Article Submission

Methods

TITLE

Predicting lineage-specific differences in open chromatin across dozens of mammalian genomes

AUTHORS

Irene M. Kaplow[1,3,*], Daniel E. Schäffer[1], Morgan E. Wirthlin[1,3], Alyssa J. Lawler[2,3], Ashley R. Brown[1,3], Michael Kleyman[1,3], and Andreas R. Pfenning[1,2,3,*]

Carnegie Mellon University Departments of [1]Computational Biology and [2]Biology and [3]Neuroscience Institute, Pittsburgh, PA, United States

*Corresponding authors

Irene M. Kaplow: ikaplow@cs.cmu.edu

Andreas R. Pfenning: apfenning@cmu.edu

ABSTRACT

Many of the differences in complex phenotypes across species are likely due to differences in gene regulatory sequences, especially distal enhancers, that influence expression of genes across development, tissues, cell types, and conditions. Linking phenotypic differences to enhancer sequence differences remains a challenge due to the high degree of sequence variation at enhancers relative to protein-coding genes and the complexity of the code that connects genome sequence to regulatory activity in a given cellular context. Here, we demonstrate that we can accurately predict differences between species in open chromatin, a proxy for enhancer activity, using convolutional neural network models trained on genome sequence across species. We first present a new negative training set and new testing criteria that are explicitly designed to measure performance for cases where orthologous regions of the genome have different open chromatin status across species. Applying those techniques, we show that our machine learning models are more accurate at predicting the conservation of open chromatin in such regions than heavily used conservation scores are. We use our models to predict open chromatin at orthologs of brain and liver open chromatin regions across hundreds of mammals and find that brain enhancers associated with neuron activity and liver enhancers associated with liver regeneration have a stronger tendency than the general population to have predicted lineage-specific open chromatin. Overall, the framework presented here provides a mechanism to annotate tissue-specific regulatory function across hundreds of genomes and to study enhancer evolution using predicted regulatory differences rather than nucleotide-level conservation measurements.

INTRODUCTION

The molecular mechanisms underlying the incredible phenotypic diversity across mammals are largely unknown. To study these mechanisms, many consortia, including the Vertebrate Genomes Project, the Genome 10K Project (Koepfli, et al. 2015), the Bat 1K Project (Teeling, et al. 2018), and the Zoonomia Project (Zoonomia Consortium 2020), are sequencing, assembling, and aligning (Armstrong, et al. 2020) genomes from hundreds of mammals, including endangered species and species that live in remote parts of the world. Using these data, we can investigate mammalian phenotypic diversity by comparing the DNA sequences of species whose most recent common ancestors lived tens of millions of years ago. A large component of phenotypic evolution is mediated by differences in *cis*-regulatory elements, many of which are enhancers that control gene expression (King and Wilson 1975; Wray 2007). Consistent with that understanding of evolution, many complex phenotypes, including vocal learning (Pfenning, et al. 2014), domestication (Trut, et al. 2009; Albert, et al. 2012), longevity (Fraser, et al. 2005; Fushan, et al. 2015), brain size (McLean, et al. 2011), vision (Partha, et al. 2017; Berger, et al. 2018), echolocation (Shen, et al. 2012), and monogamy (Young, et al. 2019), are associated with differential gene expression between species. This understanding is further supported by studies of transcription factor (TF) binding across species that identified TF binding differences that could be underlying differences in the regulatory activity of enhancers (Schmidt, et al. 2010; Schmidt, et al. 2012). Therefore, to elucidate the ways in which complex phenotypes have evolved, new methods are required that link genome sequence differences at *cis*-regulatory elements to differences in enhancer function.

Much of our knowledge of enhancers comes from regulatory genomics measurements that are associated with enhancer activity, especially the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-Seq) and DNase hypersensitivity assay for open chromatin and chromatin immunoprecipitation sequencing (ChIP-Seq) for the histone modifications H3K27ac and H3K4me1 (Thurman, et al. 2012; Vermunt, et al. 2016). These studies have demonstrated

that enhancers, relative to genes, are substantially more tissue- or cell type-specific and generally less conserved across species (Kellis, et al. 2014; Roller, et al. 2021). Thus, identifying enhancers through direct experimentation or through comparative genomic annotation are both challenging. To overcome these challenges, multiple recent studies have described machine learning models that use DNA sequences underlying likely enhancers from a small number of mammals to predict whether DNA sequences are likely to be enhancers in other mammals (Chen, et al. 2018; Minnoye, et al. 2020).

In one recent study, support vector machines (SVMs) and convolutional neural networks (CNNs) (LeCun, et al. 1989) were able to predict which 3kb windows have the enhancer-associated histone modification H3K27ac in brain, liver, and limb tissue of human, macaque, and mouse. Importantly, the study found that models trained in one mammal achieved high accuracy in another mammal in the same clade and in another mammal in a different clade, suggesting that the regulatory code in all three of these tissues is highly conserved across mammals (Chen, et al. 2018). Two other studies have obtained similar results using another proxy for enhancer activity: open chromatin regions (OCRs). One study found that training CNNs on OCRs from multiple mammals had better performance than training CNNs on OCRs from a single mammal, albeit using 131,072bp sequences as input. The boost in power from incorporating multiple species generalized to predicting TF binding strength from ChIP-seq data and gene expression from RNA-seq data (Kelley 2020). An additional study found that a combined CNN-recurrent neural network (Graves, et al. 2009; Quang and Xie 2016) trained on sequences underlying 500bp OCRs from melanoma cell lines in one species can accurately predict melanoma cell line open chromatin in other species at a wide range of genetic distances from the training species, including in parts of the genome with low sequence conservation between the training and evaluation species (Minnoye, et al. 2020).

While these studies represent major advances in cross-species enhancer prediction, they have yet to demonstrate an ability to identify sequence differences between species that are associated with differences in regulatory genomic measurements of enhancer activity. An additional study trained SVMs to predict liver enhancers using dinucleotide-shuffled enhancers as negatives. While the overall performance was good, human enhancers whose orthologs are active in Old World Monkeys but not New World Monkeys were predicted to have consistent activity across all primates, showing that models with good overall performance do not always work well on enhancer orthologs whose activity differs between species (Klein, et al. 2018).

Some of the previous studies' methods are also limited because they require long input sequences, with some requiring input sequences > 100kb long (Kelley 2020). In order for these methods to make predictions for an enhancer ortholog, the enhancer ortholog needs to be on a sufficiently long scaffold. Since many of the new genomes assembled with short reads consist primarily of scaffolds that are less than 100kb long, and some have many scaffolds that are less than 3kb long, some existing methods could not be used for predicting the activity of most enhancer orthologs in these genomes (Zoonomia Consortium 2020). Furthermore, even in genomes with chromosome-level assemblies, identifying orthologs of long sequences is often infeasible due to genome rearrangements, often undetected, that have happened during evolution (Kim, et al. 2017; Damas, et al. 2021).

To predict and evaluate candidate *cis*-regulatory element differences across species, we chose to focus on OCRs, which are only a proxy of regulatory activity, but whose high resolution has the potential to better identify specific genome sequence differences associated with putative regulatory activity. We leveraged new, controlled open chromatin experiments conducted by our laboratory (Wirthlin, et al. unpublished work), trained a set of new models to predict OCR differences across species, and developed new criteria for evaluating models for this task. We also developed a novel

method for associating previously identified candidate enhancer sets with predicted lineage-specific open and lack of open chromatin. Our methods for evaluating approaches to OCR ortholog open chromatin status prediction and for identifying predicted lineage-specific OCRs associated with candidate enhancer sets can be applied to any tissue or cell type with open chromatin data from multiple species. This allows for reliable annotation of open chromatin across the hundreds of new genomes that are being sequenced, providing a valuable resource for research communities studying gene regulation in non-traditional model organisms, especially where direct measurements aren't feasible. We anticipate that this work will encourage researchers to develop and properly evaluate new models for predicting OCR ortholog open chromatin across species and help reveal potential functional roles of lineage-specific enhancers, enabling us to uncover transcriptional regulatory mechanisms underlying the evolution of mammalian phenotypic diversity.

NEW APPROACHES

We developed a new set of criteria for evaluating machine learning models for predicting the open chromatin status of OCR orthologs that directly evaluate the utility of such models for studying gene expression evolution. First, such models need to achieve high "Lineage-specific OCR accuracy": They need to accurately predict whether differences in sequences between species are associated with differences in open chromatin status, which we evaluate by computing performance on species- and clade-specific open chromatin and lack of open chromatin. The subset of OCR orthologs with lineage-specific open chromatin patterns are strong candidates for enhancers involved in the evolution of gene expression. Second, such models need to achieve high "Tissue-specific OCR accuracy": They need to learn both sequence patterns that are specific to the tissue in which the OCRs were identified and non-tissue-specific sequence patterns shared across OCRs from multiple tissues. Third, to demonstrate evolutionary parsimony that was previously shown to apply to mammalian enhancer activity (Roller, et al. 2021), such models' predictions should have "Phylogeny-matching correlations": Predictions across large numbers of species should approximately match what would be expected based on the species' phylogeny – mean predictions should decrease and standard deviation of predictions should increase with distance to the species with the OCRs. We used these benchmarks to evaluate how well multiple negative sets work for predicting brain and liver OCR ortholog open chromatin status. For this evaluation, we used 500bp sequences from mouse, human, macaque, and rat brain and mouse, macaque, and rat liver open chromatin peaks that we generated or obtained from publicly available datasets and datasets generated by the Pfenning Lab (ENCODE Project Consortium 2012; Thurman, et al. 2012; Fullard, et al. 2018; Liu, et al. 2019; Srinivasan, et al. unpublished work; Wirthlin, et al. unpublished work) so that we could easily identify and evaluate performance on OCR orthologs in different species. We found that good performance on chromosomes held out from training does not guarantee good performance for all of our metrics and that machine learning models tend to more accurately predict whether open chromatin status is conserved than OCRs' mean sequence conservation scores, which are based on the extent to which individual nucleotides are likely to be conserved.

In addition to developing benchmarks and using them to compare models, we developed a novel method for associating previously identified candidate enhancer sets with predicted lineage-specific open and lack of open chromatin. We first cluster OCRs, where the features are their orthologs' open chromatin predictions in each species. We do this by initially using k-means clustering to create thousands of small clusters and next using affinity propagation clustering to combine the small clusters into larger clusters. We then identify clusters overlapping candidate enhancer sets more than

expected by chance. We anticipate that this method will enable researchers discovering enhancers associated with changes in conditions to connect their findings to differences in open chromatin between species.

