observed that expression level and stability increase steadily with regulatory complexity across the entire mammalian gene set. Interestingly, this trend was consistent for promoters as well as enhancers, whereas in an average mammal, promoters show a more switch-like behaviour (Figure 2). Therefore, integrating regulatory information across twenty species increases the resolution to detect the impact of multiple meta-promoters.

We next asked how the entire spectrum of regulatory conservation impacts gene expression across the full set of 17,475 orthologous genes. Meta-promoters (Figure 6d) and meta-enhancers (Figure 6e) were classified by the number of species in which they are active, and we measured the expression and evolutionary stability of associated meta-genes. Strikingly, the level and stability of gene expression tracks with the conservation of the regulatory landscape. This result extends and complements our observations on highly-conserved and recently-evolved elements, and suggests that the gradual relationship between conservation of regulatory landscapes and stability of expression is a general feature of mammalian gene regulation.

These data also illustrate the difficulty of predicting expression level or stability of specific genes, even when informed by enhancer and promoter maps from twenty mammals. In fact, we observe substantial variability within, and overlap between, all of the meta-element classes (insets of Figure 6b-e). Nevertheless, our integrated analysis reveals how regulatory complexity and conservation interplay to shape expression level and expression stability across mammalian genomes.

DISCUSSION

The majority of evolutionary differences across species are expected to be driven by alterations in gene expression rather than by changes in the protein sequences[49,50]. Previous comparative studies have shown that gene expression is globally correlated across species, with similar tissues displaying stronger correlation of gene expression than different tissues within a species[12,26]. These tissue-specific expression patterns, however, display significant evolutionary divergence. How much of this divergence results from modifications of the regulatory landscape is not fully understood.

To date, comparative approaches to understanding gene regulation have largely focused on lineage-specific innovations, identifying candidate regions driving lineage-specific phenotypes[4,8,15,17] (reviewed in [16]). Concurrently, evolutionarily conserved regulatory elements are thought to play a predominant role in gene regulation[45,51,52], while the functional relevance of less conserved elements has been the subject of speculation[21,22,53,54]. To extend these analyses, we collected an integrated dataset of gene expression output and regulatory histone marks from the same liver samples across a wide array of mammalian species. This strategy allowed us to systematically test the contributions of both landscape complexity (i.e. number of regulatory elements) and landscape conservation on gene expression evolution.

Our key finding is that the transcription of a gene is evolutionarily stabilised by the presence of many regulatory elements regardless of their conservation. In other words, gene expression level and its evolutionary resilience reflect the complexity of the regulatory landscape, both within a single species and across mammals. However, regulatory regions are not functionally equal: those highly-conserved across placental mammals exert a more powerful stabilizing effect, associating with gene expression levels that are simultaneously high and evolutionarily stable. In contrast, recently-evolved enhancers contribute more weakly to gene expression and transcriptional stability, consistent with a model whereby a fraction of new-born elements have a neutral role on gene expression evolution[55,56]. These effects are clear throughout our data, whether considering a full-scale, reference-free map of mammalian regulatory complexity, or investigating subsets of extremely conserved or divergent regulatory elements. Our discoveries extend previous reports connecting evolutionary constraint on regulatory activities with expression outputs[4,18,20,57], and are consistent with an enhanced functional importance of conserved regulatory elements[44,46,58].

There are a number of limitations to our approach. First, the precise measurements of regulatory complexity, conservation, and gene expression are partly dependent on the reference genome assembly and annotation, which are of variable completeness across our study species.

Second, our strategy to connect regulatory elements to putative targets is based on genomic proximity. This simplification can mis-assign distal enhancers; this could partly explain the noisy correlation between enhancer activity and gene expression. Experimentally linking regulatory activity to target genes would refine evaluation of how individual enhancer elements contribute to transcriptional output[59-61]. Further, this approach inherently assigns a larger number of regulatory elements to genes surrounded by larger intergenic regions, such as transcriptional regulators[62]. Whether larger intergenic regions simply produce more regulatory activity, or whether a demand for increased transcriptional control expands intergenic space around those genes remains unresolved.

Third, mapping regulatory activity in other tissues or additional signatures of regulatory activity, such as open chromatin[63], other histone modifications[64] or co-activator proteins[65], may identify other features that contribute to gene expression evolution. Nevertheless, other reports across tissues and developmental stages[4,6,10,19,21,22] consistently showed a similar regulatory plasticity as we observed in adult liver – suggesting our results are representative of regulatory evolution in most somatic tissues.

Fourth, we did not explore how the often-poorly annotated non-coding transcriptome evolves, where different regulatory principles may apply[66,67].

Finally, phylogenetic frameworks for functional genomics data remain elusive[68,69]. Integrating the evolution of gene regulation and expression with the species-tree structure promises to afford greater resolution into the regulatory rewiring of mammalian genomes, but will require denser phylogenetic sampling than this study provides and additional methodological development[70].

This study suggests a general framework of how transcriptional output and transcriptional regulation co-evolve. Active regulatory elements have long been known to additively contribute to gene expression control[44,71]. By connecting transcriptional control with gene expression in fifteen species, we demonstrate how the number of active promoters and enhancers relates to gene expression stability across mammals. Our observations are consistent with existing models of enhancer function[72-75], and provide mechanistic insight into how conserved transcriptional outputs can be achieved by complex and rapidly evolving regulatory landscapes.

METHODS

Ethics statement

The investigation was approved by the Animal Welfare and Ethics Review Board and followed the Cambridge Institute guidelines for the use of animals in experimental studies under Home Office license PPL 70/7535. Human liver samples were obtained under Human Tissue Act license 08-H0308-117 from the Addenbrooke's Hospital at the University of Cambridge with patients' consent, and was approved by the National Research Ethics Service.

Source and detail of tissues

We quantified gene expression profiles in liver samples from 25 species by RNA extraction coupled to high throughput sequencing (RNA-seq), typically from 2-4 individuals per species. In most cases, these are the same samples we previously used in ChIP-seq experiments to assess regulatory activity across twenty mammals[7]. The origin, number of replicates, sex and age for each species' samples are detailed in Table S1.

Total RNA-sequencing (RNA-seq) library preparation

Total RNA was extracted from snap-frozen liver tissue with RNAeasy Mini Kit (Qiagen). 20 mg of tissue were weighed on dry-ice and immediately homogenized in 600 microliters (ul) of RLT buffer containing 10 ul of beta-mercaptoethanol per mililiter of buffer. Tissue samples were homogenized in a Precellys 24 tissue homogenizer, using settings 5500-2x15-015 and Precellys tubes CK28-R (bertin technology). Liver homogenates were processed according to manufacturers' instructions (Qiagen RNAeasy Mini Kit) and total RNA eluted in 50ul RNAse-free water. 10 ug total RNA from each sample were treated with 4 units of Turbo DNAse (Ambion), and total RNA samples were run on an Agilent Bioanalyser (RNA nano chip) to check RNA integrity. Samples were taken forward if RIN values were above 7. Ribosomal RNA (rRNA) was depleted with Ribo-Zero rRNA removal kit (Epicentre RZC110424) as per instructions from the manufacturer, using 5 ug of DNAse-treated total RNA.

