Sequence properties underlying gene regulatory enhancers are conserved across mammals

Ling Chen[1,4], Alexandra E. Fish[2,4], John A. Capra[1,2,3]*

[1]*Department of Biological Sciences, Vanderbilt University, Nashville, TN, 37235, USA*

[2]*Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, 37235, USA*

[3]*Departments of Biomedical Informatics and Computer Science, Center for Structural Biology, Vanderbilt University, Nashville, TN, 37235, USA*

[4]Authors contributed equally

Correspondence: tony.capra@vanderbilt.edu

Abstract

Gene expression patterns and transcription factor DNA binding preferences are largely conserved across mammals; however, there is substantial turnover in active regulatory enhancers between closely related species. We investigated this seeming contradiction by quantifying the conservation of sequence patterns underlying histone-mark defined enhancers across six diverse mammalian species (human, macaque, mouse, dog, cow, and opossum). In each species, we found that machine-learning classifiers based on short DNA sequence patterns could accurately identify many adult liver and developing limb enhancers. We applied these classifiers across species and found that classifiers trained in different species performed nearly as well as classifiers trained on the target species, indicating that the underlying sequence properties predictive of enhancers are largely conserved. We also observed similar cross-species conservation in classifiers trained on human and mouse enhancers validated in transgenic reporter assays, and these classifiers learned predictive sequence properties similar to the classifiers trained on histone-mark defined enhancers. The sequence patterns most predictive of enhancers in each species matched the binding motifs for a common set of TFs enriched for expression in relevant tissues, which supports the biological relevance of the learned features. These results suggest that, though the genomic regions with enhancer activity change rapidly between species, many of the sequence properties encoding enhancer activity have been maintained across more than 180 million years of mammalian evolution.

Introduction

Enhancers are genomic regions distal to promoters that bind transcription factors (TFs) to regulate the dynamic spatiotemporal patterns of gene expression required for proper differentiation and development of multi-cellular organisms (Shlyueva et al. 2014; Kundaje et al. 2015). It is critical to understand the mechanisms underlying enhancer evolution and function, as alterations in their activity influence both speciation and disease (Maurano et al. 2012; Corradin and Scacheri 2014; Brazel and Vernimmen 2016). Recent genome-wide profiling of TF occupancy and histone modifications associated with enhancer activity revealed the regulatory landscape changes dramatically between species—both enhancer activity and TF occupancy at orthologous regions distal to promoters are extremely variable across closely related mammals (Taher et al. 2011; Woo and Li 2012; Cotney et al. 2013; Hsu and Ovcharenko 2013; Villar et al. 2014, 2015; Reilly et al. 2015). However, the expression of orthologous genes in similar tissues is largely conserved across mammals (Chan et al. 2009; Brawand et al. 2011; Merkin et al. 2012). Much of the gene regulatory machinery is also conserved; TFs and the short DNA motifs they bind are highly similar between human, mouse, and fly (Amoutzias et al. 2007; Wei et al. 2010; Cheng et al. 2014; Nitta et al. 2015). In short, there is considerable change in the enhancer activity of orthologous regions across mammals, despite the relative conservation of gene expression and TF binding preferences.

The rapid turnover in enhancer activity between orthologous regions in different species has largely been attributed to differences in the DNA sequences of the elements involved, rather than differences in the broader nuclear context (Wilson et al. 2008; Ritter et al. 2010; Schmidt et al. 2010; Li and Ovcharenko 2015; Prescott et al. 2015). Genome-wide profiles of TF binding have shown that 60–85% of binding differences in human, mouse, and dog for the TFs CEBPα and HNF4α can be explained by genetic variation that disrupts the binding motif (Schmidt et al. 2010). Genetic differences are also often responsible for differential enhancer activity between related species; for example, variation in TF motifs at orthologous enhancers was predictive of activity differences between human and chimp neural crest enhancers (Prescott et al. 2015). This suggests that, while there is turnover at orthologous sequences, sequence properties predictive of enhancer activity may still be conserved.

Until recently, investigation of the conservation of enhancer sequence properties across mammalian evolution has been hampered by a lack of known enhancers across diverse species within the same cellular context. The canonical definition of enhancer activity is the ability to drive expression in transgenic reporter assays (Banerji et al. 1981; Shlyueva et al. 2014), which cannot currently be scaled to assess regulatory potential

genome-wide. However, high-throughput assays such as ChIP-seq can assess histone modifications strongly associated with enhancer activity (Creyghton et al. 2010; Nord et al. 2013) to identify putative enhancers genome-wide in many tissues and species (Cotney et al. 2012; Villar et al. 2015). Using these enhancers, machine learning approaches have learned their sequence properties and successfully distinguished enhancers active in specific cellular contexts from both the genomic background and enhancers active in other tissues (Lee et al. 2011, 2015; Burzynski et al. 2012; Taher et al. 2012; Erwin et al. 2014; Ghandi et al. 2014). Recently, enhancer activity has been profiled in both adult liver (Villar et al. 2015) and developing limb (Cotney et al. 2012) in diverse mammals, which enabled us to evaluate the conservation of regulatory sequence properties.

In this study, we investigate the seeming contradiction between the rapid turnover of enhancer activity across mammals and the relative conservation of gene expression and TF binding preferences by applying machine learning classifiers to genome-wide enhancer datasets across mammals. We first show that short DNA sequence patterns can accurately identify many enhancers genome-wide in the adult liver and developing limb. Then, by using classifiers trained in one species to predict enhancers in the others, we demonstrate that enhancer sequence properties are conserved across species, even though the enhancer activity of specific loci is not. We then establish the robustness of this conservation to different enhancer identification techniques by showing that classifiers trained using human and mouse enhancer sequences validated in transgenic assays also generalize across species, and are similar to classifiers trained on histone-modification defined enhancers. Furthermore, the short DNA patterns most predictive of enhancer activity in each species matched a common set of binding motifs for TFs enriched for expression in relevant tissues. This suggests the patterns learned by classifiers capture biologically relevant sequences that influence TF binding. Together, these results argue that, though active gene regulatory sequences change substantially between species, sequence features encoding regulatory activity have been conserved over 180 million years of mammalian evolution. Our findings help explain the incongruence between the fast turnover of enhancer activity and the conservation of gene expression patterns, suggest avenues for improved cross-species enhancer identification, and establish a framework for future exploration of the conservation and divergence of regulatory sequence properties between species.

Results

Enhancers can be predicted from short DNA sequence patterns in mammals

Genome-wide enhancer activity across many mammalian species was recently assayed by profiling enhancer-associated histone modifications in the adult liver (Villar et al. 2015) and developing limb (Cotney et al. 2013). Certain chemical modifications to histones, such as acetylation of lysine 27 of histone H3 (H3K27ac) and lack of trimethylation of lysine 4 of H3 (H3K4me3), are significantly associated with active enhancers. Determining the genomic locations of these modifications via ChIP-seq provides a genome-wide proxy for the active enhancer landscape (Creyghton et al. 2010; Nord et al. 2013). For brevity, we refer to genomic regions with enhancer-associated histone modification combinations identified in these previous studies as "enhancers."