RESULTS

Dataset Construction for Evaluating Machine Learning Models for OCR Ortholog Open Chromatin Prediction

To demonstrate the ability to predict OCR differences across species, we created a dataset of brain OCRs and their orthologs across dozens of species. To do this, we gathered open chromatin data generated by ATAC-seq (Buenrostro, et al. 2013; Buenrostro, Wu, Chang, et al. 2015) or DNase hypersensitivity (John, et al. 2013) from two brain regions – cortex and striatum – in four species: *Homo sapiens* (ENCODE Project Consortium 2012; Thurman, et al. 2012; Fullard, et al. 2018), *Macaca Mulatta* (Wirthlin, et al. unpublished work), *Mus musculus* (Srinivasan, et al. unpublished work), and *Rattus norvegicus* (Wirthlin et al., unpublished work). We then defined our OCRs to be the 250bp in each direction of summits of non-exonic cortex open chromatin peaks that (1) overlap striatum open chromatin peaks, (2) are less than 1kb (not super-enhancers), and (3) are at least 20kb from the nearest transcription start site (TSS) so that so they would not overlap promoters. We also identified the orthologs of each OCR in each of the other species collected by the Zoonomia Project (Armstrong, et al. 2019; Zoonomia Consortium 2020) (**Fig. 1**). Rather than focus on longer regions, which have been shown to work well in some previous studies (Chen, et al. 2018; Kelley 2020), we used 500bp regions because of the ability shorter regions provide to focus on the impact of local sequence differences, because of the relative ease of obtaining orthologs of shorter regions in genomes with short scaffolds, and because predictions of enhancer activity for shorter regions are easier to experimentally validate.

We compared machine learning model performance for five different negative sets that are similar to those used for related tasks: (1) flanking regions (Arvey, et al. 2012), (2) OCRs from other tissues (Kelley 2020; Minnoye, et al. 2020), (3) about ten times as many G/C- and repeat-matched regions as positives (Chen, et al. 2018), (4) about twice as many G/C- and repeat-matched regions as positives, and (5) ten dinucleotide-shuffled versions of each positive (Alipanahi, et al. 2015) (**Supplementary fig. S1a**). We additionally created a sixth, novel negative set to force the model to learn signatures of OCRs whose orthologs' open chromatin status differs between species – sequences with closed chromatin in brain of a given species whose orthologs in another species are brain OCRs (we called this "non-OCR orthologs of OCRs" blue regions in left part of **Fig. 1**) – and included this negative set in our comparison (**Supplementary fig. S2a**). For a modeling approach, we chose CNNs (LeCun, et al. 1989; LeCun, et al. 2015) because they can model complex combinatorial relationships between sequence patterns, and changes in a single TF motif often do not cause changes in open chromatin (Slattery, et al. 2014); because they do not require an explicit featurization of the data, and many sequence patterns involved in brain open chromatin remain to be discovered; and because they can make predictions quickly relative to SVMs, the other leading approach for related tasks (Chen, et al. 2018). We did the comparison using models trained on only mouse sequences so that we could evaluate their performance on both closely and distantly related species not used for training (Chen, et al. 2018).

Optimizing and Evaluating Models for Achieving Lineage-Specific OCR Accuracy

We evaluated the machine learning model trained on each negative set on its corresponding test set. We found that all models worked well, with every model achieving an AUC > 0.85 and an AUPRC > 0.7 (**Fig. 2a, Supplementary fig. S1b**). This performance is especially impressive given that the ratios of the number of negatives to the number of positives

ranged from approximately 1.2:1 (non-OCR orthologs of OCRs) to approximately 20:1 (OCRs in other tissues). The best-performing model was the model with the negative set consisting of dinucleotide-shuffled brain OCRs (**Fig. 2a, Supplementary fig. S1b**). However, in this comparison, each model was evaluated on a different negative set, so this evaluation may not be indicative of how useful each model would be in answering questions about gene expression evolution.

We therefore also evaluated each model's lineage-specific OCR accuracy, which we did by evaluating models on the OCR orthologs whose brain open chromatin status differs between species. First, we evaluated each model on the subset of mouse brain OCRs whose orthologs in at least one other species are closed (subset of positive set for all models) and the mouse brain closed chromatin regions whose orthologs in at least one other species are open in brain. Interestingly, although all models performed decently on these genomic regions (AUC > 0.65, AUPRC > 0.55), none of the models worked as well on these genomic regions as they did on the test set corresponding to the regions used in training them, and the best-performing model on its own test set – the model trained with dinucleotide-shuffled negatives – was the worst-performing model on these regions (**Fig. 2b, Supplementary fig. S1c**). We obtained similar results for the subset of mouse brain open and closed chromatin regions whose rat orthologs have the opposite brain open chromatin status (**Supplementary fig. S1d**), showing that some of these models are capable of accurately predicting differences in brain open chromatin between closely related species. These findings reveal a bias in some existing methods for accurately predicting conserved epigenomic features and show that these biases can be mitigated by using other negative sets.

To facilitate open chromatin prediction in species for which open chromatin datasets are not available, we also evaluated whether a model trained in one species could make accurate predictions of brain open chromatin status on sequences obtained from another species (Chen, et al. 2018). However, unlike other previous work that evaluated overall performance (Chen, et al. 2018), we evaluated performance on the subset of regions in another species whose brain open chromatin status differs from that of their orthologs in the training species, as this is necessary for showing that models can accurately predict differences in brain open chromatin status between species. We therefore evaluated our models trained in mouse on macaque brain OCRs whose mouse orthologs are closed and on macaque brain closed chromatin regions whose mouse orthologs are open in brain. We found that all models achieved decent performance (AUC > 0.65, AUPRC > 0.55), with the model trained on the dinucleotide-shuffled brain OCR negatives providing the worst performance (**Fig. 2d, Supplementary fig. S1f**), further demonstrating the necessity of evaluating models on OCRs whose orthologs' open chromatin statuses differ between species. We also evaluated our models on human (**Supplementary fig. S1g**) and rat (**Supplementary fig. S1h**) regions with different brain open chromatin statuses from their mouse orthologs and found that the models generally did not work as well for such regions but still obtained decent performance and that the relative performance of different negative sets was similar to what it was for the macaque regions.

Having demonstrated the ability to train a model in one species and predict in another, we next evaluated the ability of the models to accurately predict clade-specific brain open and closed chromatin regions – regions whose brain open chromatin status is shared across species in one clade for which we have data but not shared by any species in another clade for which we have data. We focused our analyses on the two major clades for which we had experimentally determined open chromatin data (**Fig. 1**): Glires (the clade comprising all rodents and lagomorphs, including mouse and rat) and Euarchonta (the clade comprising all primates — including human and macaque — and their closest relatives, colugos and tree shrews). We found that all models obtained decent performance for Glires-specific open and closed

chromatin regions (AUC > 0.75, AUNPV-Specificity > 0.7), with the model trained on the dinucleotide-shuffled brain OCR negatives providing the worst performance and the models trained with our negative set and the larger G/C- and repeat-matched negative set providing the best performance (**Supplementary fig. S1e**). Because we trained our models on sequences from mouse (clade: Glires), we also evaluated how well our models work on clade-specific brain open and closed chromatin regions from Euarchonta species, as this clade was not used in training. We found that the models obtained slightly worse performance on Euarchonta-specific brain open and closed chromatin regions than they did on Glires-specific brain open and closed chromatin regions, with similar relative performances for the models trained on different negative sets (**Fig. 2d, Supplementary fig. S1i**). Re-calibrating the models with the positive training set and the training set from our novel negative set did not substantially change the relative performances of the models trained on different negative sets (**Supplementary notes, Supplementary tables S1-6, Supplementary fig. S2**). Since many phenotypes are clade-specific, these results demonstrate the necessity of evaluating models for OCR ortholog open chromatin prediction on clade-specific open and closed chromatin regions for clades used and not used in training.

<u>Best Overall Model Performance Does Not Guarantee Tissue-Specific OCR Accuracy</u>

In addition to evaluating model performance on OCR orthologs whose open chromatin status differs between species, we also determined if the models achieved high tissue-specific OCR accuracy. To determine if our models learned sequence patterns associated with only brain-specific open chromatin, we evaluated our models' predictions for the subset of brain OCRs that do not overlap liver OCRs and the subset of brain OCRs that overlap liver OCRs. Test set predictions from all models for both of these subsets of the positive set were usually close to one (**Fig. 3, Supplementary fig. S3a**). To determine if our models learned only sequence patterns that are indicative of general open chromatin, we also evaluated our models' predictions for the liver OCRs that do not overlap brain OCRs. We compared this to the predictions on the negative set. We found that predictions on both of these sets tended to be close to zero. However, the liver, non-brain open chromatin status predictions from the model trained with dinucleotide-shuffled OCR negatives tended to be more evenly distributed between zero and one than the liver, non-brain open chromatin status predictions for the models trained with the other negative sets (**Fig. 3, Supplementary fig. S3a**). We also did a comparison of the performances of models trained on different negative sets in which we limited the positive set to brain OCRs that overlap liver OCRs and defined the negative set as liver OCRs that do not overlap brain OCRs. We found that all models worked well (AUC > 0.75, AUPRC > 0.6) on mouse as well as on macaque and rat, which were not used in training, with the model trained on dinucleotide-shuffled brain OCR negatives providing the worst performance (**Supplementary fig. S3b**). The difference in performance may be a result of differences between the TF motifs learned by the models (**Supplementary Notes, Supplementary fig. S4**). In addition, we found that the models that tended to work better on positives were more accurate for shared brain and liver OCRs than for liver, non-brain OCRs and that the models that tended to work better on brain closed chromatin orthologs of brain OCRs were more accurate for liver, non-brain OCRs than for shared brain and liver OCRs (**Supplementary tables S8-9**, **Supplementary fig. S3c**). Calibration with the positive training set and the training set from our novel negative set had a similar effect to calibration for clade-specific open and closed chromatin regions (**Supplementary tables S8-9**, **Supplementary fig. S3c**). These results show that models with good overall performance are not always able to comprehensively learn open chromatin sequence patterns that are specific to the tissue in which they were trained.

<u>Predictions from Models of OCR Ortholog Open Chromatin Status Have Phylogeny-Matching Correlations</u>

We also determined whether our models' predictions have phylogeny-matching correlations in a way that does not require open chromatin data from multiple species. To do this, we obtained the orthologs of the mouse brain OCRs in all of the fifty-six Glires clade species in the Zoonomia Project (Armstrong, et al. 2020; Zoonomia Consortium 2020), used our machine learning models to predict the brain open chromatin statuses of these orthologs, computed the mean brain open chromatin statuses across all brain OCR orthologs in each species, and computed the correlation between mean predicted brain open chromatin status and evolutionary distance from mouse. As we expected based on the principle of evolutionary parsimony and a previous study of enhancer activity across multiple species (Roller, et al. 2021), all models showed a strong negative correlation between mean predicted brain open chromatin status and divergence from mouse (**Fig. 4**, **Supplementary fig. S5a**). Nevertheless, there is still more open chromatin at these brain OCR orthologs than would be expected from brain non-OCRs, even in the most distantly related Glires species, because all mean predictions are greater than the mean predictions for the negative test sets (**Fig. 4, Supplementary fig. S5a**). We also expected there to be a strong positive correlation between the standard deviation of open chromatin status and divergence from mouse because most brain OCR orthologs in species closely related to mouse are active in brain, while the brain open chromatin status of brain OCR orthologs in species that are more distantly related should vary. We found this expected positive correlation for all negative sets (**Supplementary fig. S5b**).