Strand-specific rRNA-depleted RNA-seq libraries were prepared with a modified version of TruSeq RNA Library Preparation kit (Illumina). Fragmentation and first-strand synthesis of rRNA-depleted RNA samples were according to the Illumina protocol. Second-strand cDNA synthesis was done with SuperScript double-stranded cDNA synthesis kit (Life Technologies) at 16C for two hours, using a 10mM dATP, dCTP, dGTP, dUTP nucleotide mix. cDNA was purified with QIAquick PCR purification kit (Qiagen) and end repair, A-tailing and adaptor ligation were performed with Illumina's protocol. Second-strand degradation was achieved by treatment with one unit of Uracil N-Glycosylase (Life Technologies) at 35C for 15 minutes, prior to PCR enrichment. Libraries were amplified according to Illumina's protocol for 13 PCR cycles, and cleaned-up with Agencourt AMPure beads (Beckman Coulter) with a 1:1 DNA:beads ratio. RNA-seq libraries were quantified with Kapa Library quantification kit (Kapa biosystems) on a QuantStudio 6 Flex instrument (Applied Biosystems), pooled in equimolar amounts and sequenced to a minimum depth of twenty million uniquely mapped reads on an Illumina HiSeq 2500 instrument. Libraries were sequenced as either paired-end 100 bp or paired-end 150bp.

RNA-seq alignment and gene expression quantification

RNA-seq reads from de-multiplexed fastq files were trimmed to 100 bp and aligned to the corresponding reference genomes and full transcript sets available in Ensembl and Ensembl Pre! v.73[25] using TopHat 2.0.13[76] with default parameters and a mate pair inner distance (-r) of 75 bp. Aligned reads were subsampled to 20 million read pairs per sample, a read depth close to saturation for protein-coding genes[77] (see also Figure S1 for a depth saturation analysis in human).

Transcript quantification was performed using Cufflinks 2.2.1[78] (default parameters), based on the transcript annotations available in Ensembl v.73[25]. Estimated gene expression levels in FPKM (fragment per kb of exon per million mapped fragments) were obtained from the Cufflinks gene summary output. The gene expression levels were further transformed into TPM (transcripts per million transcripts).

Genes annotated in human with 1-to-1 orthologs in one or more species were identified from the gene phylogenies available in Ensembl v.73[25]. In each species where a unique ortholog was identified, the mean expression level over all available replicates was used as the representative expression level for this gene. Orthologous expression levels were further normalized between species using the median of ratios to the geometric means, as described in [79].

Assignment of active regulatory regions to putative target genes

Regulatory elements were assigned to putative target genes following rules similar to those implemented in GREAT[36]. A regulatory association domain was defined for each gene as the genomic window up and downstream of the TSS, until the TSS of the next gene and within 1 Mb. Additionally, genes were exclusively assigned those regulatory elements directly at the TSS (up to 5 kb upstream and 1 kb downsteam). In general, this approach associates a single regulatory element to no more than two genes, with a few exceptions in case of overlapping genes and/or extremely close TSSs. For each gene, the TSS annotation used was that of the reference ("canonical") transcript in Ensembl v.73[25].

This procedure was performed in each species independently. The median number of promoters or enhancers assigned to each orthologous gene across species was used as the representative value for this gene in an average mammal.

Measures of gene expression divergence between test sets and matched controls

Evolutionary divergence of gene expression was measured by the Spearman correlation coefficients for orthologous expression levels between pairs of species. The relative divergence of two gene subsets was compared using the correlations within each subset across all pairs of species (Wilcoxon paired rank sum test). Species phylogenetic trees were built by hierarchical clustering based on the pairwise correlations of gene expression levels using complete linkage. The unweighted pair group method with arithmetic mean (UPGMA) gave similar results.

When estimating the relative divergence of particular gene subsets, confounding effects due to differences in expression levels distributions were controlled for by matching genes one-to-one to control genes with similar expression. Each gene set of interest was matched to a distinct set of controls in R with the MatchIt library[80] using the caliper option to prune genes that could not be matched to an appropriate control (increments of 0.1 to 0.001). Matching for the number of regulatory elements was performed similarly, either on its own or in combination with gene expression level. Of note, this matching approach is limited to a few categories and/or matching variables, as otherwise only a small number of genes can be matched for comparison.

Expressed genes (mean expression across species > 1 TPM; 10,704 genes) were also classified into evolutionarily stable and variable genes based on their coefficient of variation (standard deviation across species normalized by mean expression; bottom 50%: stable; top 50%: variable). Because these two categories had different mean expression levels (18.8 vs. 27.5 TPM; Wilcoxon rank sum test: $p < 2.10^{-16}$), we additionally matched stable and variable genes into pairs with similar mean expression across species as described above with MatchIt (4,264 genes in each group), and removed 2,176 unmatched genes from the subsets.

Recently-evolved regulatory elements

Recently-evolved regulatory elements were identified in each of the ten highest-quality reference genomes in our dataset (human, macaque, marmoset, mouse, rat, rabbit, cow, pig, dog and cat), all of which are included in the multiple whole genome alignment available from Ensembl. Regulatory elements were defined as recently-evolved when they either could not be aligned to an orthologous sequence in any of the other genomes, or when their orthologous loci in other genomes showed no significant enrichment in regulatory histone marks, as described in [7]. As previously reported, the vast majority of recently-evolved promoters corresponded to non-alignable sequences. Most recently-evolved enhancers could be aligned to orthologous loci in other species, but these orthologous locations showed no evidence of regulatory activity[7].

Genes recurrently targeted by recently-evolved promoters were defined as genes associated with a recently-evolved promoter in five species or more out of the ten (i.e. median number of recently-evolved promoters across species ≥ 0.5; 1,075 genes). Genes recurrently targeted by recently-evolved enhancers were defined as genes associated with three or more recently-evolved enhancers in more than five species (i.e. median number of recently-evolved enhancers across species ≥ 3; 530 genes). Conversely, control genes rarely targeted by recently-evolved promoters were defined as genes associated with no recently-evolved promoter in five species or more; genes rarely targeted by recently-evolved enhancers had one or no recently-evolved enhancers in five species or more.

All-vs-all inter-species analysis of promoter and enhancer activity

Regulatory elements identified in each species were first mapped to their orthologous loci in each of the ten highest-quality reference genomes in our dataset (human, macaque, marmoset, mouse, rat, rabbit, cow, pig, dog and cat). These ten species are all cross-mappable against each other via a single multi-species alignment. Elements from the other species are mapped via pairwise alignments to one or more of these ten reference species, as described in [7]. In each of these ten coordinate systems, regulatory elements active in two or more species were considered to be orthologous and merged into a consensus element when their genomic coordinates overlapped by 50% or more, using the *bedmap* utility from BEDOPS v.2.4.20 (option --fraction-either 0.5)[81]. This procedure resulted in ten independent maps of regulatory activity; each integrating regulatory elements from the 10 or more species aligned to this reference. All ten independent maps were then merged into a master regulatory map containing meta-elements. This map thereby integrates all regulatory elements identified in all species as long as a given element had an orthologous locus in at least one of the ten reference species.