For each species and tissue, we evaluated how well short DNA sequence patterns identified enhancers. We quantified DNA sequence patterns present in each genomic region by computing its *k*-mer spectrum—the observed frequencies of all possible nucleotide substrings of length *k*. Using ten-fold cross validation, we then trained support vector machine (SVM) classifiers on the *k*-mer spectra to distinguish enhancers from random genomic regions matched to the enhancers on various attributes, such as length, GC-content, and repeat-content, as appropriate. We evaluated the performance of each classifier by computing the average area under receiver operating characteristic (auROC) and precision-recall (auPR) curves over the ten cross-validation runs (Figure 1; Methods).

We first evaluated the ability of classifiers trained on 5-mer spectra to identify liver enhancers in six representative mammals: human, macaque, mouse, cow, dog and opossum. All classifiers could distinguish active liver enhancers from length-matched background regions; auROCs ranged from 0.78 in dog to 0.84 in mouse (Figure 2a, PR curves in Figure S1a). Next, we trained 5-mer spectrum SVM classifiers to predict enhancers active in limb for human, macaque, and mouse. Again, classifiers accurately distinguished enhancers from background (Figure 2b; auROC of ~0.89 in each species, PR curves in Figure S1b). These results illustrate that SVMs trained only on DNA sequence patterns can distinguish many enhancers from background sequences across a variety of mammals for two tissues and developmental time-points.

Sequence properties predictive of enhancers are conserved across species

We then investigated whether learned DNA sequence patterns predictive of enhancer activity were conserved across mammals by testing whether classifiers trained in one species could distinguish enhancers from the genomic background in another species. First, we applied the human liver classifier to five other species with liver enhancer data: macaque, mouse, cow, dog, and opossum. We quantified cross-species performance using the relative auROC—the auROC of the enhancer classifier trained on species A and applied to species B, divided by the auROC derived from ten-fold cross validation by the classifier trained and tested on species B. In other words, the relative auROC is the proportion of within-species performance achieved by a classifier trained in a different species. The classifier trained on human liver enhancers predicted liver enhancers in other mammals nearly as accurately as classifiers trained in each species (Figure 3a-b, PR curves in Figure S1c), and its relative performance decreased only slightly across species (Figure 3b, relative auROCs > 95.5%). When expanded to all pair-wise combinations of species, classifiers accurately predicted enhancers in every mammalian species tested, regardless of the specific species they were trained in; the average relative auROC was 96.0% (Figure 3b; raw AUCs in Figure S2a-b). Classifiers generalized significantly better to more closely related species; generalization was significantly inversely correlated with the species' evolutionary divergence, as quantified by substitutions per neutrally evolving site (Figure S3, Pearson's $r = -0.585$, $P = 0.022$). The small fraction of shared enhancers between species did not drive the cross-species generalization of the classifiers (Figure S4). Furthermore, classifiers trained to identify enhancers in developing limb also accurately generalized across species (average relative auROC = 95.0%; Figure 3c-d, raw AUCs in Figure S5). The ability of classifiers to generalize to other species illustrates the conservation of sequence properties predictive of enhancers across mammals.

In addition, we investigated whether conclusions drawn from histone-modification derived enhancers generalize to enhancers identified via transgenic assays from the VISTA enhancer database. We included six tissues (limb, forebrain, midbrain, hindbrain, heart and branchial arch) with a sufficient number of enhancers in human and mouse. Consistent with the results from classifiers trained on histone-modification defined enhancers, the classifiers trained on VISTA human enhancers accurately predicted VISTA mouse enhancers of corresponding tissue from genomic background, and vice versa (Figure S6, average relative auROC = 96.3%). This suggests that sequence patterns in enhancers confirmed via reporter assays are conserved between human and mouse. Moreover, the histone-modification trained limb classifiers accurately predicted VISTA enhancers

(auROC = 0.83 in human, 0.76 in mouse) competitively with the VISTA-trained limb classifier itself (auROC = 0.81 in human, 0.78 in mouse), suggesting that sequence properties predictive of histone-modification defined enhancers are also predictive of transgenic assay validated enhancers.

Given the accurate cross-species predictions, we expected that classifiers trained on different species would produce similar regulatory potential scores for the same sequence. The SVM computes continuous scores that classify sequences as positive (enhancer) or negative (non-enhancer) based on the sequence's 5-mer spectrum. Larger absolute values of this score indicate higher confidence predictions. The scores from the human classifier applied to human enhancers were significantly positively correlated with the scores from non-human classifiers (Figure 4; Spearman's ρ between 0.90 for macaque and 0.66 for opossum). Notably, there was a weaker correlation with scores from mouse and opossum classifiers due to a subset of human enhancers with high GC-content, which received lower scores from the mouse and opossum classifiers (Figure 4b, e). Upon further investigation, we found that mouse and opossum were depleted of enhancers with high GC content, consistent with their genomic GC-content patterns (Romiguier et al. 2010), compared to other species (Figure 4f) and therefore their classifiers did not recognize high GC content enhancers present in other species. Thus, classifiers trained in different species largely produce consistent scores for a given sequence; however, this is influenced by species-level differences such as GC content and evolutionary divergence.

Overall, these results show that the DNA sequence profiles of enhancer sequences captured by species-specific 5-mer spectrum SVM classifiers are predictive of enhancers in other mammalian species in corresponding tissues. The strong generalization of performance and correlation of scores assigned to specific sequences by classifiers trained in different species indicates that sequence properties predictive of enhancers are conserved across mammals.

Short DNA sequence patterns remain predictive of enhancer activity after controlling for GC content

GC content is positively correlated with enhancer scores (Figure 4). To test if the generalization of enhancer prediction models across species was driven by the high GC content of enhancers, we trained GC-controlled classifiers using negative sets of random genomic regions matched on GC content. The predictive power of the GC-controlled classifiers was substantial (average auROC of 0.75 for liver and 0.79 for limb; Figures S7a and S8a), but as expected, less than the corresponding classifiers without GC-control (average auROC of 0.81 for liver

and 0.89 for limb; Figure 2). Nevertheless, GC-controlled classifiers maintained strong cross-species generalization: liver classifiers had an average relative auROC of 94.8% when applied to the other five species (Figures 5a and S7); limb classifiers had an average relative auROC of 95.0% when applied across species (Figures S8e). The enhancer scores given to individual sequences by the GC-controlled classifiers were significantly correlated, and as expected, high GC-content sequences no longer received consistently high scores (Figure S9). Ultimately, the strong performance of the GC-controlled classifiers suggests that enhancers differ from the genomic background in higher order sequence patterns beyond GC-content, and that those patterns are conserved.

The generalization of each species' GC-controlled classifier had the same pattern as the classifiers without GC-control: the human classifier had the best generalization (average relative auROC = 96.1%), while the opossum had the worst (average relative auROC = 92.8%). In these GC-controlled analyses, we observed a stronger inverse correlation between the relative performance across species and sequence divergence (Figure S10, Pearson's $r = -0.77$, $P = 0.001$) than in the non-GC-controlled analysis (Figure S4, Pearson's $r = -0.585$, $P = 0.022$). This indicates that genomic differences in GC content distribution and overall evolutionary divergence influence the conservation of the sequence patterns predictive of putative enhancers.