<u>Machine Learning Models Trained Using Data from Multiple Species Can Accurately Predict OCR Orthologs' Open Chromatin Statuses</u>

Based on other research showing that training models with data from multiple species can improve OCR prediction accuracy (Kelley 2020), we trained additional machine learning models using open chromatin data from multiple species. For each of brain and liver, we used the open chromatin data from all of the species that we had collected as positives (four species for brain, three species for liver) and the orthologs of all these OCRs in the other species for which we had data that did not overlap brain or liver open chromatin, respectively, as negatives. Before training multi-species models, we trained mouse-only models for liver and found that they achieved good performance for all of our metrics (**Supplementary notes, Supplementary fig. S6, Supplementary table S10**). We then trained brain and liver multi-species models and found that they achieved high lineage-specific and tissue-specific accuracy, where performance was generally better than the performance for any of the models trained on only mouse sequences (**Figs. 5a-b**). In addition, we determined if the multi-species brain and liver models' predictions had phylogeny-matching correlations by using them to predict the OCR ortholog open chromatin status of mouse brain and liver OCRs, respectively, across Glires and found strong negative correlations between divergence from mouse and mean OCR ortholog open chromatin status predictions (**Figs. 5c-d**). We also found strong positive correlations between divergence from mouse and standard deviations of OCR ortholog open chromatin status predictions (**Supplementary figs. S7a-b**). For both these models and our mouse-only models, our phylogeny-matching correlation results could not be fully explained by genome quality (**Supplementary notes, Supplementary fig. S8, Supplementary tables S11-12**). Finally, we evaluated our mouse-only and multi-species liver models on Laurasiatheria-specific liver OCRs and liver non-OCRs (Halstead, et al. 2020), where no Laurasiatheria were used in training either model, and found that the multi-species liver model had better performance than the mouse-only liver model (**Fig. 5e**). Like our mouse-only models, our multi-species brain and liver models also learned TF motifs for TFs that are known to be involved in brain and liver, respectively (**Supplementary notes, Supplementary figs. S7c-**

d). Overall, these results suggest that machine learning models trained on data from multiple species can accurately predict open chromatin statuses of OCR orthologs.

Machine Learning Models Make More Accurate Predictions than Mean Conservation Scores

We compared the test set predictions of our multi-species models to those made by mean conservation scores. First, we found that our model predictions for orthologs in other species of mouse brain and liver OCRs whose OCR status is conserved tends to be higher than for orthologs in other species of mouse brain and liver OCRs whose OCR status is not conserved (**Supplementary tables S13-14**). Then, for each tissue, we ranked the macaque OCR orthologs based on their model predictions, with the highest rank corresponding to the highest score or open chromatin status prediction. For the OCR status-conserved OCRs, we evaluated whether these OCRs tended to have higher ranks for our multi-species model predictions than they did for mean conservation scores; we found that the ranks were significantly higher for our predictions (brain predictions vs. PhastCons scores: 7.80×10^{-5}, brain predictions vs. PhyloP scores: 1.53×10^{-6}, liver predictions vs. PhastCons scores: 3.68×10^{-22}, liver predictions vs. PhyloP scores: 1.25×10^{-27}, **Fig. 6a**). For each tissue, we also evaluated whether the OCR orthologs without open chromatin tended to have lower ranks for our multi-species model predictions than they do for mean conservation scores, and we found that the ranks were significantly lower for our predictions (brain predictions vs. PhastCons scores: 4.34×10^{-5}, brain predictions vs. PhyloP scores: 8.08×10^{-9}, liver predictions vs. PhastCons scores: 7.15×10^{-14}, liver predictions vs. PhyloP scores: 9.82×10^{-22}, **Fig. 6a**). We also did this for human and rat brain orthologs of mouse brain OCRs with conserved and non-conserved open chromatin statuses, for rat liver orthologs of mouse liver OCRs with conserved and non-conserved open chromatin statuses, and for all of these groups of OCR orthologs with our mouse-only models and obtained similar results (**Supplementary notes, Supplementary figs. S9-10, Supplementary tables S13-16**).

Some of the OCR orthologs for which our multi-species brain model correctly predicted brain open chromatin conservation in spite of low mean sequence conservation or for which our models correctly predicted lack of brain open chromatin conservation in spite of high mean sequence conservation are near genes that have been shown to play important roles in the brain. For example, there is a region on mouse chromosome 2 – part of our test set – that has low mean sequence conservation according to PhastCons (Siepel, et al. 2005) and PhyloP (Pollard, et al. 2010) but high brain experimentally identified and predicted open chromatin conservation between mouse and macaque (**Fig. 6b**) and whose mouse and macaque orthologs are located near the gene Stx16. This gene is involved in vesicle trafficking in most tissues, including the brain (Chua and Tang 2009), and may play a role in Alzheimer's disease (Ray and Zhang 2010); in fact, its role in axon regeneration is conserved between mammals and *C. elegans* (Chen, Liu, et al. 2016). Although this region near Stx16 has generally low sequence conservation, running TomTom (Gupta, et al. 2007) on the 22bp sequence with high conservation revealed a subsequence that is similar to the Fos motif, which is also found in the macaque ortholog. Since our machine learning model used sequence similarity to the Fos motif in making predictions (**Supplementary fig. S7c**), the machine learning model was likely able to automatically determine that it should use this sequence in making its prediction. In addition, there is a region on mouse chromosome 2 that has high mean sequence conservation but low experimentally identified and predicted open chromatin conservation between mouse and macaque (**Fig. 6c**) and whose mouse and macaque orthologs are located near the gene Lnpk. This gene is an endoplasmic reticulum junction stabilizer (Mookherjee, et al. 2019) that has been shown to play a role in brain and limb development (Spitz, et al. 2003), and mutations in this gene have been associated with neurodevelopmental disorders (Breuss, et al. 2018). It is possible that

this region of the genome near Lnpk has a high degree of sequence conservation because it has functions in other tissues. These results demonstrate the potential benefits of using OCR ortholog open chromatin status predictions instead of mean conservation scores for studying the evolution of the expression of important genes in a tissue of interest.

In addition, some of the accurate multi-species model liver open chromatin conservation predictions disagreed with the mean conservation scores. For instance, there is a region on mouse chromosome 2 that has high experimentally identified and predicted liver open chromatin conservation but low sequence conservation (**Fig. 6d**) and whose mouse and macaque orthologs are located near Rxra. This gene is a TF involved in regulating lipid metabolism (Zhang, et al. 2015; Stossi, et al. 2019; Xue, et al. 2020), TF-MoDISco identified a sequence similar to its motif as being important in our liver models (**Supplementary fig. S7d, Supplementary fig. S9b**), and its liver expression is stable across fifteen mammals (Berthelot, et al. 2018). Although this region near Rxra has generally low sequence conservation, the 15bp segment with high conservation is similar to the motif for Ctcf according to TomTom (Gupta, et al. 2007), and that motif is also found in the macaque ortholog. Since our machine learning model used sequence similarity to the Ctcf motif in making predictions (**Supplementary fig. S7d**), the machine learning model was likely able to automatically determine that it should use this sequence in making its prediction. There is also a region on mouse chromosome 1, which is part of our test set, whose mouse ortholog is an OCR and macaque ortholog is not an OCR according to our data and predictions in spite of being highly conserved (**Fig. 6e**) and whose mouse and macaque orthologs are near Fn1. This gene has been implicated in liver fibrosis (Hernandez-Gea and Friedman 2011; Chen, Wang, et al. 2016; Govaere, et al. 2019), and a multi-species liver RNA-seq study found that it has higher expression in mouse liver relative to livers of other mammals and birds and lower expression in primate livers relative to livers of other mammals and birds (Brawand, et al. 2011). For both of these OCRs, the H3K27ac signal conservation in the same regions (Villar, et al. 2015) is similar to the open chromatin status conservation, suggesting that the open chromatin status conservation is indicative of enhancer activity conservation (**Figs. 6d-e**). These results suggest that using predicted OCR ortholog open chromatin status instead of conservation has the potential to be beneficial for understanding gene expression evolution in multiple tissues.

<u>Lineage-Specific Brain and Liver OCRs Are Associated with Neuron and Liver Functions</u>

Since our models can accurately predict lineage-specific OCRs, we evaluated whether lineage-specific brain and liver OCRs were associated with brain and liver functions. First, we used publicly available RNA-seq data from the livers of thirteen placental mammals with high-quality genomes (Berthelot, et al. 2018) to identify genes with Rodent-specific expression; we obtained 15 genes with false discovery rate (FDR) < 0.05. We then used our predictions to identify liver OCRs with predicted Rodent-specific activity. One of the genes with rodent-specific expression that is near OCRs with predicted liver-specific open chromatin is Atp9a (**Supplementary fig. S11**). This gene has high expression in mouse liver according to Mouse Genome Informatics (Smith, et al. 2019), it has little or no RNA or protein expression in human liver according to all of the studies referenced in the Human Protein Atlas (Uhlen, et al. 2015), and its protein is expressed in the rat canicular membrane (Chaubey, et al. 2016), a part of hepatocytes involved in bile production. This suggests that OCRs with predicted lineage-specific open chromatin may be regulating genes with lineage-specific expression.

To further investigate the roles of predicted lineage-specific OCRs, we clustered each of the brain and liver OCRs, where the features were the predicted activity in each of the Boreoeutherians from the Zoonomia Project (Zoonomia Consortium 2020). For each of brain and liver, we obtained a little over one hundred clusters with different patterns of predicted open chromatin across species (brain clusters:

https://github.com/pfenninglab/OCROrthologPrediction/clusters/brain, liver clusters: https://github.com/pfenninglab/OCROrthologPrediction/liver). We then determined whether each brain cluster that was open in mouse overlapped mouse candidate enhancers associated with neuron firing (Beagan, et al. 2020) more than expected by chance (**Supplementary table S17**) and each brain cluster that was open in human overlapped human candidate enhancers associated with neuron activity (Boulting, et al. 2021) more than expected by chance (**Supplementary table S18**). Interestingly, the candidate enhancers from each of these sets intersected clusters with predicted lineage-specific open chromatin or predicted lineage-specific lack of open chromatin more than expected by chance. Specifically, mouse neuron bicuculline (Bic)-specific candidate enhancers, where Bic induces neuron firing, were enriched for overlapping two predicted Murinae-specific brain open chromatin clusters – cluster 43 (**Fig. 7**) and cluster 27 (**Supplementary fig. 12a**). We think that these results are unlikely to be explained by the number of usable orthologs or conservation because mouse brain OCRs overlapping Bic-specific candidate enhancers do not have significantly fewer usable orthologs or lower conservation according to PhastCons (Siepel, et al. 2005) or PhyloP (Pollard, et al. 2010) than mouse brain OCRs in general. In contrast, mouse activity-invariant candidate enhancers were enriched for overlapping the predicted open chromatin in all species cluster (cluster 1), a noisy cluster without a clear pattern of predicted open chromatin (cluster 37), a noisy predicted Yangochioptera-specific brain non-open chromatin cluster (cluster 81), and a noisy predicted Primate-specific brain non-open chromatin cluster (cluster 88). Likewise, human candidate enhancers from GABAergic neurons made from hiPSCs from four genotypes that had increased activity two hours after exposure to potassium chloride (KCl), where KCl induces neuron activity, were enriched for overlapping a predicted Carnivora-, Perissodactyla- and Euarchonta-specific brain open chromatin cluster (cluster 74, **Fig. 7**), a predicted Hystricognathi-specific brain non-open chromatin cluster (cluster 11, **Supplementary fig. 12b**), and a predicted Muroidea- and Pecora-specific brain non-open chromatin (cluster 48, **Supplementary fig. 12b**). These results are unlikely to be explained the number of usable orthologs or conservation because human brain OCRs overlapping this set of GABAergic neuron candidate enhancers tended to have more usable orthologs and higher conservation according to both PhastCons (Siepel, et al. 2005) and PhyloP (Siepel, et al. 2005) than human brain OCRs overall. In contrast, candidate enhancers from the same source that had decreased activity two hours after exposure to KCl were enriched for overlapping the predicted open chromatin in all species cluster (cluster 1) and a predicted Ruminantia-specific brain non-open chromatin cluster (cluster 82). These results suggest that there may be a relationship between enhancer response to neuron activity and whether the enhancer's activity tends to be specific to the lineage in which it was identified.