Meta-elements in the master regulatory map were assigned to putative target meta-genes based on the collection of predicted targets in each individual species. Specifically, a gene was considered a predicted target if it was identified as such in at least half of the species where the regulatory element is active. Similarly, meta-elements were identified as meta-promoters or meta-enhancers based on their predominant histone marking across the orthologous elements integrated in the map.

Data availability

RNA-sequencing data has been deposited under Array Express accession number E-MTAB-4550, with the exception of three human and four mouse datasets (previously reported in E-MTAB-4052). ChIP-seq data from twenty mammalian species were previously reported in ArrayExpress accession number E-MTAB-2633. Links to raw data files and processed data are available at http://www.ebi.ac.uk/research/flicek/publications/FOG20..

Python and R scripts used to process the data are available upon request.

AUTHOR CONTRIBUTIONS

CB, DV, PF, DTO designed experiments; DV performed experiments; CB, DV analyzed the data; JEH provided tissue samples; CB, DV, PF, DTO wrote the manuscript; PF, DTO oversaw the work. All authors read and approved the final manuscript.

COMPETING INTEREST STATEMENT

Paul Flicek is a member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd.

ACKNOWLEDGEMENTS

We thank the Cambridge Institute Biological Resource Unit and Genomics Core for technical support, Claudia Kutter, Christine Feig, Maša Roller and Tim Rayner for useful comments and discussions, and the EMBL-EBI systems team for management of computational resources. This research was supported by Cancer Research UK grant 20412; the European Research Council grant 615584; the Welcome Trust grant numbers WT108749/Z/15/Z, WT098051, WT202878/A/16/Z and WT202878/B/16/Z; the French Institut National de la Santé et de la Recherche Médicale (INSERM) and the European Molecular Biology Laboratory. Cetacean samples were collected by the UK Cetacean Strandings Investigation Programme, funded by Defra and the Governments of Scotland and Wales.

REFERENCES

1 Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics* **13**, 613-626, doi:10.1038/nrg3207 (2012).

2 Moorthy, S. D. *et al.* Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome research* **27**, 246-258, doi:10.1101/gr.210930.116 (2017).

3 Shin, H. Y. *et al.* Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nature genetics* **48**, 904-911, doi:10.1038/ng.3606 (2016).

4 Cotney, J. *et al.* The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154**, 185-196, doi:10.1016/j.cell.2013.05.056 (2013).

5 Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381-1392, doi:S0092-8674(12)00574-0 [pii]
10.1016/j.cell.2012.04.029 (2012).

6 Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007-1012, doi:10.1126/science.1246426 (2014).

7 Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566, doi:10.1016/j.cell.2015.01.006 (2015).

8 Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159, doi:10.1126/science.1260943 (2015).

9 Young, R. S. *et al.* The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome research* **25**, 1546-1557, doi:10.1101/gr.190546.115 (2015).

10 Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature genetics* **42**, 631-634, doi:10.1038/ng.600 (2010).

11 Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036-1040, doi:10.1126/science.1186176 (2010).

12 Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343-348, doi:10.1038/nature10532 (2011).

13 Chan, E. T. *et al.* Conservation of core gene expression in vertebrate tissues. *J Biol* **8**, 33, doi:jbiol130 [pii]
10.1186/jbiol130 (2009).

14 Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593-1599, doi:10.1126/science.1228186 (2012).

15 Prescott, S. L. *et al.* Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell* **163**, 68-83, doi:10.1016/j.cell.2015.08.036 (2015).

16 Reilly, S. K. & Noonan, J. P. Evolution of Gene Regulation in Humans. *Annual review of genomics and human genetics* **17**, 45-67, doi:10.1146/annurev-genom-090314-045935 (2016).

17 Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K. & Gilad, Y. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS genetics* **7**, e1001316, doi:10.1371/journal.pgen.1001316 (2011).

18 Wong, E. S. *et al.* Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome research* **25**, 167-178, doi:10.1101/gr.177840.114 (2015).

19 Paris, M. *et al.* Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. *PLoS genetics* **9**, e1003748, doi:10.1371/journal.pgen.1003748 (2013).

20 Cain, C. E., Blekhman, R., Marioni, J. C. & Gilad, Y. Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics* **187**, 1225-1234, doi:10.1534/genetics.110.126177 (2011).

21 Emera, D., Yin, J., Reilly, S. K., Gockley, J. & Noonan, J. P. Origin and evolution of developmental enhancers in the mammalian neocortex. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E2617-2626, doi:10.1073/pnas.1603718113 (2016).

22 Arnold, C. D. *et al.* Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nature genetics* **46**, 685-692, doi:10.1038/ng.3009 (2014).

23 Chatterjee, S., Bourque, G. & Lufkin, T. Conserved and non-conserved enhancers direct tissue specific transcription in ancient germ layer specific developmental control genes. *BMC developmental biology* **11**, 63, doi:10.1186/1471-213X-11-63 (2011).

24 Necsulea, A. & Kaessmann, H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nature reviews. Genetics* **15**, 734-748, doi:10.1038/nrg3802 (2014).

25 Flicek, P. *et al.* Ensembl 2013. *Nucleic acids research* **41**, D48-55, doi:10.1093/nar/gks1236 (2013).

26 Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome biology* **16**, 287, doi:10.1186/s13059-015-0853-4 (2015).

27 Perry, G. H. *et al.* Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome research* **22**, 602-610, doi:10.1101/gr.130468.111 (2012).

28 Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends in genetics : TIG* **29**, 569-574, doi:10.1016/j.tig.2013.05.010 (2013).

29 Odom, D. T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378-1381, doi:10.1126/science.1089769 (2004).

30 She, X. *et al.* Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC genomics* **10**, 269, doi:10.1186/1471-2164-10-269 (2009).

31 Israel, J. W. *et al.* Comparative Developmental Transcriptomics Reveals Rewiring of a Highly Conserved Gene Regulatory Network during a Major Life History Switch in the Sea Urchin Genus Heliocidaris. *PLoS biology* **14**, e1002391, doi:10.1371/journal.pbio.1002391 (2016).

32 Santos-Rosa, H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407-411, doi:10.1038/nature01080 (2002).

33 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39**, 311-318, doi:10.1038/ng1966 (2007).

34 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).

35 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283, doi:10.1038/nature09692 (2011).

36 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

37 Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature genetics* **48**, 488-496, doi:10.1038/ng.3539 (2016).

38 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

39 Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E. G., Singaravelu, K. & Beyer, A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. *PLoS computational biology* **9**, e1003342, doi:10.1371/journal.pcbi.1003342 (2013).