Short DNA sequence patterns remain predictive of enhancer activity after controlling for repetitive elements

Approximately half of the base pairs in most mammalian genomes are comprised of repetitive sequence elements derived from transposable elements with characteristic sequence patterns or simpler low complexity repeats. Repetitive elements can contribute to the birth of new regulatory elements (Rebollo et al. 2012; Chuong et al. 2013; Su et al. 2014), as their sequence patterns often include TF binding motifs. To evaluate the influence of repetitive elements on the ability to distinguish enhancers from the background and the observed conservation of sequence properties across species, we trained classifiers to distinguish enhancers that did not overlap a repeat element (only 3.3% of all enhancers in human) from matched non-repetitive regions from the genomic background. Neither the ability to distinguish enhancers from the background in a species, nor the ability of predictive sequence properties to generalize across species, was substantially reduced (Figure S11). This demonstrates that, while repetitive elements contribute to enhancer activity, the conservation of sequence properties predictive of enhancers is not contingent on their presence.

To examine the influence of repetitive elements across all observed enhancer sequences, we also trained classifiers to distinguish all enhancers regions from genomic background regions matched for both GC-content and the proportion of overlap with a repeat element. The performance of these classifiers decreased (average auROC of 0.73; Figure S12a) relative to when not controlling for repeat overlap (average auROC of 0.75; Figure S7a) or neither repeats or GC-content (average auROC of 0.81; Figure 2). This indicates that both features play a role in enhancer function. Most importantly though, the repeat and GC-controlled classifiers still generalized across species (average relative auROC = 94.0%, Figures 5b and S12), which demonstrates that enhancer sequence properties beyond both GC and repeat content are conserved across species.

Enhancer sequence properties are more similar across the same tissue in different species than across different tissues in the same species

Gene expression patterns are significantly more similar in corresponding tissues across species than between different tissues in the same species (Chan et al. 2009; Brawand et al. 2011; Merkin et al. 2012). We demonstrated that enhancer sequence properties are strongly conserved in the same tissue across species (Figure 2). We hypothesized that, as for gene expression, enhancer sequence properties would be more similar in the same tissue across species (cross-species) than between different tissues in the same species (cross-tissue). To test this, we performed two cross-tissue analyses. The first examined the cross-tissue performance of the liver (Villar et al. 2015) and limb (Cotney et al. 2013) classifiers over the three species with enhancers in both datasets: human, macaque, and mouse. For each species, the liver-trained classifier was applied to that species' limb enhancers, and vice versa. Cross-species performance (all pairwise relative auROCs) was significantly higher than cross-tissue performance (Figure 6). This held for both GC-controlled and non-GC-controlled classifiers. To evaluate this across a broader array of cellular contexts, we performed a second cross-tissue analysis using human enhancers identified in 11 diverse cellular contexts by the Roadmap Epigenomics Project (Kundaje et al. 2015) (Methods). We applied the human liver classifier to these contexts, and again observed that cross-species performance was significantly higher than the cross-tissue performance (Figure 6).

The ability of enhancers to regulate gene expression is often contingent on both cell-type specific attributes, such as expression patterns of TFs (Vaquerizas et al. 2009), and properties that are shared across active enhancers in general. The stronger performance of the trained classifiers in the cross-species compared to cross-

tissue prediction tasks demonstrates that they capture cell-type-specific sequence attributes and that these features are conserved across species.

The most predictive 5-mers in different species match binding motifs for many of the same transcription factors

To interpret the biological relevance of the sequence patterns learned by the trained enhancer prediction models in each species, we analyzed the similarity of the sequence properties in their functional context: TF binding motifs. We matched the 5% (n = 52) most enhancer-associated 5-mers learned by the human GC-controlled liver classifier to a database of 205 known TF motifs (Mathelier et al. 2014) using TOMTOM (Figure 7a). The enhancer-associated 5-mers were significantly more likely to match a TF (46.1% matched at least one TF; one-tailed $P = 0.0035$, binomial test) than expected at random (27.7%). The 5% (n=52) most background-associated 5-mers were not significantly different from random (21.6% matched at least one TF, two-tailed $P = 0.43$, binomial test). This illustrates that the classifiers learned sequence patterns with regulatory potential.

Next, we investigated whether the TF binding motifs matched by enhancer-associated 5-mers were shared between species. For each species' GC-controlled liver classifier, we matched the top 5% of enhancer-associated 5-mers to TF motifs. The highly weighted 5-mers in the human-trained classifier matched 121 TF motifs. Of these, the binding motifs for 33 TF were also matched by enhancer-associated 5-mers in all other species (Figure 7b, Supplementary Table 1). This is significant enrichment for shared TF motifs among the enhancer-associated 5-mers; only 0.59 TF motifs were shared across all species on average over 100 random sets of 5% of 5-mers from each species (Figure 7c). Similarly, only one TF motif (MZF1) was shared among all the species' most background-associated 5-mers. The GC-controlled limb classifiers also shared more TFs among the top 5% of enhancer-associated 5-mers than expected from random sets (12 vs. 8.78, respectively). However, it is likely that the smaller number of available species for developing limb enhancers, our limited knowledge of binding motifs for TFs active in developing limb and the heterogeneity of developing limb tissue reduced power to detect sharing compared to liver. We obtained similar results when comparing the TFs matched by 5-mers from non-GC-controlled SVM models (Figure S13).

To evaluate the relevance of the shared TF motifs to liver function, we evaluated the expression patterns of the TFs across 12 tissues (Bernstein et al. 2012). Shared TFs among liver enhancer-associated 5-mers were significantly enriched for liver expression (Table 1, $P = 0.011$, one-tailed Fisher's exact test). Many of the shared

TFs play an essential role in liver function. For instance, they are enriched for activity in the TGF-β signaling pathway compared to non-shared TFs; the enrichment is mainly due to members of the AP-1 (JUN, FOS, and MAF subfamilies) and SMAD families (Methods) (The Gene Ontology Consortium 2000, 2015). TGF-β signaling is a central regulatory mechanism that is disrupted in all stages of chronic liver disease (Dooley and ten Dijke 2012). Further, mice deficient in c-JUN or MAF have an embryonic lethal liver phenotype (Eferl et al. 1999; Yamazaki et al. 2012). This demonstrates that the sequence patterns learned in each species capture similar motifs that are recognized by TFs that important to the relevant tissue context.

Discussion

In this study, we demonstrated that machine-learning classifiers based on short DNA sequence patterns could distinguish many liver and limb enhancers from the genomic background in diverse mammalian species. We then showed that, in spite of significant changes in the enhancer landscape between species, sequence-based enhancer prediction models exhibited minimal decreases in performance when applied across species. This indicates that sequence properties predictive of enhancer activity captured by these models are conserved across mammals. Furthermore, the DNA patterns most predictive of activity across species matched a common set of TF binding motifs with enrichment for expression in the relevant tissues. Also, sequence properties predictive of histone-mark defined enhancers were also predictive of enhancers confirmed in transgenic reporter assays. These results suggest the presence of conserved regulatory mechanisms that have maintained sequence constraints on enhancers for more than 180 million years.