We also investigated whether each liver cluster that was active in mouse overlapped mouse candidate enhancers associated with liver regeneration (Wang, et al. 2020) more than expected by chance (**Supplementary table S19**). We found that candidate liver enhancers that have increased activity four weeks after hepatocyte repopulation relative to the control were enriched for overlapping a cluster with predicted Murinae-specific open chromatin – cluster 29 – as well as two clusters with predicted Muroidea-specific open chromatin – cluster 36 and cluster 100 (**Supplementary fig. 12c**). We think that these results are unlikely to be explained by the number of usable orthologs or conservation because OCRs overlapping this candidate enhancer set do not have significantly fewer usable orthologs or lower conservation according to PhastCons (Siepel, et al. 2005) or PhyloP (Pollard, et al. 2010) than liver OCRs overall. We also found that candidate liver enhancers that have increased activity one week after hepatocyte repopulation relative to the control were enriched for overlapping a cluster with noisy predicted Primate-specific liver non-open chromatin (cluster 83). In contrast to these

findings, liver candidate enhancers with decreased activity four weeks after hepatocyte repopulation relative to the control were not enriched for overlapping any clusters, and liver candidate enhancers with decreased activity four weeks after hepatocyte repopulation relative to one week after hepatocyte repopulation were enriched for overlapping a cluster without a clear pattern of predicted open chromatin (cluster 39). These results suggest that there may be a relationship between liver regeneration in Murinae and Murinae-specific open chromatin.

DISCUSSION

The ability to link differences in genome sequence across species to the evolution of complex traits is becoming tractable due to increased availability in genomic datasets and advances in computational methodologies. Although gene regulatory mechanisms have been strongly linked with the evolution of a number of complex traits (King and Wilson 1975; Pfenning, et al. 2014), enhancers, which play a crucial role in gene regulation, have been difficult to study due to their tissue-specificity (Kundaje, et al. 2015) and relatively poor sequence conservation across species (Hoffman, et al. 2013; Kellis, et al. 2014). In recent years, computational models have been constructed that can predict enhancer-associated regulatory genomics features with relatively high accuracy (Ghandi, et al. 2014; Zhou and Troyanskaya 2015). Their success has largely been a result of them not requiring an explicit featurization of the input sequence, as well-characterized sequence features like TF motifs encompass only a small subset of the sequence features involved in enhancer activity (Slattery, et al. 2014). However, in spite of these models' overall strong performance, it has yet to be demonstrated that they can achieve high lineage-specific accuracy. Indeed, when we evaluated the performance of CNN models for predicting open chromatin, their performance was often substantially stronger for regions with conserved open chromatin relative to those whose open chromatin status differed across species. We were able to improve this performance by shifting our negative sets in model training from negative sets like dinucleotide-shuffled OCRs to negative sets consisting of closed chromatin regions whose orthologs in other species are OCRs or large numbers of regions with matching G/C-content. It is possible that other negative sets, such as genome-wide negatives or combinations of negative sets that work well, could further improve performance, but such negative sets would require substantially more training time than our novel negative set, making training and tuning such models potentially prohibitive for researchers who do not have access to multiple graphical processing units (GPUs). In addition, should models that using other negative sets or overcoming some of the limitations of our models (**Supplementary Notes**) be developed, our evaluation metrics can be used to determine the utility of such models in identifying mechanisms associated with gene expression evolution.

Efforts of the Vertebrate Genomes Project, the Zoonomia Project (Zoonomia Consortium 2020), the Bat1K Project (Teeling, et al. 2018), the Genome 10K Project (Koepfli, et al. 2015), and others are making the genomes of hundreds of vertebrate species available to the scientific community. In parallel, computational tools are being developed to improve assembly, alignment, and identification of orthologous genes and regulatory elements (Hickey, et al. 2013; Zhang, et al. 2020). Leveraging these resources, we demonstrated the ability of our computational models to predict OCR ortholog open chromatin status in dozens of species beyond those that we used for training, including correctly predicting differences in open chromatin status between OCR orthologs. Furthermore, we found a relationship between evolutionary distance and predicted open chromatin status differences in the brain and the liver, which is further supported by a previous study showing a similar relationship using histone modification ChIP-seq data from a smaller number of mammals (Roller, et al. 2021). Although there are OCR orthologs, such as species-specific OCRs and OCRs with convergently evolved open

chromatin (Partha, et al. 2017), whose open chromatin status across species is not associated with phylogenetic distance, we think that such OCRs in most tissues are a minority. This evaluation does not require experimentally determined OCR data from multiple species; it requires only genomes from a large number of species, which are becoming increasingly available (Koepfli, et al. 2015; Zoonomia Consortium 2020).

In addition to showing that our novel models achieved lineage-specific OCR accuracy, achieved tissue-specific OCR accuracy, and made predictions that had phylogeny-matching correlations, we also showed that they were substantially more effective at predicting conservation or lack of conservation of open chromatin than mean conservation scores generated by PhastCons (Siepel, et al. 2005) or PhyloP (Pollard, et al. 2010). In some cases, the OCR orthologs whose open chromatin status conservation or lack of conservation were correctly predicted by our models but not by mean conservation scores were near important genes in the tissue whose expression has the same cross-species trend as the open chromatin (Brawand, et al. 2011; Berthelot, et al. 2018), demonstrating the potential machine learning model predictions have to identify candidate OCRs involved in regulating important genes. We think our predictions work better than mean conservation scores because conservation-based methods for predicting open chromatin status conservation consider every nucleotide within an OCR to be equally important (Siepel, et al. 2005; Pollard, et al. 2010; Partha, et al. 2017), whereas CNNs automatically learn which nucleotides within a sequence are most strongly associated with open chromatin. Thus, if most of an OCR is not conserved but the motif of a TF that is important for establishing open chromatin is conserved, then the region may have conserved open chromatin in spite of having overall low conservation. Furthermore, the high rate of turnover of specific transcription factor binding sites suggests that the regulatory function of a region can be conserved in regions for which nucleotide-level conservation is difficult to detect (Odom, et al. 2007; Schmidt, et al. 2010). In addition, a region might be open in multiple tissues and be highly conserved because its open chromatin status in some tissues is highly conserved even though its open chromatin status in a tissue of interest is not conserved; thus, this further demonstrates the necessity of showing that models achieve high tissue-specific accuracy.

Using 500bp sequences, as compared to other approaches that require sequences of 3kb (Chen, et al. 2018) or over 100kb (Kelley 2020), to train our OCR ortholog open chromatin status prediction models has advantages beyond enabling us to meet most of the criteria that we recommended. First, this enables us to make predictions in species whose assemblies have short scaffolds, which include many of the species in the Zoonomia Project (Zoonomia Consortium 2020), and species whose genomes are substantially rearranged relative to the species with the open chromatin data, such as gibbons (Carbone, et al. 2014). In addition, it allows us to identify short regions with sequence differences that are likely to have caused open chromatin differences, which is useful for downstream work to identify mechanisms underlying evolutionary differences in open chromatin. Finally, using short sequences allows us to experimentally validate predictions in a reporter assay (Shen, et al. 2016; Klein, et al. 2018). While our negative set requires open chromatin data from at least two and ideally at least three species, we found that using a large number of random, G/C-matched negatives (Chen, et al. 2018) works well for training models on 500bp sequences with open chromatin data from only one species.

In addition to meeting our evaluation criteria, the predictions made by our models revealed that many open chromatin regions may be lineage-specific or may be closed in specific lineages and that these lineage-specific patterns of open chromatin may be functional. We identified a predicted rodent-specific liver open chromatin region near Atp9a, a gene with rodent-specific expression in the liver that may play a role in bile production in rats (Chaubey, et al. 2016). In addition, we developed a novel approach to deciphering the putative lineage-specificity of candidate enhancers identified

in other studies in which we clustered our OCRs using predicted activity as features and identified clusters with more overlap with these candidate enhancers than expected by chance. Using this approach, we found that mouse candidate enhancers activated during neuron firing (Beagan, et al. 2020) are enriched for overlapping clusters with predicted Murinae-specific open chromatin, and human candidate enhancers activated during neuron activity (Boulting, et al. 2021) are enriched for overlapping clusters with predicted Euarchonta-specific gains or Muroidea-specific losses of open chromatin, suggesting that there may be Muroidea-specific gene expression and enhancer signatures of Muroidea neuron activity. Although no study, to our knowledge, has evaluated this for large numbers of species, a few studies have compared neuron activity between mice and one or two Euarchonta and found striking differences, including lower spiking frequencies in fast-spiking mouse cortical neurons relative to human and rhesus macaque (Wang, et al. 2016) and lack of expression of h-Channels in mouse excitatory neurons relative to human (Kalmbach, et al. 2018). We also found that candidate enhancers activated during liver regeneration (Wang, et al. 2020) are enriched for overlapping predicted Murinae-specific open chromatin. We found clusters with predicted lineage-specific open chromatin for additional linages, predicted lineage-specific losses in open chromatin for additional lineages, and predicted convergent gains and losses in open chromatin; investigating the roles of the open chromatin regions in these clusters could be an exciting path towards identifying mechanisms underlying linage-specific differences between mammalian brains and livers.