40 Odom, D. T. *et al.* Core transcriptional regulatory circuitry in human hepatocytes. *Molecular systems biology* **2**, 2006 0017, doi:10.1038/msb4100059 (2006).

41 Mikkelsen, T. S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156-169, doi:10.1016/j.cell.2010.09.006 (2010).

42 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077, doi:10.1126/science.1232542 (2013).

43 Jubb, A. W., Young, R. S., Hume, D. A. & Bickmore, W. A. Enhancer Turnover Is Associated with a Divergent Transcriptional Response to Glucocorticoid in Mouse and Human Macrophages. *Journal of immunology* **196**, 813-822, doi:10.4049/jimmunol.1502009 (2016).

44 Cheng, Y. *et al.* Principles of regulatory information conservation between mouse and human. *Nature* **515**, 371-375, doi:10.1038/nature13985 (2014).

45 Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482, doi:10.1038/nature10530 (2011).

46 McLean, C. & Bejerano, G. Dispensability of mammalian DNA. *Genome research* **18**, 1743-1751, doi:10.1101/gr.080184.108 (2008).

47 Kvon, E. Z. *et al.* Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**, 633-642 e611, doi:10.1016/j.cell.2016.09.028 (2016).

48 Royo, J. L. *et al.* Transphyletic conservation of developmental regulatory state in animal evolution. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 14186-14191, doi:10.1073/pnas.1109037108 (2011).

49 Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics* **8**, 206-216, doi:10.1038/nrg2063 (2007).

50 King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).

51 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

52 Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature reviews. Genetics* **13**, 59-69, doi:10.1038/nrg3095 (2011).

53 Lowdon, R. F., Jang, H. S. & Wang, T. Evolution of Epigenetic Regulation in Vertebrate Genomes. *Trends in genetics : TIG* **32**, 269-283, doi:10.1016/j.tig.2016.03.001 (2016).

54 Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome research* **24**, 1963-1976, doi:10.1101/gr.168872.113 (2014).

55 Cooper, G. M. & Brown, C. D. Qualifying the relationship between sequence conservation and molecular function. *Genome research* **18**, 201-205, doi:10.1101/gr.7205808 (2008).

56 Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 6131-6138, doi:10.1073/pnas.1318948111 (2014).

57 Zhou, X. *et al.* Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome biology* **15**, 547, doi:10.1186/s13059-014-0547-3 (2014).

58 Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nature reviews. Genetics* **15**, 221-233, doi:10.1038/nrg3481 (2014).

59 Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature genetics* **47**, 598-606, doi:10.1038/ng.3286 (2015).

60 Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome research* **25**, 582-597, doi:10.1101/gr.185272.114 (2015).

61 Kieffer-Kwon, K. R. *et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155**, 1507-1520, doi:10.1016/j.cell.2013.11.039 (2013).

62 Zeitlinger, J. & Stark, A. Developmental gene regulation in the era of genomics. *Developmental biology* **339**, 230-239, doi:10.1016/j.ydbio.2009.12.039 (2010).

63 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

64 Pradeepa, M. M. *et al.* Histone H3 globular domain acetylation identifies a new class of enhancers. *Nature genetics* **48**, 681-686, doi:10.1038/ng.3550 (2016).

65 May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nature genetics* **44**, 89-93, doi:10.1038/ng.1006 (2011).

66 Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635-640, doi:10.1038/nature12943 (2014).

67 Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS genetics* **8**, e1002841, doi:10.1371/journal.pgen.1002841 (2012).

68 Dunn, C. W., Luo, X. & Wu, Z. Phylogenetic analysis of gene expression. *Integrative and comparative biology* **53**, 847-856, doi:10.1093/icb/ict068 (2013).

69 Rohlfs, R. V., Harrigan, P. & Nielsen, R. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Molecular biology and evolution* **31**, 201-211, doi:10.1093/molbev/mst190 (2014).

70 Dunn, C. W., Zapata, F., Munro, C., Siebert, S. & Hejnol, A. Pairwise comparisons across species are problematic when analyzing functional genomic data. *bioRxiv*, doi:https://doi.org/10.1101/107177 (2017).

71 Visel, A. *et al.* Functional autonomy of distant-acting human enhancers. *Genomics* **93**, 509-513, doi:10.1016/j.ygeno.2009.02.002 (2009).

72 Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of cellular biochemistry* **94**, 890-898, doi:10.1002/jcb.20352 (2005).

73 Biggin, M. D. Animal transcription networks as highly connected, quantitative continua. *Developmental cell* **21**, 611-626, doi:10.1016/j.devcel.2011.09.008 (2011).

74 Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170-1187, doi:10.1016/j.cell.2016.09.018 (2016).

75 Panne, D. The enhanceosome. *Current opinion in structural biology* **18**, 236-242, doi:10.1016/j.sbi.2007.12.002 (2008).

76 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

77 Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome research* **21**, 2213-2223, doi:10.1101/gr.124321.111 (2011).

78 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

79 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

80 Ho, D., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* **42**, 1-28 (2011).

81 Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919-1920, doi:10.1093/bioinformatics/bts277 (2012).