Confidently identifying and experimentally validating enhancers remains challenging. We showed that enhancer sequence properties are conserved across species using enhancers identified via two complementary techniques: histone modification profiling and transgenic assays. Each of these approaches has strengths and weaknesses. The histone modification based enhancer predictions enable genome-wide characterizations across many species, but this approach is prone to false positives. On the other hand, the transgenic assays clearly demonstrate the competence of a sequence to drive gene expression, but are restricted to relatively few sequences from two species that are tested at one developmental stage. By showing the cross-species conservation is maintained in both categories, and that models trained on each set perform similarly, we argue the conservation of enhancer sequence properties is robustness to the methodology used to define enhancers.

The design of this study will serve as a useful framework for further examining the conservation of regulatory sequence patterns across species. We trained sequence-based machine learning models within a species, and then applied them to other species; this approach can be applied on a genome-wide scale, and is not dependent on knowledge of TF binding motifs, and allows some flexibility in the weights assigned to each feature while directly testing the generalization of overall sequence patterns. Identification of enhancers in more divergent species would enable us better quantify how deeply conserved enhancer sequence properties are. This remains an open question, as more divergent animal species have very little conservation of TF co-associations at putative enhancers (Boyle et al. 2014) despite conservation of TF binding preferences. Identification of enhancers

in the same cellular context for more closely related species would enable the investigation of lineage-specific regulatory sequence patterns. Thus, additional comparative studies of regulatory sequence features in more species are needed to better understand both recent and ancient influences on regulatory sequences.

While the classifiers correctly distinguished many enhancers from the genomic background, the performance was not perfect. Many factors contribute to this, including: false positives in the training data, noise from the low resolution of the histone modification peaks (i.e., they include non-functional sequence flanking the enhancer), and the nature of the sequence features in our models. Additionally, our models only evaluated properties captured by *k*-mer spectra; however, our framework could be expanded to quantify the conservation of higher-order sequence patterns such as *k*-mer spacing, order, combinations, and hierarchies through the application of more flexible sequence models (Ghandi et al. 2014; Zhou and Troyanskaya 2015). The framework could also be adapted to investigated other functionally relevant factors, such as histone modifications and DNA shape (Slattery et al. 2014; Villar et al. 2014).

We demonstrated that short DNA sequence patterns can distinguish many active enhancers from the genomic background in diverse mammalian species, and that the predictive DNA patterns learned in one species generalize very well to other mammals. The commonality of sequence elements predictive of enhancer activity across mammals suggests that much of what we learn about enhancer biology, particularly at the basic sequence level, in model organisms could be extrapolated to humans. Sequence-based cross-species enhancer prediction could be of particular use in studying difficult to obtain human tissues and providing preliminary annotations in uncharacterized species and tissues. Furthermore, there is the potential to combine sequence-based models with successful cross-species enhancer prediction strategies based on functional genomics data (Capra 2015). Nonetheless, much work remains to understand how regulatory and evolutionary processes have interacted to produce a conserved core of sequence features in mammalian enhancer DNA sequences and to understand the functional signatures encoded in them.

Methods

Genomic data

All work presented in this paper is based on hg19, rheMac2, mm10 (mouse liver dataset), mm9 (mouse limb dataset), bosTau6, canFam3 and monDom5 DNA sequence data from the UCSC Genome Browser. All mm9 regions were mapped to mm10 using the *liftOver* tool from the UCSC Kent tools (Kuhn et al. 2013). Liver gene annotations are from Ensembl v73; limb gene annotations are from Ensembl v67 (Flicek et al. 2014). The sequence divergence between each pair of species was computed from the neutral model built from fourfold degenerate sites in the 100-way multiple species alignment from UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/).

Enhancer and random genomic background datasets

We used two multi-species histone-modification-defined enhancer datasets in this study. The first consisted of liver enhancers identified by genome-wide ChIP-seq profiling of histone modifications (H3K27ac without H3K4me3) in 20 species from five mammalian orders, including human, macaque, mouse, cow, dog, and opossum (Villar et al. 2015). The second dataset contained human, macaque, and mouse enhancers identified from profiling the H3K27ac modification in developing limb tissue (Cotney et al. 2013). For limb enhancers, we excluded regions in exons or within 1 kb of a transcription start site. For each species, we combined the enhancer regions from different limb development stages.

In the training and evaluation of the classifiers (see next section), the active enhancers in each species were the positive training examples. For non-GC-controlled analyses, the negative training examples were 10 sets of length and chromosome matched random genomic regions. In the GC-controlled analyses, the negative training examples were 10 sets of length, chromosome, and GC-content matched random regions from the genomic background. For the repeat controlled analysis, we obtained the repetitive elements coordinated from RepeatMasker (Smit et al. 2013) and generated 10 sets of random regions from the genomic background matched on length, chromosome, GC-content, and proportion overlap with repetitive elements as negative training examples. For all analyses, we did not consider enhancers or random regions that fell in genome assembly gaps (UCSC gap track). For human and mouse, we also excluded the ENCODE blacklist regions (Bernstein et al. 2012) (https://sites.google.com/site/anshulkundaje/projects/blacklists). For within-species classification, we

performed ten-fold cross validation and averaged the auROC and auPR over the ten runs. For cross-species classification, we trained on the whole dataset in the training species and evaluated the performance on a random half of the dataset in the test species due to computational limitations of the kebabs package.

In addition to the histone-modification-defined enhancers, we also analyzed enhancers validated in transgenic reporter assays in embryonic day 11.5 mouse embryos from VISTA (Visel et al. 2007). We investigated all six tissues with at least 50 positive enhancer elements in both species: forebrain, midbrain, hindbrain, limb, heart and branchial arch. These enhancers comprised the positive training examples. For each tissue, we additionally generated 10 sets of length and chromosome matched random genomic as negative training examples.

To determine how well classifiers generalized across tissue types in addition to limb and liver, we used human enhancers identified by the Roadmap Epigenomics Project (Kundaje et al. 2015) in ten tissues from diverse body systems: hippocampus middle (brain, E071), pancreas (exocrine-endocrine, E098), gastric (GI, E094), left ventricle (heart, E095), lung (E096), ovary (reproductive, E097), bone marrow derived mesenchymal stem cell cultured cells (stromal-connective, E026) and CD14 primary cells (white blood, E029). We defined enhancers in these tissues as H3K27ac without H3K4me3 regions. For each tissue, we generated 10 sets of not-GC-controlled and GC-controlled negative training examples as above.

Spectrum kernel SVM classification

A support vector machine (SVM) is a discriminative classifier that learns a hyperplane to separate the positive and negative training data in feature space. We used *k*-mer spectrum kernel to quantify sequence features for the SVM (Leslie et al. 2002); reverse complements of *k*-mers are different *k*-mers in this study. Binary classification, evaluation, and calculation of feature weights were performed with the kebabs R package (Palme et al. 2015). We report all analyses with $k = 5$, but classifier performance and generalization were similar for $k = 4–7$ (Supplementary note). More flexible models, such as the mismatch (Leslie et al. 2002; Palme et al. 2015) and gappy pair kernels (Mahrenholz et al. 2011; Bodenhofer et al. 2009), did not significantly increase the performance (Supplementary note).

Transcription factor motif analysis

5-mers were matched to known TF binding motifs in the JASPAR 2014 Core vertebrate database (Mathelier et al. 2014) using the TOMTOM package (Gupta et al. 2007) with default parameters. The sharing of 5-mers and TFs across species was visualized using jVenn (Bardou et al. 2014).