The metrics that we presented for evaluating OCR ortholog open chromatin status prediction methods and our approach to identifying candidate enhancers from other studies enriched for overlapping OCRs with predicted lineage-specific open chromatin to any tissue or cell type with data from open chromatin assays in multiple species. Beyond the data employed in this study, there are publicly available open chromatin data from both human and mouse from many tissues (ENCODE Project Consortium 2012; Liu, et al. 2019) and from dog (Megquier, et al. 2019), cow, and pig (Giuffra, et al. 2019) from multiple tissues. An exciting extension would be to train new machine learning models to predict the open chromatin status of OCR orthologs in additional species and tissues, making use of these datasets as well as the wealth of similar OCR datasets soon to be released. As single-cell ATAC-seq (scATAC-seq) (Buenrostro, Wu, Litzenburger, et al. 2015) becomes more widely used, these methods could be applied to predicting OCR ortholog open chromatin status in specific cell types and determining if specific cell types have been lost in some species.

The results from our method can also be used in a forward genomics approach (Hiller, et al. 2012) to help identify the mechanisms underlying the evolution of the expression of genes of interest or of phenotypes that have evolved through gene expression. This can be done by identifying OCR orthologs whose differences in predicted open chromatin status correspond to differences in gene expression or phenotypes. Many multi-species gene expression datasets in tissues with data from an assay that can serve as a proxy for enhancers are publicly available (Sudmant, et al. 2015; Berthelot, et al. 2018), and more will likely be generated in the near future (Giuffra, et al. 2019). Additional multi-species single-cell RNA-seq datasets are being generated from some of these tissues (Tosches, et al. 2018; Zhu, et al. 2018) and will likely soon be supplemented by scATAC-seq. Many of these tissues have been associated with phenotypes that have evolved through gene expression (Pfenning, et al. 2014; Fushan, et al. 2015), so we can use the gene expression data along with predictions from models like ours to gain insights how these phenotypes evolved. Such insights may also reveal mechanisms underlying diseases associated with these phenotypes (Minnoye, et al. 2020).

MATERIALS AND METHODS

Constructing Negative Set for Non-OCR Orthologs of OCRs

Data generation and processing, construction of positive sets, and construction of other negative sets are described in **Supplementary Materials and Methods**. We obtained the non-OCR orthologs of OCRs by obtaining the orthologs of OCRs in each species in each other species for which we had data from the same tissue and filtering the orthologs to include only those that did not overlap open chromatin in the same tissue. For example, mouse chr4:127435564-127436049 does not overlap any OCRs in mouse cortex or striatum, but its human ortholog, chr1:34684126-34684689, is an OCR in both cortex and striatum, so this mouse region was a member of our novel negative set. To ensure that we would have enough negatives for training the model, we created a less conservative set of OCRs, which we called "loose OCRs." For each species, tissue combination, the loose OCRs are the base peaks that are non-exonic, at least 20kb from a TSS, and at most 1kb long (same criteria used in constructing the positive set) and intersect at least 1 peak from the pooled reads across replicates from each of other datasets that are used for the species, tissue combination; we obtained these loose OCRs using bedtools (Quinlan and Hall 2010). We defined the peak summit of a loose OCR to be the peak summit of the corresponding base peak.

We identified orthologs of our loose OCRs in each other species with open chromatin data from the same tissue using halLiftover (Hickey, et al. 2013) followed by HALPER (Zhang, et al. 2020). We used these tools instead of liftOver (Kent, et al. 2003) because they map regions using Cactus alignments (Paten, et al. 2011), which, unlike the pairwise alignment liftOver chains, contain many species and account for a wide range of structural rearrangements, including inversions. We ran halLiftover with default parameters on our loose OCRs and their peak summits using the Zoonomia version 1 Cactus alignment (Armstrong, et al. 2020). We then constructed contiguous loose OCR orthologs from the outputs of halLiftover by running HALPER with parameters -max_frac 2.0, -min_len 50, and -protect_dist 5. Finally, we used bedtools subract with the -A option (Quinlan and Hall 2010) to remove the loose OCR orthologs in each species that overlapped peaks from the pooled reads across replicates from any of the datasets from the same tissue in that species.

For the models trained only on mouse sequences, we used only the mouse orthologs of loose OCRs in non-mouse species. For the multi-species brain models, we used the mouse, human, macaque, and rat orthologs of loose brain OCRs from each of the other species, and for the multi-species liver models, we used the mouse, macaque, and rat orthologs of loose liver OCRs from each of the other species. When constructing negatives of the non-OCR orthologs of OCRs, we used sequence underlying the ortholog of the base peak summit +/- 250bp and that sequence's reverse complement. Our negative:positive training set ratios were approximately 1.16:1 for the brain model trained on only mouse sequences, 0.704:1 for the liver model trained on only mouse sequences, 1.49:1 for the multi-species brain model, and 0.822:1 for the multi-species liver model.

Evaluating Machine Learning Models

Evaluating Models' Lineage-Specific OCR Accuracy

Training, validation, and test set construction and machine learning model training are described in **Supplementary Materials and Methods**. To obtain OCR orthologs that are open in one species but not in another, we used as positives the sequences and reverse complements of sequences underlying OCRs from one species whose orthologs in the other species have closed chromatin and as negatives sequences and reverse complements of sequences underlying non-OCRs whose orthologs in at least one other species are OCRs. More specifically, for evaluating mouse OCRs whose open chromatin status differs in other species, we used halLiftover (Hickey, et al. 2013) followed by HALPER (Zhang, et

al. 2020) with the same parameters we used previously to identify mouse OCR orthologs in human, macaque, and rat for brain and macaque and rat for liver that did not overlap any of the peaks from the pooled replicates from any of the datasets that we used for identifying OCRs. This gave us 1,570 positive test set examples for evaluating brain models and 3,738 positive test set examples for evaluating liver models. For evaluating mouse non-OCRs whose open chromatin status differs in other species, we used the same approach that we took to construct our novel negative set except that we identified non-OCR orthologs of OCRs instead of loose OCRs. This gave us 2,062 negative test set examples for evaluating brain models and 4,080 negative test set examples for evaluating liver models (**Fig. 2b, Supplementary fig. S1c, Supplementary fig. S6b**). For evaluating mouse OCRs whose rat orthologs are not OCRs, we identified mouse OCRs whose rat orthologs do not overlap pooled peaks across replicates for any rat OCRs from any of the datasets used to define OCRs; this gave us 674 positive test set examples for evaluating brain models and 2,482 positive test set examples for evaluating liver models. For evaluating mouse closed chromatin regions whose orthologs are open in rat, we used the subset of our novel negative set that came from rat OCR orthologs (not rat loose OCR orthologs); this gave us 890 negative test set examples for evaluating brain models and 2,050 negative test set examples for evaluating liver models (**Supplementary fig. S1d, Supplementary fig. S6b**). For evaluating macaque OCRs whose open chromatin status differs in mouse, we identified macaque OCRs whose mouse orthologs do not overlap peaks from the pooled replicates of any of the datasets we used for identifying mouse OCRs; this gave us 734 positive test set examples for evaluating brain models and 2,384 positive test set examples for evaluating liver models. For identifying macaque non-OCRs whose OCR status differs in mouse, we identified mouse OCR orthologs in macaque that do not overlap peaks from the pooled replicates from any of the datasets we used for identifying macaque OCRs; this gave us 788 negative test set examples for evaluating brain models and 2,228 negative test set examples for evaluating liver models (**Fig. 2c**, **Supplementary fig. S1f, Supplementary fig. S6b**). We obtained human and rat OCRs and non-OCRs whose OCR status is different in mouse using same process that we used for macaque. We obtained 416 positive and 896 negative test set examples for evaluating brain models (**Supplementary fig. S1g**). For evaluating rat OCRs and non-OCRs whose mouse orthologs have different OCR statuses, we obtained 990 positive and 676 negative test set examples for evaluating brain models and 2,050 positive and 2,482 negative test set examples for evaluating liver models (**Supplementary fig. S1h, Supplementary fig. S6b**). Identification of clade- and species-specific OCRs is described in **Supplementary Materials and Methods**.

Evaluating Models' Tissue-Specific OCR Accuracy

To evaluate the performance of models trained in one tissue on OCRs from another tissue, we defined our positives to be OCRs that are shared between the two tissues (shows our models were not learning only the sequences involved in tissues-specific open chromatin), and we defined our negatives to be OCRs in the evaluation tissue that do not overlap OCRs in the training tissue (shows our models were not learning only sequences involved in general open chromatin). More specifically, we used bedtools intersectBed with options -wa and -u (Quinlan and Hall 2010) to identify OCRs from our training tissue that overlap OCRs from the evaluation tissue. For brain models, we obtained 1,040 positives for mouse, 846 positives for macaque, and 1,770 positives for rat (**Fig. 3, Supplementary fig. S3**). For liver models, we obtained 2,012 positives for mouse, 946 positives for macaque, and 1,130 positives for rat (**Supplementary fig. S6c**). We used bedtools subtractBed with option -A (Quinlan and Hall 2010) to identify liver OCRs that do not overlap any open chromatin from pooled replicates in any of the datasets that were used to generate the brain OCRs and brain OCRs that do not overlap open chromatin from pooled replicates in any of the datasets that were used to generate liver OCRs. For brain,

we obtained 3,382 negatives for mouse, 1,898 negatives for macaque, and 3,518 negatives for rat (**Fig. 3, Supplementary fig. S3**). For liver, we obtained 2,212 negatives for mouse, 1,428 negatives for macaque, and 2,942 negatives for rat (**Supplementary fig. S6c**). We combined the data from all three species for evaluating the multi-species models (**Figs. 5a-b**).

For the negative set comparison, we also compared the distributions of test set predictions for the brain OCRs that do not overlap liver OCRs, the brain OCRs that overlap liver OCRs, the liver OCRs that do not overlap brain OCRs, and the negative set. We defined these groups of OCRs as we did for other evaluations, and we used predictions for sequences and their reverse complements. We compared the distributions for the brain OCRs that overlap liver OCRs to the liver OCRs that do not overlap brain OCRs using a Wilcoxon rank-sum test and multiplied the p-values by 6 to do a Bonferroni correction.

Evaluating if Models' Predictions Had Phylogeny-Matching Correlations

To evaluate the relationship between OCR ortholog open chromatin status and phylogenetic distance, we identified test set mouse OCR orthologs in all of the fifty-six Glires species from Zoonomia, predicted the open chromatin statuses of those orthologs, and computed the correlation between those predictions and the species' phylogenetic divergences from mouse. This provides us with an approximate measure of how predicted OCR ortholog open chromatin statuses change over evolution. We identified the test set mouse OCR orthologs and OCR summit orthologs in Glires using halLiftover (Hickey, et al. 2013) with the Zoonomia version 1 Cactus alignment (Paten, et al. 2011; Armstrong, et al. 2020); we used brain OCRs for evaluating brain OCR models and liver OCRs for evaluating liver OCR models. We next constructed contiguous orthologs from the outputs of halLiftover using HALPER (Zhang, et al. 2020) with parameters -max_frac 2.0, -min_len 50, and -protect_dist 5. We constructed inputs for our models from the contiguous OCR orthologs by using bedtools fastaFromBed (Quinlan and Hall 2010) with fasta files downloaded from NCBI (NCBI Resource Coordinators 2018; Zoonomia Consortium 2020) and the UCSC Genome Browser (Kent, et al. 2002) to obtain the sequences underlying their summit orthologs +/- 250bp. We constructed the reverse complements of sequences, used our models to predict each sequence and its reverse complement's open chromatin status, and averaged the predictions between the forward and reverse strands. We then removed all predictions from OCRs with orthologs in less than one quarter of species. After that, for each model, we computed the mean OCR ortholog open chromatin status prediction and the standard deviation of predictions across all remaining OCR orthologs in each species. We finally computed the Pearson and Spearman correlations between these means and standard deviations of predictions and the millions of years since divergence from mouse, which we obtained from TimeTree (Kumar, et al. 2017). We did this for brain OCR orthologs using brain models trained on mouse sequences from each negative set and the multi-species brain model as well as for liver OCR orthologs using the liver model trained on only mouse sequences and the multi-species liver model.