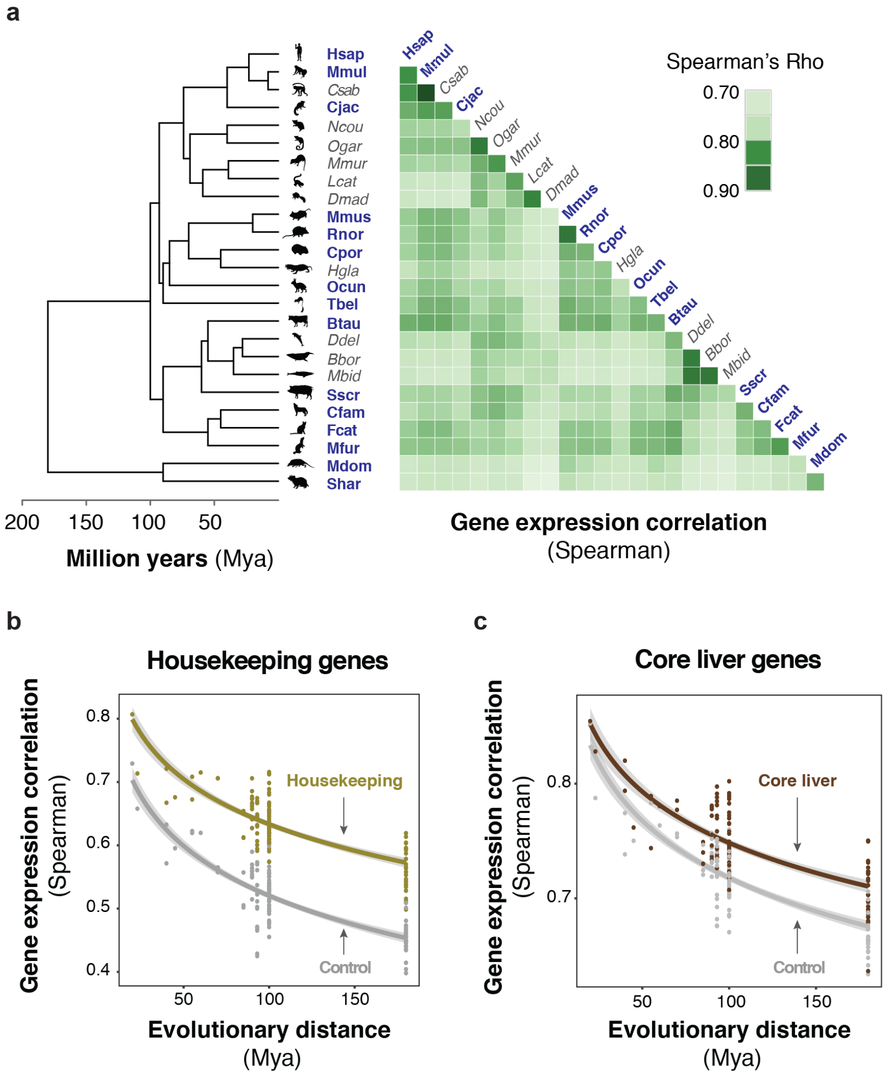

Figure 1: Liver gene expression levels are highly conserved across 25 mammalian species

(a) Pairwise correlations of normalized expression levels for 17,475 one-to-one orthologous genes in livers isolated from 25 mammalian species show high conservation of gene expression. Shading of individual tiles in the heatmap depict pairwise correlation coefficients between species (Spearman's Rho). Known phylogenetic relationships and species divergences are represented by an evolutionary tree (left of Y-axis), which includes twenty-three placental species (in four orders) and two marsupial species (in two orders). In bolded blue: species with higher-quality reference genomes; in grey: species with either draft or proxy reference genomes (excluded from analysis, Methods and Figure S2).

(b-c) Housekeeping[28] and core liver genes[29] show slower expression divergence, compared to controls matched for gene expression levels. Pairwise correlation values were plotted against evolutionary distance for housekeeping (gold, **b**; 3,612 genes) and core liver genes (brown, **c**; 2,224 genes), and compared to the correlation values of control genes with the same distribution of mean expression levels across species (grey). Control genes were matched in expression to either housekeeping (**b**) or core liver genes (**c**), and are thus different sets for the two panels. Lines correspond to linear regression trends (after log transform of the time axis), with 95% confidence intervals in grey shading, and were added for visualisation purposes. Regression R^2 are reported in Table S2.

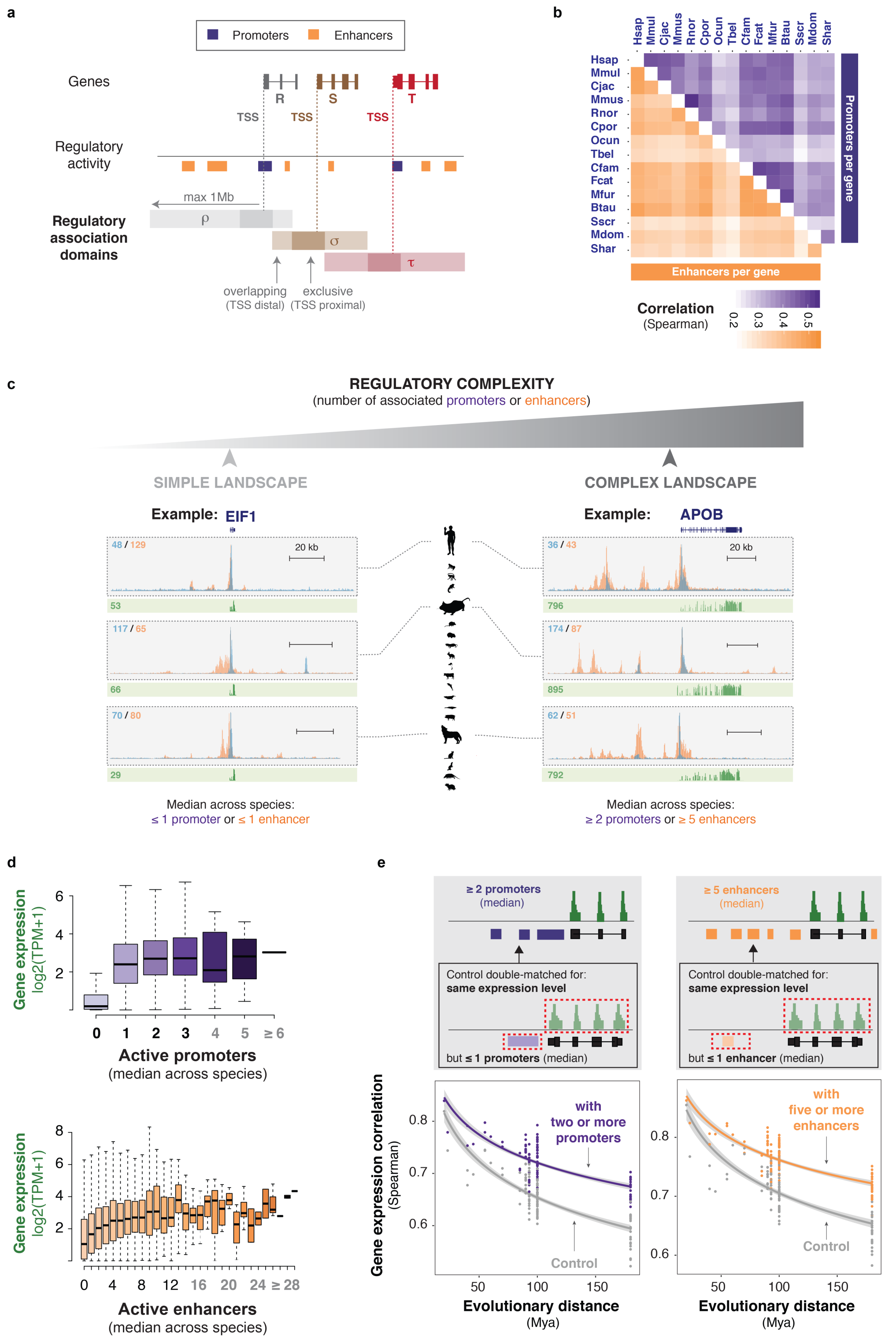

Figure 2: The number of promoters and enhancers corresponds with gene expression stability across evolution

(a) Genes are associated with all active regulatory elements sitting between their TSS and the TSS of the next gene on either side, within a limit of 1Mb. Regulatory elements sitting directly on the TSS of a gene (max. 5kb upstream and 1kb downstream) were exclusively associated to that gene (darker shading, exclusive TSS proximal). The cartoon example illustrates this procedure for three genes R, S, and T and their regulatory association domains ρ, σ and τ.

(b) Numbers of promoters and enhancers associated to a gene are correlated across species. Shading of individual tiles corresponds to pairwise tie-corrected Spearman correlation coefficients for numbers of promoters and enhancers associated to orthologous genes across 15 mammalian species.