Transcription factor expression data

We obtained RNA-seq data for TFs across 12 tissues from the Gene Expression Atlas (https://expressionatlas.org/hg19/adult/). The FPKM (Fragments Per Kilobase of transcript per Million mapped reads) was converted to binary expressed/non-expressed calls (non-zero reads were converted to 1).

Availability

Source code for cross-species enhancer prediction and generating all results presented here is freely available at: https://github.com/lingchen42/EnhancerCodeConservation.

Acknowledgements

We thank D. Kostka, D. Rinker, and C. Simonti for helpful discussions and comments on the manuscript. We thank P. Flicek for clarifications about the liver enhancer data. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. This work was supported in part by US National Institutes of Health (NIH) grant (1R01GM115836 to J.A.C) and an Innovation Catalyst Award from the March of Dimes Prematurity Research Center Ohio Collaborative.

Author Contributions

J.A.C. conceived and supervised the project. L.C. collected the data and led the analyses. A.E.F. contributed analyses. All authors interpreted the data and wrote the manuscript.

Figures

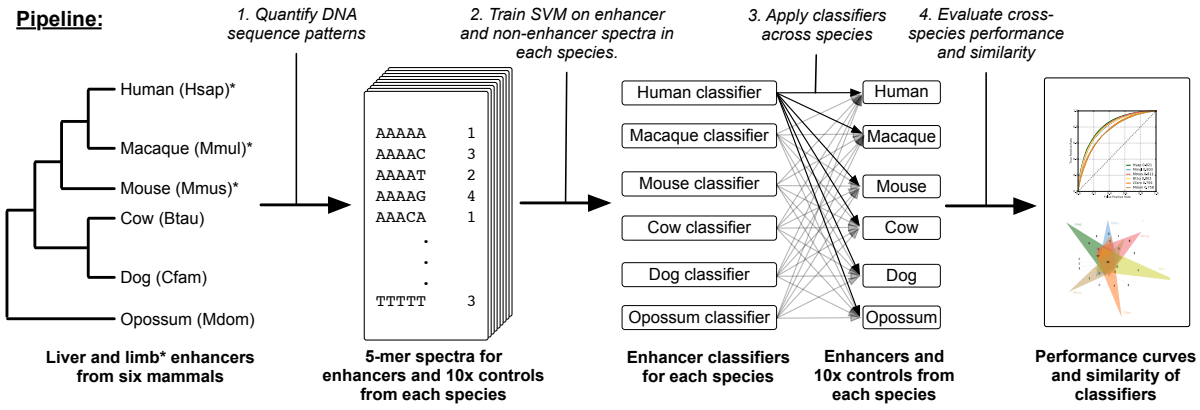

Figure 1. Overview of the framework for evaluating DNA patterns predictive of enhancer activity across diverse mammals. Starting with liver and limb enhancers and ten sets of non-enhancer DNA sequences from six mammals, the first step of the pipeline quantified each of these genomic regions by their 5-mer spectrum—the frequency of occurrence of all possible length five DNA sequence patterns. Using the spectra as features, we trained a spectrum kernel support vector machine (SVM) to distinguish enhancers from non-enhancers in each species and evaluated their performance with ten-fold cross validation. Then, we applied classifiers trained on one species to predict enhancer activity in all other species. Finally, we evaluated the performance of cross-species prediction compared to within species prediction and compared the most predictive features in classifiers from different species. Limb enhancer data were only available for human, macaque, and mouse.

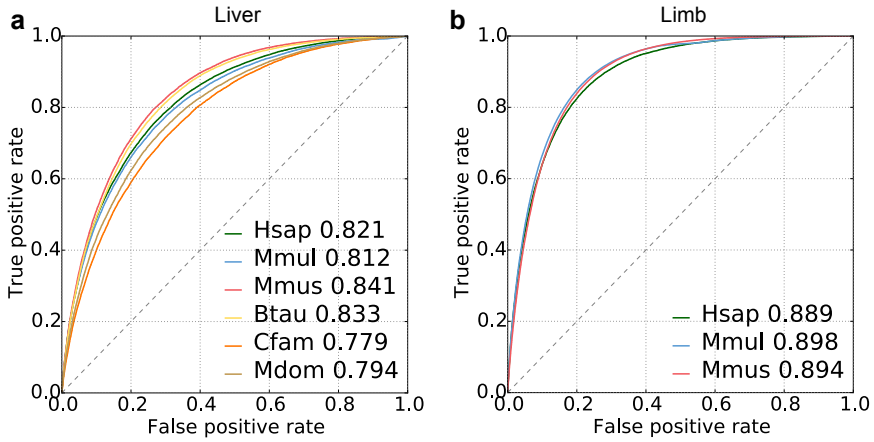

Figure 2. Performance of DNA sequence-based enhancer identification in diverse mammals. (a) ROC curves for classification of liver enhancers vs. the genomic background in six diverse mammals: human (Hsap), macaque (Mmul), mouse (Mmus), cow (Btau), dog (Cfam), and opossum (Mdom). (b) ROC curves for classification of developing limb enhancers in human, macaque, and mouse. Area under the curve (AUC) values are given after the species name. Ten-fold cross validation was used to generate all ROC and PR curves (Figure S1a, b).