<u>Comparing Predictions to Mean Conservation Scores</u>

We compared the predictions to mean conservation scores by identifying OCR orthologs with conserved and non-conserved open chromatin status between species (**Supplementary Materials and Methods**), computing the mean conservation scores of those OCR orthologs (**Supplementary Materials and Methods**), and comparing those scores to the predicted open chromatin status of those OCR orthologs. We evaluated whether the predictions were more effective than the mean conservation scores at differentiating between open chromatin status-conserved and open chromatin status non-conserved OCR orthologs. We first averaged the predictions of the sequence underlying the non-mouse OCR

ortholog's summit +/-250bp and its reverse complement so that each OCR ortholog would have a single prediction value. We next combined our open chromatin status-conserved and open chromatin status non-conserved OCR orthologs and ranked them according to each of PhastCons score, PhyloP score, and OCR ortholog open chromatin status prediction. We then did a Wilcoxon sign-rank test to compare the ranking distributions of the open chromatin status-conserved OCR orthologs between the OCR ortholog open chromatin status predictions and each type of conservation score. We also did this for the ranking distributions of the open chromatin status non-conserved OCR orthologs. We did this for predictions made by the models trained using only mouse sequences and by the multi-species models. Finally, we did a Bonferroni correction by multiplying all p-values by 40 (2 conservation score comparisons for each of open chromatin status-conserved and open chromatin status non-conserved OCR orthologs in 2 tissues for 5 species, tissue pairs).

Clustering OCRs:

To generate OCR clusters for OCRs from a tissue, we mapped the OCRs from each species across all of the Boreoeutherians from Zoonomia (Zoonomia Consortium 2020) except for *Manis tricuspis* (**Supplementary Materials and Methods**), filtered the OCRs (**Supplementary Materials and Methods**), and clustered the OCRs. We clustered the remaining OCRs using the prediction of each species's OCR ortholog's open chromatin status as a feature and treating species without usable orthologs as missing data. We first used k-means clustering with k = 9,000 (brain) or 12,000 (liver) to cluster the OCRs into small clusters; cluster centroid values for each species were defined as the mean of all OCR ortholog activity predictions for that species in the cluster. We then used affinity propagation clustering (Frey and Dueck 2007), implemented in scikit-learn (Pedregosa and Varoquaux 2011), with preference = -0.6 to cluster the outputs of the k-means clustering, which we defined as the centroids of the small clusters, into larger clusters. In both clustering steps, we defined the distance between two OCRs as 1 minus the cosine similarity between the vectors of enhancer activity predictions in the species for which both had usable orthologs. We ultimately obtained 102 brain clusters and 103 liver clusters.

Identifying Enhancer Sets with More Overlap with Clusters than Expected by Chance:

To identify enhancer sets with more overlap with a cluster than expected by chance, we first obtained the enhancer sets from the Supplementary information of the relevant manuscripts. For the mouse neuron firing candidate enhancers, we used the regions in Tables S11, S12 and S13 from a recent paper that identified mouse enhancers associated with neuron firing (Beagan, et al. 2020) and used liftOver (Kent, et al. 2003) to map the coordinates from mm9 to mm10. For the human GABAergic neuron activity candidate enhancers, we used the regions in Supp Data 12 from a recent paper that identified human enhancers associated with neuron activity (Boulting, et al. 2021) and used liftOver to map the coordinates from hg19 to hg38 (Kent, et al. 2003). For the mouse liver regeneration candidate enhancers, we used various subsets of regions from Supplementary Table 1 from a recent paper that identified mouse enhancers associated with liver regeneration (Wang, et al. 2020); for each category of regions in **Supplementary table S19**, we required the FDR to be less than 0.05 and the log2 fold-change to be greater than 1 (for ↑) or less than 1 (for ↓). We then used LOLA (Sheffield and Bock 2015) to run a hypergeometric test to evaluate the statistical significance of the overlap of each enhancer set with each cluster from the relevant tissue that is open in the relevant species; our query was our enhancer set, our database was our list of bed files with the relevant cluster locations, our universe was our list of OCRs in the relevant species, tissue combination, and we set "redefineUserSets" to TRUE. We did a Bonferroni correction by multiplying all p-values by 391, which was the total number of tests.

We wanted to determine if our enrichment for overlaps with clusters with lineage-specific open chromatin or lack of open chromatin could be partially explained by OCRs overlapping enhancer sets having fewer usable orthologs or lower conservation. To evaluate this, we used bedtools (Quinlan and Hall 2010) to identify the OCRs from each relevant tissue, species combination overlapping each enhancer set. We next computed the number of usable orthologs, the average PhastCons (Siepel, et al. 2005) score for the peak summit +/- 250bp, and the average PhyloP (Pollard, et al. 2010) score for the peak summit +/- 250bp for all OCRs from the relevant tissue, species combination as well as for the OCRs overlapping each enhancer set. Then, for each enhancer set, we used a Wilcoxon rank-sum test to compare the numbers of usable orthologs and the PhastCons and PhyloP scores between the OCRs overlapping the enhancer set and the full set of OCRs. We considered the difference to not be statistically significant if the nominal p-value was greater than or equal to 0.05.

ACKNOWLEDGEMENTS

We would like to thank the other members of the Pfenning Lab for useful discussions and suggestions and J. Ma and Y. Zhang for feedback on the manuscript. We would also like to thank the members of the Paten Lab and the Zoonomia Project for assistance with using Cactus multi-species alignments and providing us with early access to these alignments. I.M.K. was supported by Carnegie Mellon University Computational Biology Department Lane Fellowship. I.M.K., M.E.W., and A.R.P. were supported by the Alfred P. Sloan Foundation Research Fellowship. I.M.K., D.E.S., M.E.W., A.J.L., A.R.B., M.K., and A.R.P were supported by the NIH NIDA DP1DA046585. A.J.L. was supported by the NSF Graduate Research Fellowship Program under grants DGE1252522 and DGE1745016. M.K. was supported by The Carnegie Mellon Neuroscience Institute Presidential Fellowship. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), through the Pittsburgh Supercomputing Center Bridges Compute Cluster, which was supported by National Science Foundation grant number TG-MCB190067. Most animal silhouettes used in figures were obtained from PhyloPic (http://phylopic.org/).

DATA AVAILABILITY

Mouse liver ATAC-seq data in this manuscript can be accessed in Gene Expression Omnibus accession GSE159815 with token gzkbqwyobvmvfsd. Code and cluster images can be found at https://github.com/pfenninglab/OCROrthologPrediction, and additional code and cluster information can be obtained from the authors upon request.

AUTHORS' CONTRIBUTIONS

I.M.K. and A.R.P. designed the study with assistance from M.E.W. A.J.L. collected the mouse liver ATAC-seq data with assistance from A.R.B. I.M.K. processed the open chromatin and histone modification data with assistance from M.E.W. I.M.K. did the machine learning model training, machine learning model evaluation, and other computational method implementation except for the clustering. I.M.K. and A.R.P. designed the machine learning model evaluation metrics with assistance from M.K. and M.E.W. I.M.K. and D.E.S. developed the clustering method with assistance from M.K. and A.R.P., and D.E.S. implemented the clustering method. I.M.K. wrote the manuscript with assistance from D.E.S. and A.J.L. All authors reviewed and helped revise the manuscript.

FIGURE CAPTIONS

Figure 1: OCR Ortholog Open Chromatin Status Prediction Framework Overview

We trained a convolutional neural network (CNN) for predicting brain open chromatin using sequences underlying brain open chromatin region (OCR) orthologs in a small number of species and used the CNN to predict brain OCR ortholog open chromatin status across all of the species in the Zoonomia Project.

Figure 2: Lineage-Specific OCR Accuracy Evaluation – Performance of Models Trained on Brain OCRs and Different Negative Sets

a) Performance on test sets from negative sets used in model training. **b)** Performance on mouse brain test set open chromatin regions (OCRs) whose orthologs in at least one other species are not brain OCRs and mouse brain test set non-OCRs whose orthologs in at least one other species are brain OCRs. **c)** Performance on macaque brain OCRs whose orthologs in mouse are not brain OCRs and are on test set chromosomes and macaque brain non-OCRs whose orthologs in mouse are brain OCRs and are on test set chromosomes. **d)** Performance on Euarchonta-specific brain OCRs and Euarchonta-specific brain non-OCRs whose mouse orthologs are on test set chromosomes. *'s indicate the species from which sequences were obtained for making predictions. AUC stands for area under the receiver operating characteristic curve, AUPRC stands for area under the precision-recall curve, Rep. stands for repeat, and Orths. stands for orthologs.

Figure 3: Tissue-Specific OCR Accuracy Evaluation – Predictions from Models Trained on Brain OCRs on Brain, Liver OCRs

We made test set predictions on brain open chromatin regions (OCRs) that do not overlap liver OCRs, brain OCRs that overlap liver OCRs, liver OCRs that do not overlap brain OCRs, and negative sets. p-Values were computed with a Wilcoxon rank-sum test and a Bonferroni correction across all negative sets. Rep. stands for repeat, and Orths. stands for orthologs.

Figure 4: Phylogeny-Matching Correlations Evaluation – Divergence from Mouse versus Predictions in Glires

We compared the relationship between the phylogenetic distance between each Glires species and mouse and the mean brain open chromatin region (OCR) ortholog open chromatin status prediction across each Glires species. The yellow curve is the best fit exponential function of the form $y = ae^{bx}$. The yellow dotted line is the average prediction across the negatives in the test set. MYA stands for millions of years ago.