(c) Examples of genes with simple (*EIF1)* and complex (*APOB)* regulatory landscapes in liver. Regulatory complexity was measured as the median number of promoters and enhancers associated to each gene across species. Histone modification ChIP-seq fold enrichments are shown in blue (H3K4me3) and orange (H3K27ac), and RNA-seq reads in green, for three representative species (human, mouse and dog). Numbers in blue and orange: maximum fold enrichment; numbers in green: gene expression values (TPM, normalised across species).

(d) Expression distributions (mean expression across species) are shown for genes associated with increasing numbers of active promoters (purple) or enhancers (orange) in an average mammal. Active enhancers associated to a gene have an additive effect, whereas promoters show a more switch-like effect on gene expression levels. Classes containing fewer than thirty genes are greyed.

(e) The number of associated promoters and enhancers contributes to evolutionary stability of gene expression. **Grey insets**: Expression divergence across species is compared between (i) genes associated to multiple promoters or enhancers (top) and (ii) control genes with the same expression level but associated to few promoters or enhancers (one or none, bottom). **Plots**: Pairwise Spearman correlation coefficients of expression levels between species were plotted against evolutionary distance for genes associated with multiple promoters (left; 1,688 genes) or enhancers (right; 1,479 genes), and compared to control gene sets. In both cases the number of associated promoters or enhancers corresponds to the median number across species. Lines are as described in Figure 1b-c.

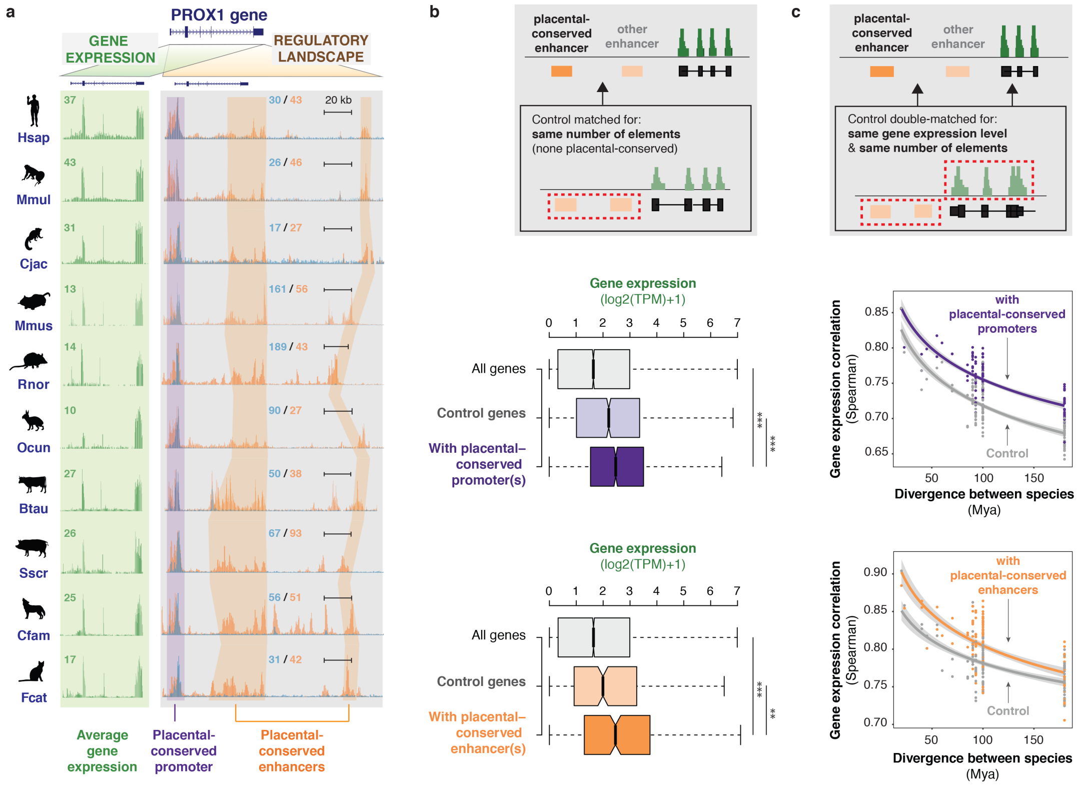

Figure 3: Conserved regulatory activity is associated with both high and stable gene expression levels

(a) Example of gene expression and regulatory landscapes around the PROX1 gene in livers from ten placental mammals. Each row shows PROX1 expression (left, green background) and activity of promoters and enhancers around the PROX1 locus in one species (H3K4me3 (blue) and H3K27ac (orange) ChIP-seq signals, grey background; as described in Figure 2C). A placental-conserved promoter and two placental-conserved enhancers at this locus are highlighted.

(b) Genes associated with placental-conserved promoters and enhancers show high expression levels. **Grey inset:** The contribution of placental-conserved regulatory activity to gene expression was evaluated using control genes associated with the same number of active promoters or enhancers, none of which are placental-conserved. **Boxplots** show the distribution of mean expression levels across species for all 1-to-1 orthologs (all genes); for genes associated with placental-conserved elements (dark purple for promoters, 2,384 genes; dark orange for enhancers, 387 genes); and for control genes (pale purple for promoters, pale orange for enhancers). ***: $p < 0.001$, **: $p < 0.01$, Wilcoxon rank sum test.

(c) Genes associated to placental-conserved promoters and enhancers exhibit slow expression divergence across species. **Grey inset:** The contribution of placental-conserved regulatory activity to gene expression conservation was evaluated using control genes with similar expression levels and associated with the same number of active promoters or enhancers, none of which are placental-conserved. **Plots**: Pairwise Spearman correlation coefficients of expression levels between species were plotted against evolutionary distance, for genes associated with placental-conserved promoter(s) (purple) or enhancer(s) (orange) and control gene sets. Lines are as described in Figure 1b-c.