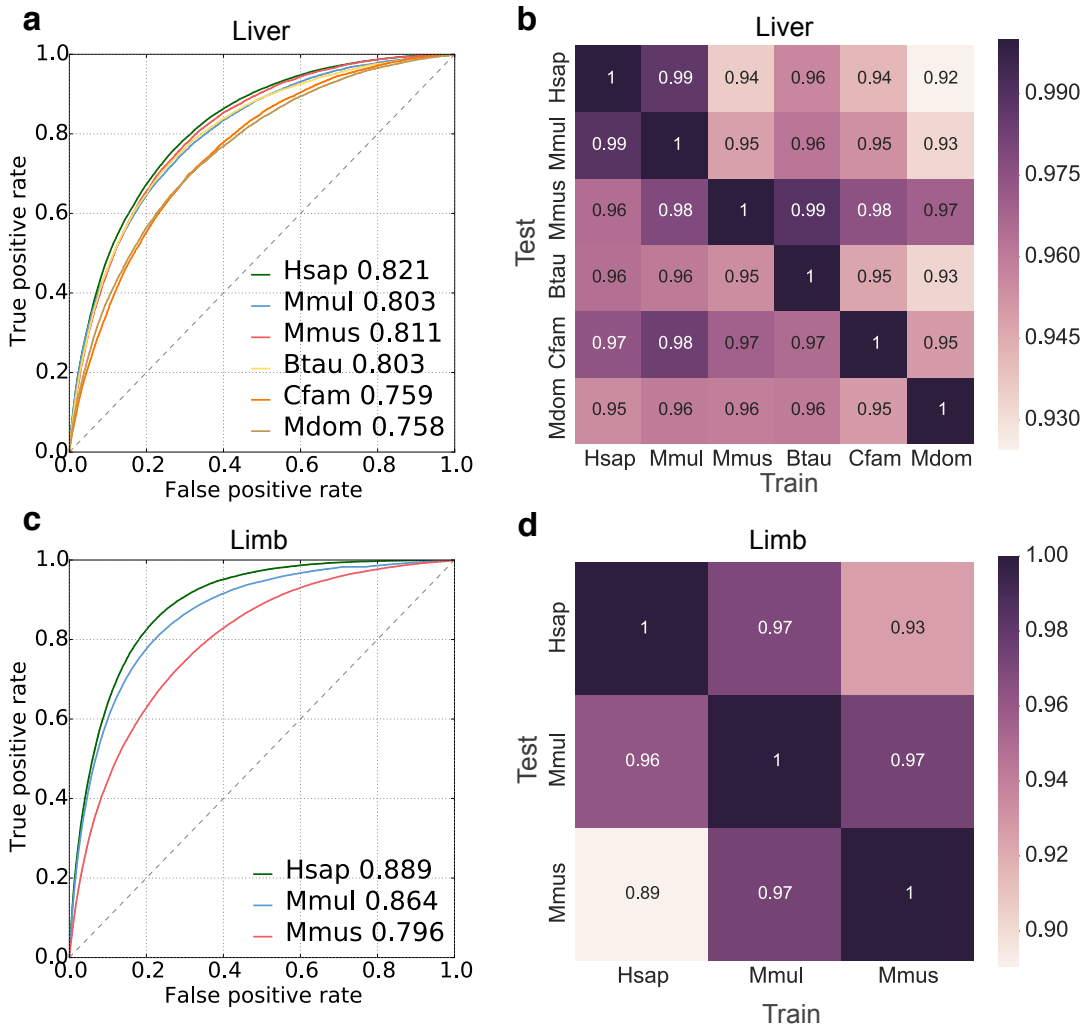

Figure 3. Human-trained enhancer classifiers accurately predicted liver and limb enhancers in diverse mammals. (a) ROC curves of the performance of the human liver enhancer classifier applied to the human (Hsap), macaque (Mmul), mouse (Mmus), cow (Btau), dog (Cfam) and opossum (Mdom) datasets. Area under the curve (auROC) values are given after the species name. (b) Heat map showing the relative auROC of liver enhancer classifiers applied across species compared to the performance of classifiers trained and evaluated on the same species (Figure 2a). The classifiers were trained on the species listed on the x-axis and tested on species on the y-axis. (c) ROC curves showing the performance of the human limb enhancer classifier on human, macaque and mouse. (d) Heat map showing the relative auROC of limb enhancer classifiers applied across species compared to the performance of classifiers trained and evaluated on the same species (Figure 2b). The raw auROC and auPR values for all comparisons are given in Figure S3 and Figure S4.

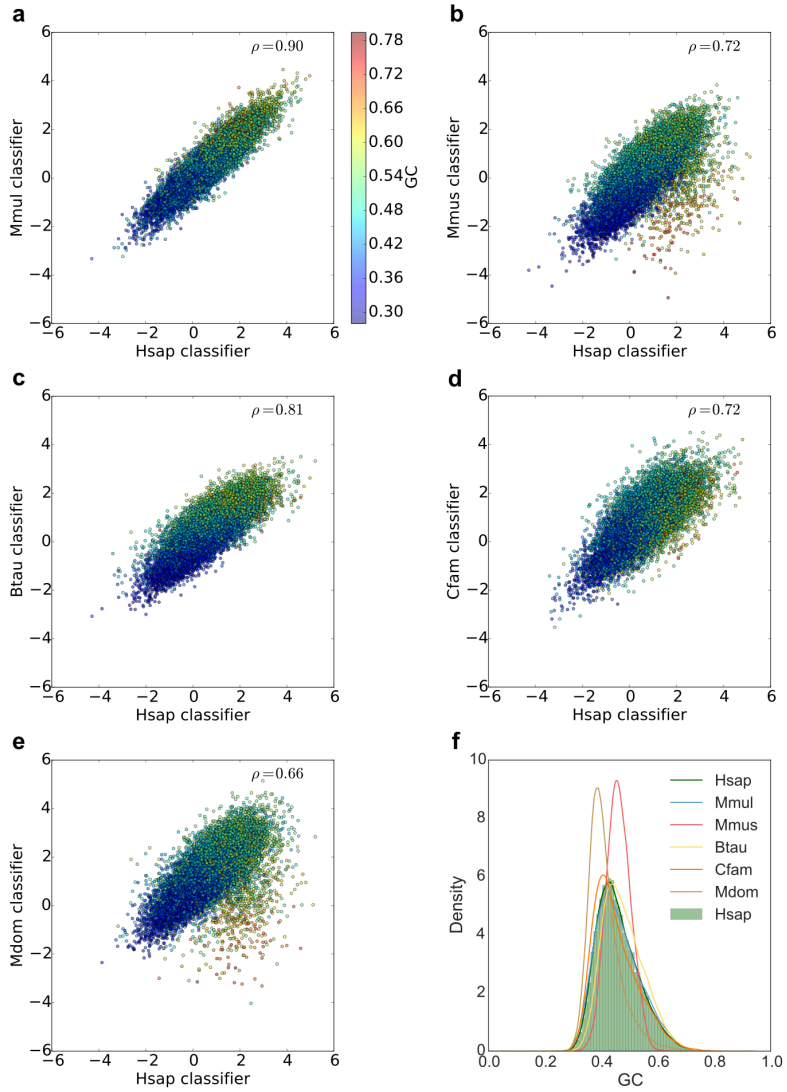

Figure 4. The predictions of enhancer classifiers trained in different species were strongly correlated. Scatter plots showing the correlation between scores assigned to human enhancers by the human-trained classifier and the classifiers trained on other species: (a) Human vs. Macaque. (b) Human (Hsap) vs. Mouse (Mmus) (c) Human vs. Cow (Btau) (d) Human vs. Dog (Cfam) (e) Human vs. Opossum (Mdom). Each dot represents a human liver enhancer sequence. The enhancer score assigned by the human-trained classifier is plotted on the x-axis, and the score assigned by the classifier trained on the other specified species is plotted on the y-axis. The color indicates the GC content. Correlation is quantified by Spearman's rank correlation coefficient (ρ). (f) The GC content distribution of liver enhancers in human, macaque, mouse, cow, dog, and opossum. Human, macaque, cow and dog enhancers have a similar GC distribution. Mouse and opossum have less variation in GC content and are depleted of high GC enhancers compared to the other species.

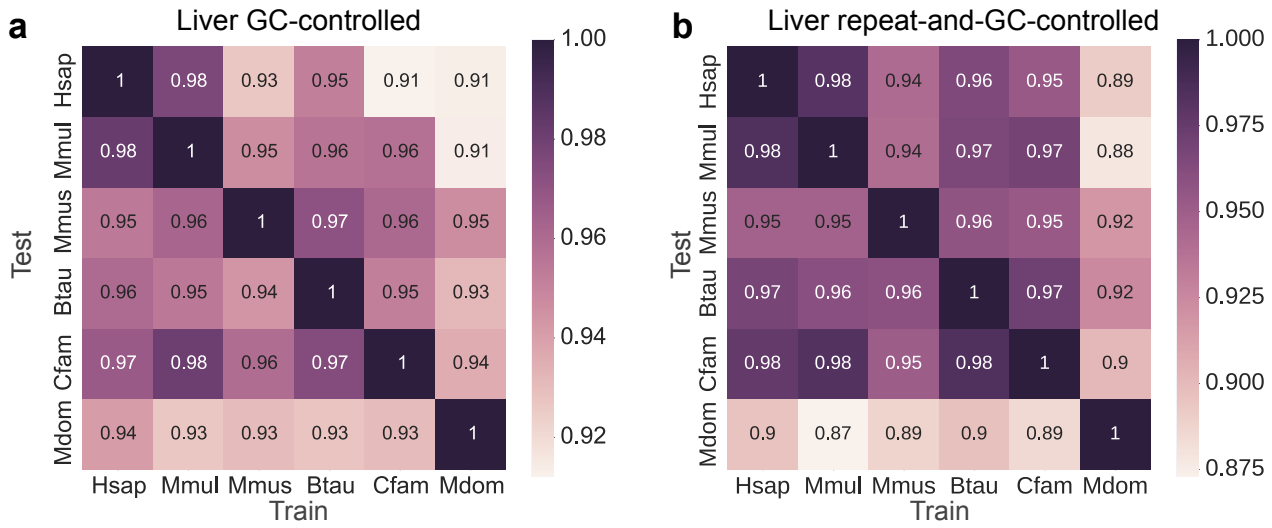

Figure 5. Enhancer sequence properties remain conserved across diverse mammals after controlling for both GC-content and repetitive elements. The heat maps give the cross-species relative auROCs for SVM classifiers trained on 5-mer spectra to identify enhancers in the species along the x-axis, and then used to predict enhancers in the species on the y-axis. The "negative" training regions from the genomic background were matched to the enhancers': (a) GC-content, and (b) GC-content and proportion overlap with repetitive elements.

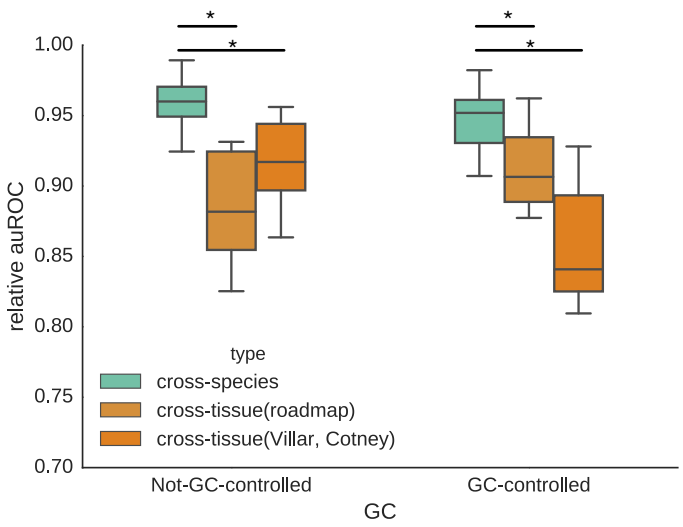

Figure 6. Enhancer classifiers generalize more accurately across the same tissue in different species than across different tissues in the same species. In the not-GC-controlled analysis, the cross-species performance (average relative auROC = 95.8%) is significantly better than the cross-tissue (roadmap) performance (88.4%, two-tailed t test, $P = 0.001$) and the cross-tissue (Villar, Cotney) performance (91.6%, two-tailed t test, $P = 0.03$). This holds true for the GC-controlled analysis. The cross-species performance (average relative auROC = 94.6%) is significantly better than the cross-tissue (roadmap) performance (91.2%, two-tailed t test, $P = 0.02$) and the cross-tissue (Villar, Cotney) performance (85.8%, two-tailed t test, $P = 0.04$).

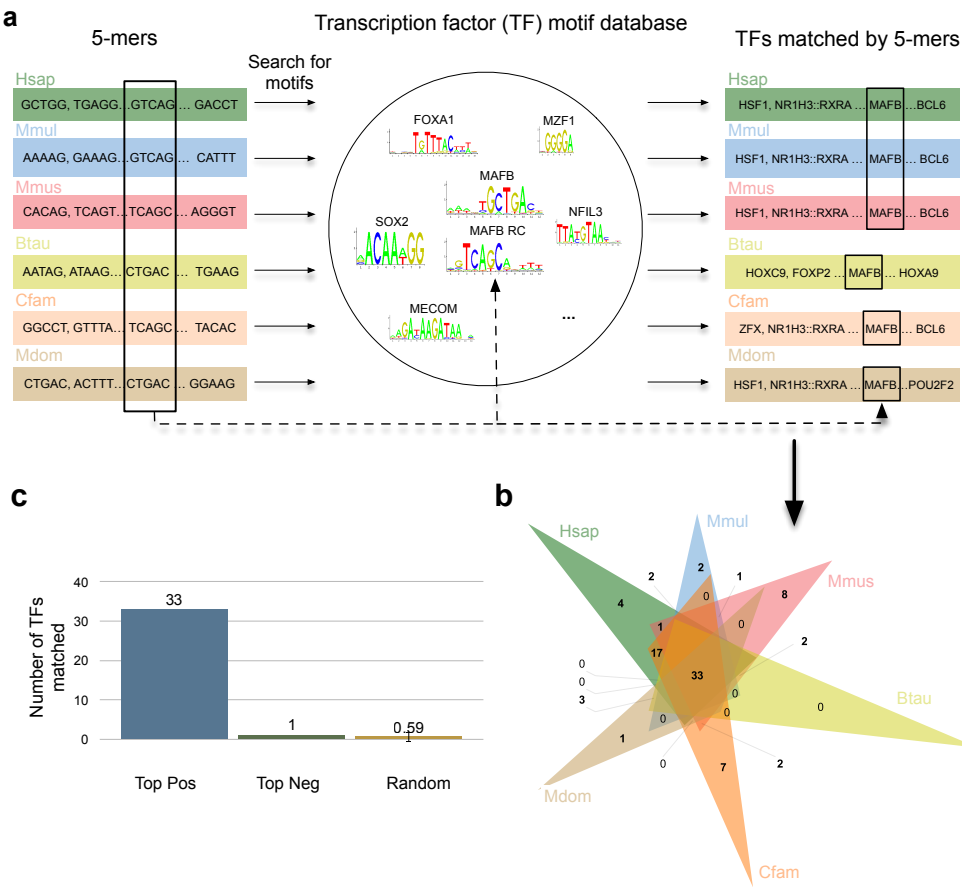

Figure 7. The DNA sequence patterns most predictive of liver activity across species matched a common set of transcription factors. (a) Transcription factor analysis workflow. For each species enhancer classifier, we found TF motifs matched by the top 5% positively weighted 5-mers. Note that different 5-mers (marked with black box on the left) can match the same motif, e.g., MAFB and its reverse complement (RC). The overlap of 5-mers and matched TFs were then compared across each species' classifier. (b) Venn diagram of the sharing of the TF motifs matched by the top 5% positive 5-mers from each GC-controlled liver classifier. The total number of TFs matched by top 5-mers in each species was: 121 (human), 104 (macaque), 100 (mouse), 81 (cow), 118 (dog), 102 (opossum). Similar results were observed for the non-GC-controlled classifier (Figure S15a). (c) The number of TFs matched by all species based on 5-mers in top positive, top negative, and 100 random sets of 5% of all possible 5-mers. The 33 TF motifs shared among the high-weight set for each species is thus significantly more than expected.