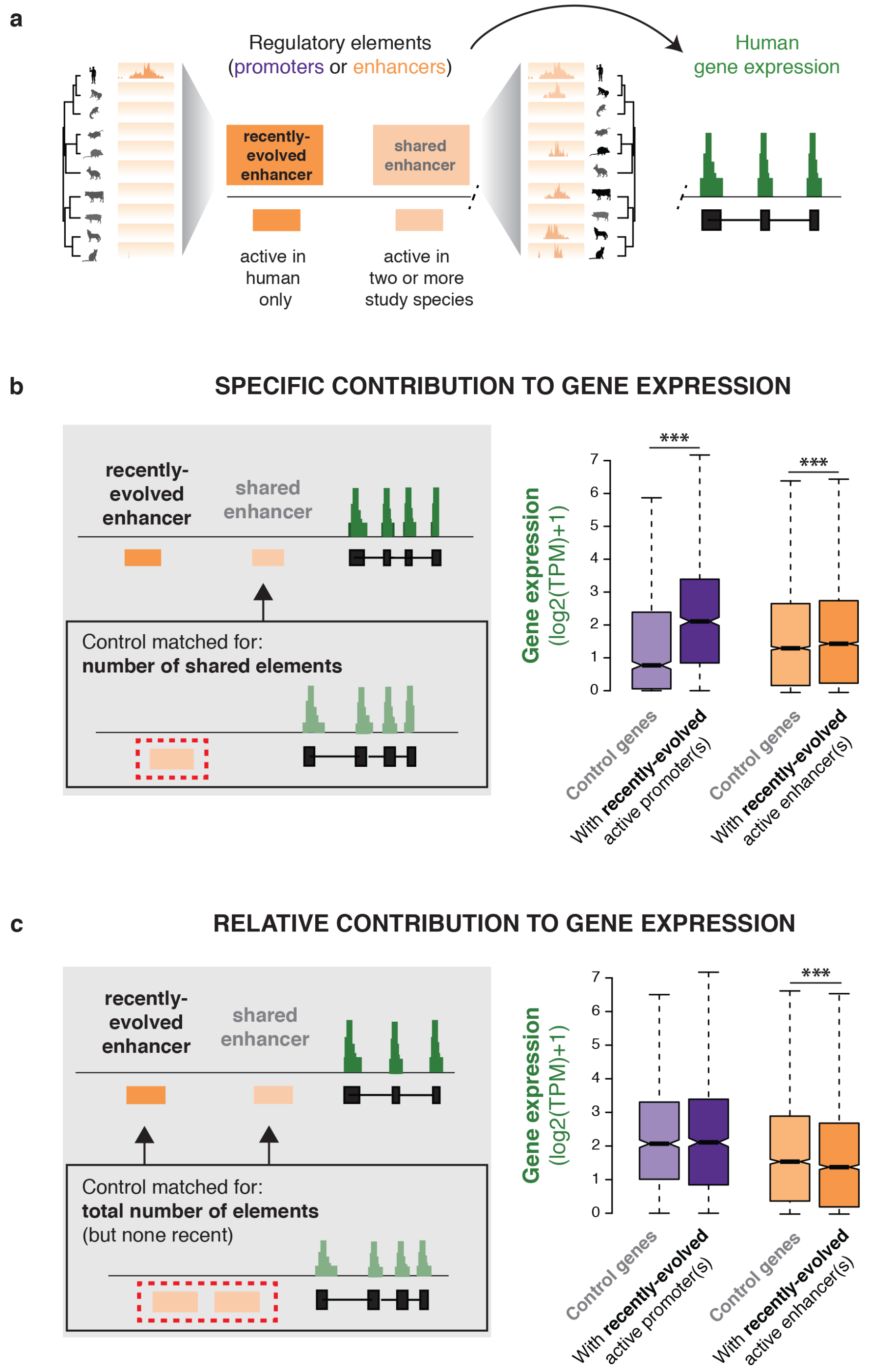

Figure 4: Recently-evolved enhancer activities weakly contribute to gene expression levels

(a) The contribution of recently-evolved regulatory elements (active in a single study species, here human) to gene expression was analysed. Genes with recently-evolved regulatory elements are typically also associated with shared regulatory elements (active in two or more species).

(b) When compared to control genes with the same number of shared regulatory elements, human genes associated with additional recently-evolved promoter(s) or enhancer(s) exhibit significantly higher expression levels (***: p < 0.001, Wilcoxon rank sum test; promoters: 995 matched genes; enhancers: 5,173 matched genes).

(c) When compared to control genes with the same total number of regulatory elements, human genes associated with recently-evolved enhancer(s) (orange) are expressed at lower levels (***: $p < 0.001$, Wilcoxon rank sum test; 3,054 matched genes). Recently evolved promoters are as active as shared ones (purple; 995 matched genes).

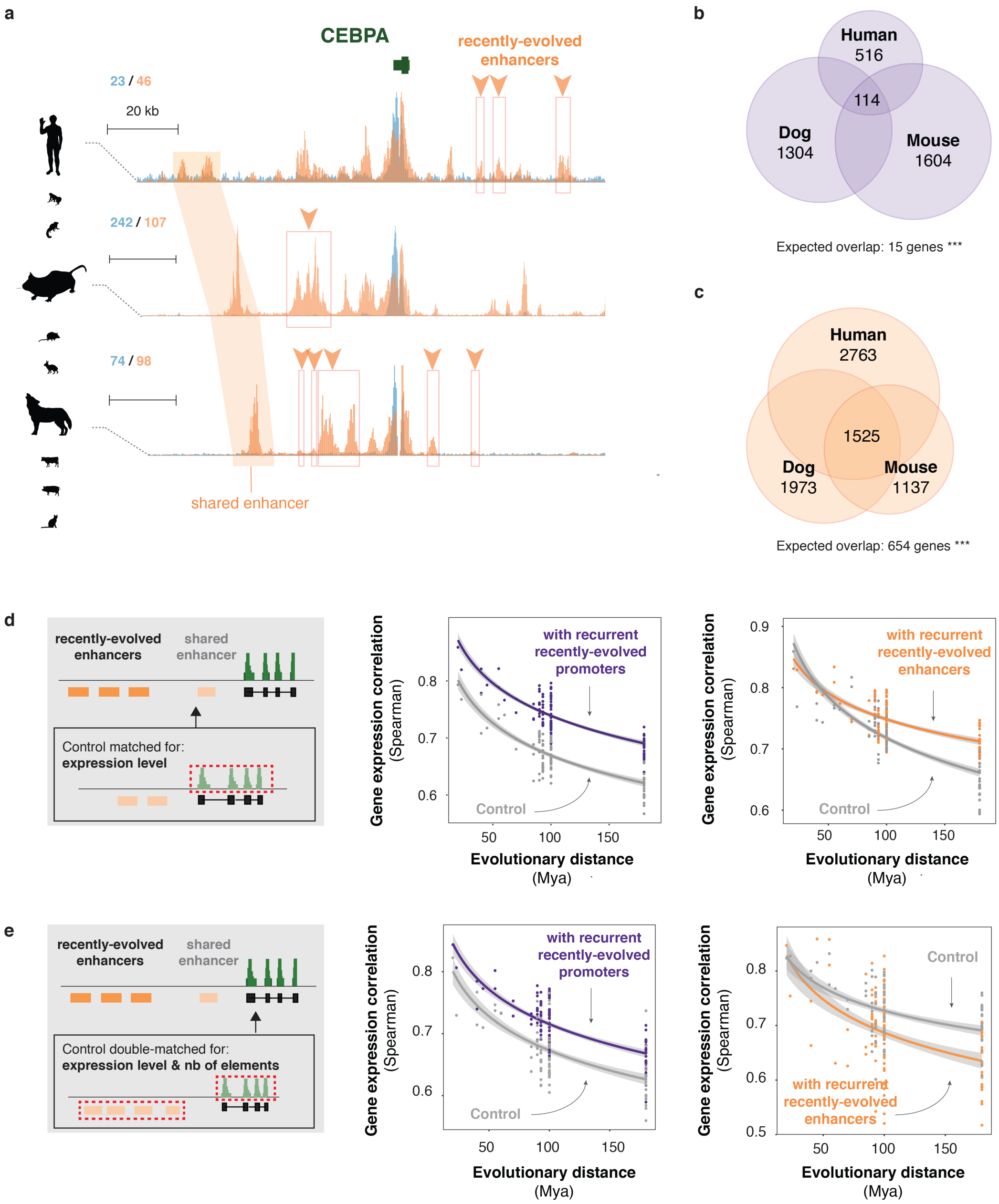

Figure 5: Recurrent recently-evolved regulatory elements contribute to gene expression stability

(a) Example of recurrent association of a gene with recently-evolved enhancers in multiple species. Genomic tracks show the regulatory landscape around the liver-specific gene *CEBPA* in human, mouse and dog (H3K4me3 (blue) and H3K27ac (orange) ChIP-seq signals; as described in Figure 2C). Recently-evolved enhancer activity in the three species is delineated with orange boxes and arrowheads. An orthologous enhancer with conserved activity across species is highlighted with orange shading.

(b-c) Genes associated with recently-evolved regulatory activity significantly overlap across three reference species (**b**: promoters; **c**: enhancers; ***: p < 0.001, Chi-squared test). Numbers in Venn diagrams correspond to the number of genes with recently-evolved elements in all three species (center) and restricted to a single species. Overlaps between pairs of species are not shown.

(d) Genes recurrently associated with recently-evolved elements across species exhibit high conservation of expression. Pairwise Spearman correlation coefficients of expression levels between species were plotted against evolutionary distance for genes recurrently associated with recently-evolved promoters (purple; 1,208 matched genes) or enhancers (orange; 729 matched genes) across multiple species, and control genes with similar mean expression levels across species. Lines are as described in Figure 1b-c.

(e) Compared to control genes with similar expression levels and regulatory complexity, genes associated with recurrent recently-acquired promoter activity in multiple species diverge more slowly in expression (purple; 1,207 matched genes). Recently-evolved enhancers however are weaker at stabilising gene expression evolution: genes recurrently associated with recently-evolved enhancers across species exhibit higher divergence than control genes with similar expression levels and number of enhancers (orange; 207 matched genes). Plots as above.

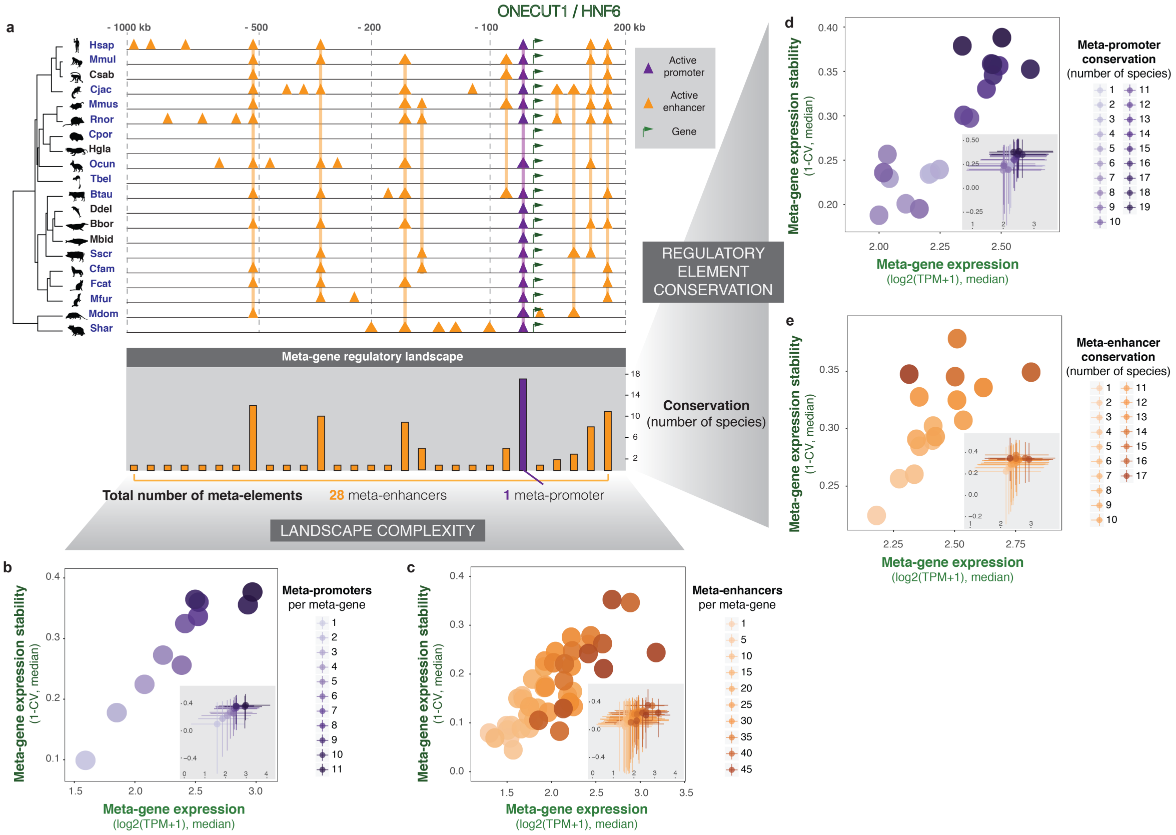

Figure 6: **An integrated summary of the evolution of mammalian regulatory complexity**

(a) Representative example of the reference-free approach to connect promoter and enhancer activity with gene expression across species. Tracks in each of twenty species show an indicative landscape of active promoters and enhancers around the ONECUT1 gene, with orthologous regions linked across species by vertical lines. The reference-free mapping of these regulatory elements across species results in a meta-gene regulatory landscape that includes a single meta-promoter and 28 meta-enhancers (bottom barplot, x-axis). For each meta-element, evolutionary conservation is recorded as the number of species where promoter or enhancer activity is detected (y-axis).

(b-c) The number of meta-promoters (**b**, purple) or meta-enhancers (**c**, orange) in the meta-gene landscape correlates with increased expression levels (x-axis) and expression stability (y-axis). Meta-genes were categorised according to the number of meta-promoters or meta-enhancers in their regulatory landscape. For each category, the median gene expression level is plotted against the median expression stability (1-CV, where CV—coefficient of variation across species). Insets in each plot show the spread of the distributions (interquartile ranges). Classes containing fewer than 30 meta-genes are not shown.

(d-e) The evolutionary conservation of meta-promoters (**d**, purple) and meta-enhancers (**e**, orange) correlates with increasingly high and stable gene expression. Individual meta-promoters or meta-enhancers were classified according to the number of species where their activity is detected, and the median expression levels and expression stability of their putative target meta-genes were plotted as above. Insets as above. Classes containing fewer than 30 meta-genes are not plotted.