Table 1. The TFs with motifs shared among the top 5-mers across all species are significantly enriched for liver expression ($P = 0.011$, one-tailed Fisher's exact test).

	Shared TFs	*Not shared TFs*
Liver expressed	26	89
Not liver expressed	7	70
Percent Liver expressed	78.8%	56.0%

References

Amoutzias GD, Veron a. S, Weiner J, Robinson-Rechavi M, Bornberg-Bauer E, Oliver SG, Robertson DL. 2007. One billion years of bZIP transcription factor evolution: Conservation and change in dimerization and DNA-binding site specificity. *Mol Biol Evol* **24**: 827–835.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a ??-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308.

Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. 2014. jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* **15**: 293.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L. 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**: 453–456.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Brazel AJ, Vernimmen D. 2016. The complexity of epigenetic diseases. *J Pathol* **238**: 333–344.

Burzynski GM, Reed X, Taher L, Stine ZE, Matsui T, Ovcharenko I, McCallion AS. 2012. Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control. *Genome Res* **22**: 2278–2289.

Capra JA. 2015. Extrapolating histone marks across developmental stages, tissues, and species: an enhancer prediction case study. *BMC Genomics* **16**: 104.

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M. 2009. Conservation of core gene expression in vertebrate tissues. *J Biol* **8**: 33.

Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* **515**: 371–375.

Chuong EB, Rumi MAK, Soares MJ, Baker JC. 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet* **45**: 325–329. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3789077&tool=pmcentrez&rendertype=abstract.

Corradin O, Scacheri PC. 2014. Enhancer variants: evaluating functions in common disease. *Genome Med* **6**: 85.

Cotney J, Leng J, Oh S, DeMare LE, Reilly SK, Gerstein MB, Noonan JP. 2012. Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res* **22**: 1069–1080.

Cotney J, Leng J, Yin J, Reilly SK, Demare LE, Emera D, Ayoub AE, Rakic P, Noonan JP. 2013. XThe evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154**: 185–196. http://dx.doi.org/10.1016/j.cell.2013.05.056.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**: 21931–21936.

Dooley S, ten Dijke P. 2012. TGF-β in progression of liver disease. *Cell Tissue Res* **347**: 245–56.

Eferl R, Sibilia M, Hilberg F, Fuchsbichler A, Kufferath I, Guertl B, Zenz R, Wagner EF, Zatloukal K. 1999. Functions of c-Jun in liver and heart development. *J Cell Biol* **145**: 1049–1061.

Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA. 2014. Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput Biol* **10**: e1003677.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res* **42**: 749–755.

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**: e1003711.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8**: R24. http://genomebiology.com/2007/8/2/R24.

Hsu C-H, Ovcharenko I. 2013. Effects of gene regulatory reprogramming on gene expression in human and mouse developing hearts. *Philos Trans R Soc Lond B Biol Sci* **368**: 20120366.

Kuhn RM, Haussler D, James Kent W. 2013. The UCSC genome browser and associated tools. *Brief Bioinform* **14**: 144–161.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–329.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47**: 955–61.

Lee D, Karchin R, Beer MA. 2011. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* **21**: 2167–2180.

Leslie C, Eskin E, Noble WS. 2002. The spectrum kernel: a string kernel for SVM protein classification. *Pac Symp Biocomput* **575**: 564–575.

Li S, Ovcharenko I. 2015. Human enhancers are fragile and prone to deactivating mutations. *Mol Biol Evol* **32**: 2161–2180.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C, Chou A, Ienasescu H, et al. 2014. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**: D142--7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965086&tool=pmcentrez&rendertype=abstract.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-)* **337**: 1190–1195.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science (80-)* **338**: 1593–1599.

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, et al. 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **4**: 1–20. http://elifesciences.org/content/4/e04837.

Nord AS, Blow MJ, Attanasio C, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry M, Afzal V, et al. 2013. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**: 1521–1531.

Palme J, Hochreiter S, Bodenhofer U. 2015. KeBABS: an R package for kernel-based analysis of biological sequences. *Bioinformatics* 1–3.

Prescott SL, Srinivasan R, Marchetto MC, Gage FH, Swigut T, Selleri L, Gage FH, Swigut T, Wysocka J. 2015. Enhancer Divergence and cis -Regulatory Evolution in the Human and Chimp Neural Crest Article Enhancer Divergence and cis -Regulatory Evolution in the Human and Chimp Neural Crest. 68–83.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**: 21–42.

Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP. 2015. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science (80-)* **347**: 1155–1159. http://www.sciencemag.org/content/347/6226/1155.abstract%5Cnhttp://www.sciencemag.org/content/347/6226/1155.full.pdf.

Ritter DI, Li Q, Kostka D, Pollard KS, Guo S, Chuang JH. 2010. The importance of Being Cis: Evolution of Orthologous Fish and Mammalian enhancer activity. *Mol Biol Evol* **27**: 2322–2332.

Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res* **20**: 1001–1009.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-jimenez CP, Mackay S, et al. 2010. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription

Factor Binding. *Science (80-)* **328**: 1036–1041.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**: 272–86.

Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordan R, Rohs R. 2014. Absence of a simple code: How transcription factors read the genome. *Trends Biochem Sci* **39**: 381–399.

Smit A, Hubley R, Green P. 2013. RepeatMasker Open-4.0. 2013-2015 . *http://www.repeatmasker.org*. http://repeatmasker.org.

Su M, Han D, Boyd-Kirkup J, Yu X, Han J-DJ. 2014. *Evolution of Alu Elements toward Enhancers*.

Taher L, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, Nobrega MA, McCallion AS, Ovcharenko I. 2011. Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res* **21**: 1139–1149.

Taher L, Narlikar L, Ovcharenko I. 2012. Clare: Cracking the LAnguage of regulatory elements. *Bioinformatics* **28**: 581–583.

The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet* **25**: 25–29.

The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43**: D1049–D1056.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**: 252–263.

Villar D, Berthelot C, Flicek P, Odom DT, Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M. 2015. Enhancer Evolution across 20 Mammalian Species. *Cell* **160**: 554–566.

Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat Rev Genet* **15**: 221–233.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser — a database of tissue-specific human enhancers. **35**: 88–92.

Wei G-H, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, et al. 2010. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J* **29**: 2147–60. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2905244&tool=pmcentrez&rendertype=abstract.

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJ, Fisher EMC, Tavaré S, Odom DT. 2008. Species-specific transcription in mice carrying human chromosome 21. *Science* **322**: 434–8.

Woo YH, Li WH. 2012. Evolutionary conservation of histone modifications in mammals. *Mol Biol Evol* **29**: 1757–1767.

Yamazaki H, Katsuoka F, Motohashi H, Engel JD, Yamamoto M. 2012. Embryonic lethality and fetal liver apoptosis in mice lacking all three small Maf proteins. *Mol Cell Biol* **32**: 808–16.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**: 931–4. http://dx.doi.org/10.1038/nmeth.3547%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26301843.