Figure 5: Multi-Species Model Performance

a) Performance of multi-species brain model on full test set, subset of test set consisting of clade-specific brain open chromatin regions (OCRs) and non-OCRs, subset of test set consisting of species-specific brain OCRs and non-OCRs, and subset of positive test set consisting of OCRs shared between brain and liver as well as subset of liver OCRs from test set genomic regions that do not overlap brain OCRs. AUC stands for area under the receiver operating characteristic curve, and AUPRC stands for area under the precision-recall curve. **b)** Performance of multi-species liver model on full test set, subset of test set consisting of clade-specific liver OCRs and non-OCRs, subset of test set consisting of species-specific liver OCRs and non-OCRs, and subset of positive test set consisting of OCRs shared between liver and brain as well as subset of brain OCRs from test set genomic regions that do not overlap liver OCRs. We reported area under the negative predictive value (NPV)-specificity (Spec.) curve instead of the area under the precision-recall curve (AUPRC) because these test sets have more positives than negatives. AUC stands for area under the receiver operating characteristic curve.

c) Divergence from mouse versus mean multi-species brain model predictions across mouse brain OCR orthologs in Glires. The green curve is the best fit exponential function of the form $y = ae^{bx}$. The green dotted line is the average prediction across test set negatives. MYA stands for millions of years ago. **d)** Divergence from mouse versus mean multi-species liver model predictions across mouse liver OCR orthologs in Glires. The red curve is the best fit exponential function of the form $y = ae^{bx}$. The red dotted line is the average prediction across the negatives in the test set. MYA stands for millions of years ago. **e)** Performance of mouse-only liver model versus multi-species liver model on Laurasiatheria-specific liver enhancers and Laurasiatheria-specific liver non-enhancers whose mouse orthologs are in the test set. AUC stands for area under the receiver operating characteristic curve, and AUPRC stands for area under the precision-recall curve.

Figure 6: Mean Conservation Score or Open Chromatin Status Prediction Rankings versus Open Chromatin Conservation – Macaque

a) Cumulative distribution functions (CDFs) of the reverse rankings of the PhastCons and PhyloP mouse mean conservation scores and the multi-species brain and liver models' open chromatin status predictions of the macaque orthologs of the mouse open chromatin regions (OCRs) whose open chromatin status is and is not conserved. The black dashed line is what we would expect from random rankings, and the blue dashed arrows show what we would expect from rankings that perfectly correspond to open chromatin status conservation. **b)** 7-week-old mouse cortex and striatum and macaque orofacial motor cortex ("Cortex") and putamen ("Striatum") open chromatin signal for a mouse brain OCR that is 50,328 bp away from the Stx16 transcription start site (TSS). **c)** 7-week-old mouse cortex and striatum and macaque orofacial motor cortex ("Cortex") and putamen ("Striatum") open chromatin signal for a mouse brain OCR that is 144,474 bp away from the Lnpk TSS. **d)** Our mouse liver and macaque liver open chromatin signal for a mouse liver OCR that is 24,814 bp away from the Rxra TSS. **e)** Our mouse liver and macaque liver open chromatin signal for a mouse liver OCR that is 154,404 bp away from the Fn1 TSS. In **b-e**, regions are mouse cortex open chromatin peak summits +/-250bp and their macaque orthologs, signals are from pooled reads across biological replicates, and liver H3K27ac ChIP-seq data comes from E-MTAB-2633 (Villar, et al. 2015).

Figure 7: Examples of Brain OCR Clusters with Predicted Lineage-Specific Open Chromatin Associated with Mouse Neuron Firing and Human Neuron Activity

We clustered the brain open chromatin region (OCRs), where the features were the brain predictions in each Boreoeutherian species from Zoonomia, and then identified clusters whose regions had significant overlap with regions associated with mouse neuron firing and human neuron activity. Mouse neuron firing enhancers had significant overlap with a predicted Murinae-specific OCR cluster (cluster 43), and human neuron activity enhancers had significant overlap with a predicted Euarchonta-, Perissodactyla- and Carnivora-specific non-OCR cluster (cluster 74).

REFERENCES

Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, Blanco-Aguiar JA, Plyusnina IZ, Trut L, Villafuerte R, et al. 2012. A comparison of brain gene expression levels in domesticated and wild animals. PLoS Genet 8:e1002962.

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33:831-838.

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et al. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature 587:246-251.

Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome Res 22:1723-1734.

Beagan JA, Pastuzyn ED, Fernandez LR, Guo MH, Feng K, Titus KR, Chandrashekar H, Shepherd JD, Phillips-Cremins JE. 2020. Three-dimensional genome restructuring across timescales of activity-induced neuronal gene expression. Nat Neurosci 23:707-717.

Berger MJ, Wenger AM, Guturu H, Bejerano G. 2018. Independent erosion of conserved transcription factor binding sites points to shared hindlimb, vision and external testes loss in different mammals. Nucleic Acids Res 46:9299-9308.

Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. 2018. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nat Ecol Evol 2:152-163.

Boulting GL, Durresi E, Ataman B, Sherman MA, Mei K, Harmin DA, Carter AC, Hochbaum DR, Granger AJ, Engreitz JM, et al. 2021. Activity-dependent regulome of human GABAergic neurons reveals new patterns of gene regulation and neurological disease heritability. Nat Neurosci 24:437-448.

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

Breuss MW, An N, Song Q, Nguyen T, Stanley V, James KN, Musaev D, Chai G, Wirth SA, Anzenberg P, et al. 2018. Mutations in LNPK, Encoding the Endoplasmic Reticulum Junction Stabilizer Lunapark, Cause a Recessive Neurodevelopmental Syndrome: The American Journal of Human Genetics. American Journal of Human Genetics 103:296-304.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods 10:1213-1218.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol 109:21.29.21-29.

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523:486-490.

Carbone L, Harris RA, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. Nature 513:195-201.

Chaubey PM, Hofstetter L, Roschitzki B, Stieger B. 2016. Proteomic Analysis of the Rat Canalicular Membrane Reveals Expression of a Complex System of P4-ATPases in Liver. PLoS One 11:e0158033.

Chen G, Wang R, Chen H, Wu L, Ge RS, Wang Y. 2016. Gossypol ameliorates liver fibrosis in diabetic rats induced by high-fat diet and streptozocin. Life Sci 149:58-64.

Chen L, Fish AE, Capra JA. 2018. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. PLoS Comput Biol 14:e1006484.

Chen L, Liu Z, Zhou B, Wei C, Zhou Y, Rosenfeld MG, Fu X-D, Chisholm AD, Jin Y. 2016. CELF RNA binding proteins promote axon regeneration in C. elegans and mammals through alternative splicing of Syntaxins. eLife.

Chua CEL, Tang BL. 2009. Syntaxin 16 is enriched in neuronal dendrites and may have a role in neurite outgrowth. Molecular Membrane Biology 25:35-45.

Damas J, Corbo M, Lewin HA. 2021. Vertebrate Chromosome Evolution. Annu Rev Anim Biosci 9:1-27.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

Fraser HB, Khaitovich P, Plotkin JB, Pääbo S, Eisen MB. 2005. Aging and gene expression in the primate brain. PLoS Biol 3:e274.

Frey BJ, Dueck D. 2007. Clustering by passing messages between data points. Science 315:972-976.

Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru MD, Reach SM, Motl J, Ehrlich ME, Hurd YL, Roussos P. 2018. An atlas of chromatin accessibility in the adult human brain. Genome Res 28:1243-1252.

Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov AV, Yim SH, Buffenstein R, Lee SR, Chang KT, Rhee H, et al. 2015. Gene expression defines natural changes in mammalian lifespan. Aging Cell 14:352-365.

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

Giuffra E, Tuggle CK, Consortium F. 2019. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annu Rev Anim Biosci 7:65-88.

Govaere O, Cockell S, Van Haele M, Wouters J, Van Delm W, Van den Eynde K, Bianchi A, van Eijsden R, Van Steenbergen W, Monbaliu D, et al. 2019. High-throughput sequencing identifies aetiology-dependent differences in ductular reaction in human chronic liver disease. J Pathol 248:66-76.

Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J. 2009. A novel connectionist system for unconstrained handwriting recognition. IEEE Trans Pattern Anal Mach Intell 31:855-868.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W. 2007. Quantifying similarity between motifs. Genome Biology 8:R24.

Halstead MM, Kern C, Saelao P, Wang Y, Chanthavixay G, Medrano JF, Van Eenennaam AL, Korf I, Tuggle CK, Ernst CW, et al. 2020. A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. BMC Genomics 21:698.

Hernandez-Gea V, Friedman SL. 2011. Pathogenesis of liver fibrosis. Annu Rev Pathol 6:425-456.

Hickey G, Paten B, Earl D, Zerbino D, Haussler D. 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics 29:1341-1342.

Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G. 2012. A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Rep 2:817-823.

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al. 2013. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res 41:827-841.

John S, Sabo PJ, Canfield TK, Lee K, Vong S, Weaver M, Wang H, Vierstra J, Reynolds AP, Thurman RE, et al. 2013. Genome-scale mapping of DNase I hypersensitivity. Curr Protoc Mol Biol Chapter 27:Unit 21.27.

Kalmbach BE, Buchin A, Long B, Close J, Nandi A, Miller JA, Bakken TE, Hodge RD, Chong P, de Frates R, et al. 2018. h-Channels Contribute to Divergent Intrinsic Membrane Properties of Supragranular Pyramidal Neurons in Human versus Mouse Cerebral Cortex. Neuron 100:1194-1208.e1195.

Kelley DR. 2020. Cross-species regulatory sequence activity prediction. PLoS Comput Biol 16:e1008050.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. 2014. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A 111:6131-6138.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100:11484-11489.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler aD. 2002. The Human Genome Browser at UCSC. In. Genome Research. p. 996-1006.

Kim J, Farré M, Auvil L, Capitanu B, Larkin DM, Ma J, Lewin HA. 2017. Reconstruction and evolutionary history of eutherian chromosomes. Proc Natl Acad Sci U S A 114:E5379-E5388.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107-116.

Klein JC, Keith A, Agarwal V, Durham T, Shendure J. 2018. Functional characterization of enhancer evolution in the primate lineage. Genome Biol 19:99.

Koepfli KP, Paten B, O'Brien SJ, Scientists GKCo. 2015. The Genome 10K Project: a way forward. Annu Rev Anim Biosci 3:57-111.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol 34:1812-1819.

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 518:317-330.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature 521:436-444.

LeCun Y, Jackel LD, Boser B, Denker JS, Graf HP, Guyon I, Henderson D, Howard RE, Hubbard W. 1989. Handwritten digit recognition: applications of neural network chips and automatic learning. IEEE Communications Magazine 27:41-46.

Liu C, Wang M, Wei X, Wu L, Xu J, Dai X, Xia J, Cheng M, Yuan Y, Zhang P, et al. 2019. An ATAC-seq atlas of chromatin accessibility in mouse tissues. Sci Data 6:65.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471:216-219.

Megquier K, Genereux DP, Hekman J, Swofford R, Turner-Maier J, Johnson J, Alonso J, Li X, Morrill K, Anguish LJ, et al. 2019. BarkBase: Epigenomic Annotation of Canine Genomes. Genes (Basel) 10:43.

Minnoye L, Taskiran II, Mauduit D, Fazio M, Van Aerschot L, Hulselmans G, Christiaens V, Makhzami S, Seltenhammer M, Karras P, et al. 2020. Cross-species analysis of enhancer logic using deep learning. Genome Res 30:1815-1834.

Mookherjee D, Majumder P, Mukherjee R, Chatterjee D, Kaul Z, Das SD, Sougrat R, Chakrabarti S, Chakrabarti O. 2019. Cytosolic aggregates in presence of non-translocated proteins perturb endoplasmic reticulum structure and dynamics. Traffic 20:943-960.

NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 46:D8-D13.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet 39:730-732.

Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL. 2017. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. Elife 6:e25884.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011. Cactus: Algorithms for genome multiple sequence alignment. Genome Res 21:1512-1528.

Pedregosa F, Varoquaux G. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825-2830.

Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell PV, Ganapathy G, et al. 2014. Convergent transcriptional specializations in the brains of humans and song-learning birds. Science 346:1256846.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110-121.

Quang D, Xie X. 2016. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res 44:e107.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Ray M, Zhang W. 2010. Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks. BMC Systems Biology 4:1-11.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47.

Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond AM, Ramachanderan R, Harewood L, Odom DT, Flicek P. 2021. LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. Genome Biol 22:62.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves n, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148:335-348.

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 328:1036-1040.

Sheffield NC, Bock C. 2015. LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics 32:587-589.

Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG, Corbo JC. 2016. Massively parallel cis-regulatory analysis in the mammalian central nervous system. Genome Res 26:238-255.

Shen YY, Liang L, Li GS, Murphy RW, Zhang YP. 2012. Parallel evolution of auditory genes for echolocation in bats and toothed whales. PLoS Genet 8:e1002788.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LDW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15:1034-1050.

Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. 2014. Absence of a simple code: How transcription factors read the genome. Trends in Biochemical Sciences. 39:381-399.

Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, Baldarelli RM, Beal JS, Campbell J, Corbani LE, et al. 2019. The mouse Gene Expression Database (GXD): 2019 update. Nucleic Acids Res 47:D774-D779.

Spitz F, Gonzalez F, Duboule D. 2003. A Global Control Region Defines a Chromosomal Regulatory Landscape Containing the HoxD Cluster. Cell 113:405-417.

Stossi F, Dandekar RD, Johnson H, Lavere P, Foulds CE, Mancini MG, Mancini MA. 2019. Tributyltin chloride (TBT) induces RXRA down-regulation and lipid accumulation in human liver cells. PLoS One 14:e0224405.

Sudmant PH, Alexis MS, Burge CB. 2015. Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biol 16:287.

Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, Consortium BK. 2018. Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. Annu Rev Anim Biosci 6:23-46.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. Nature 489:75-82.

Tosches MA, Yamawaki TM, Naumann RK, Jacobi AA, Tushev G, Laurent G. 2018. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. Science 360:881-888.

Trut L, Oskina I, Kharlamova A. 2009. Animal evolution during domestication: the domesticated fox as a model. Bioessays 31:349-360.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. 2015. Tissue-based map of the human proteome. Science 347:1260419-1260419.

Vermunt MW, Tan SC, Castelijns B, Geeven G, Reinink P, de Bruijn E, Kondova I, Persengiev S, Bontrop R, Cuppen E, et al. 2016. Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. Nat Neurosci 19:494-503.

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer evolution across 20 mammalian species. Cell 160:554-566.

Wang AW, Wang YJ, Zahm AM, Morgan AR, Wangensteen KJ, Kaestner KH. 2020. The Dynamic Chromatin Architecture of the Regenerating Liver. Cell Mol Gastroenterol Hepatol 9:121-143.

Wang B, Ke W, Guang J, Chen G, Yin L, Deng S, He Q, Liu Y, He T, Zheng R, et al. 2016. Firing Frequency Maxima of Fast-Spiking Neurons in Human, Monkey, and Mouse Neocortex. Front Cell Neurosci 10:239.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell 158:1431-1443.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206-216.

Xue Y, Guo C, Hu F, Zhu W, Mao S. 2020. PPARA/RXRA signalling regulates the fate of hepatic non-esterified fatty acids in a sheep model of maternal undernutrition. Biochim Biophys Acta Mol Cell Biol Lipids 1865:158548.

Young RL, Ferkin MH, Ockendon-Powell NF, Orr VN, Phelps SM, Pogány Á, Richards-Zawacki CL, Summers K, Székely T, Trainor BC, et al. 2019. Conserved transcriptomic profiles underpin monogamy across vertebrates. Proc Natl Acad Sci U S A 116:1331-1336.

Zhang R, Wang Y, Li R, Chen G. 2015. Transcriptional Factors Mediating Retinoic Acid Signals in the Control of Energy Metabolism. Int J Mol Sci 16:14210-14244.

Zhang X, Kaplow IM, Wirthlin M, Park TY, Pfenning AR. 2020. HALPER facilitates the identification of regulatory element orthologs across species. Bioinformatics 36:4339-4340.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12:931-934.

Zhu Y, Sousa AMM, Gao T, Skarica M, Li M, Santpere G, Esteller-Cucala P, Juan D, Ferrández-Peral L, Gulden FO, et al. 2018. Spatiotemporal transcriptomic divergence across human and macaque brain development. Science 362:eaat8077.

Zoonomia Consortium. 2020. A comparative genomics multitool for scientific discovery and conservation. Nature 547:240-245.

Figure 1

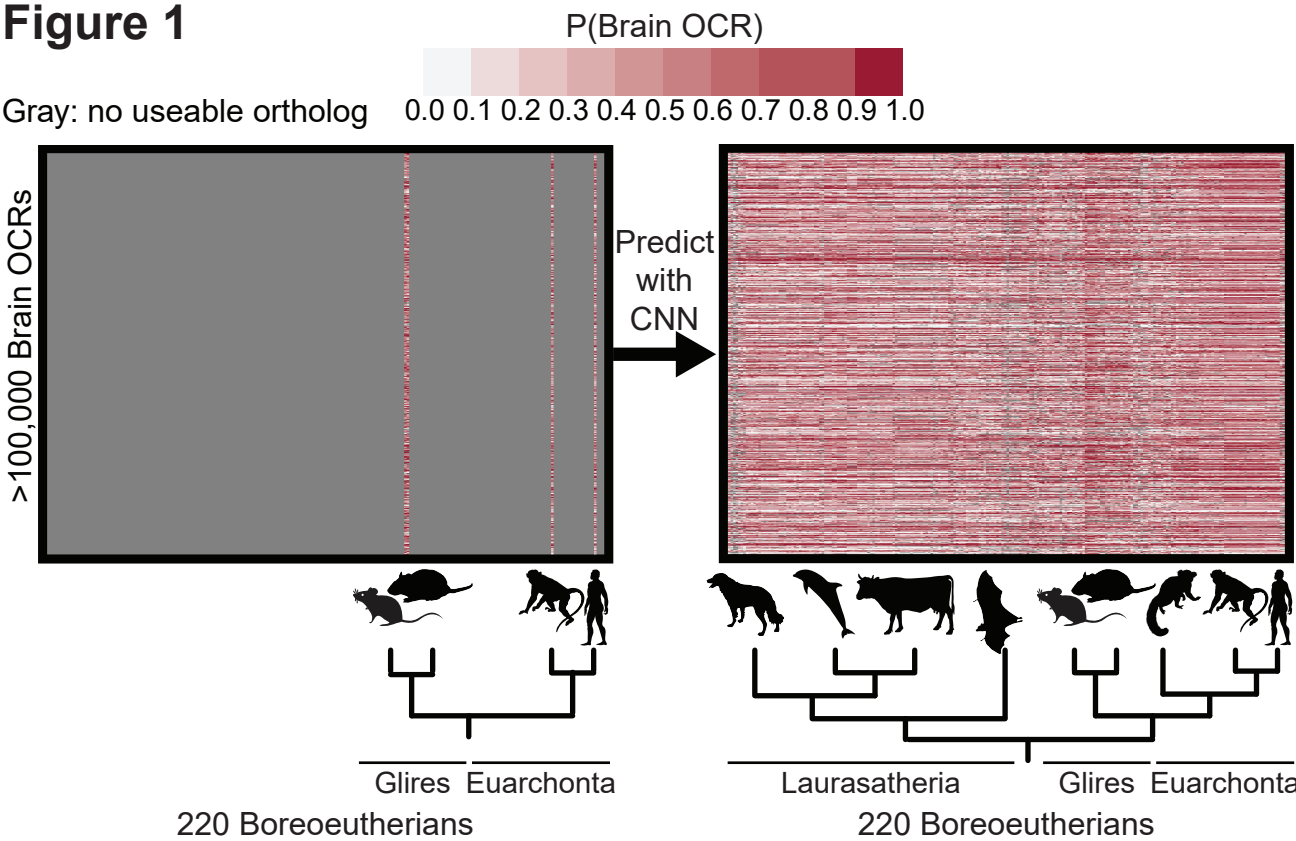
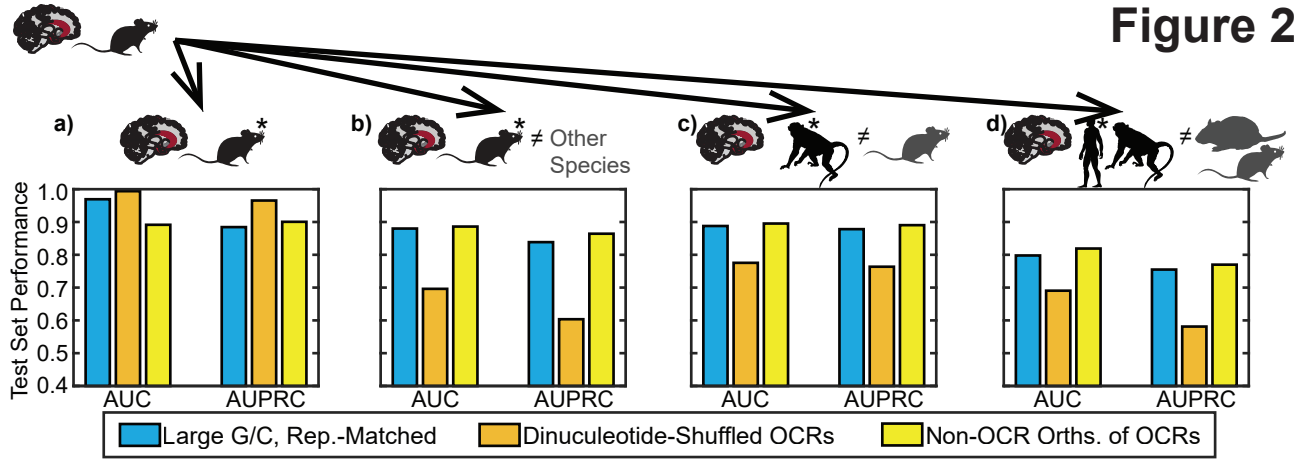

P(Brain OCR)
Gray: no useable ortholog
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
>100,000 Brain OCRs
Predict with CNN
Glires
Euarchonta
220 Boreoeutherians
Laurasatheria
Glires
Euarchonta
220 Boreoeutherians

Figure 2


a)
b)
c)
d)
*
≠ Other Species
≠
Test Set Performance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
AUC
AUPRC
Large G/C, Rep.-Matched
Dinuculeotide-Shuffled OCRs
Non-OCR Orths. of OCRs

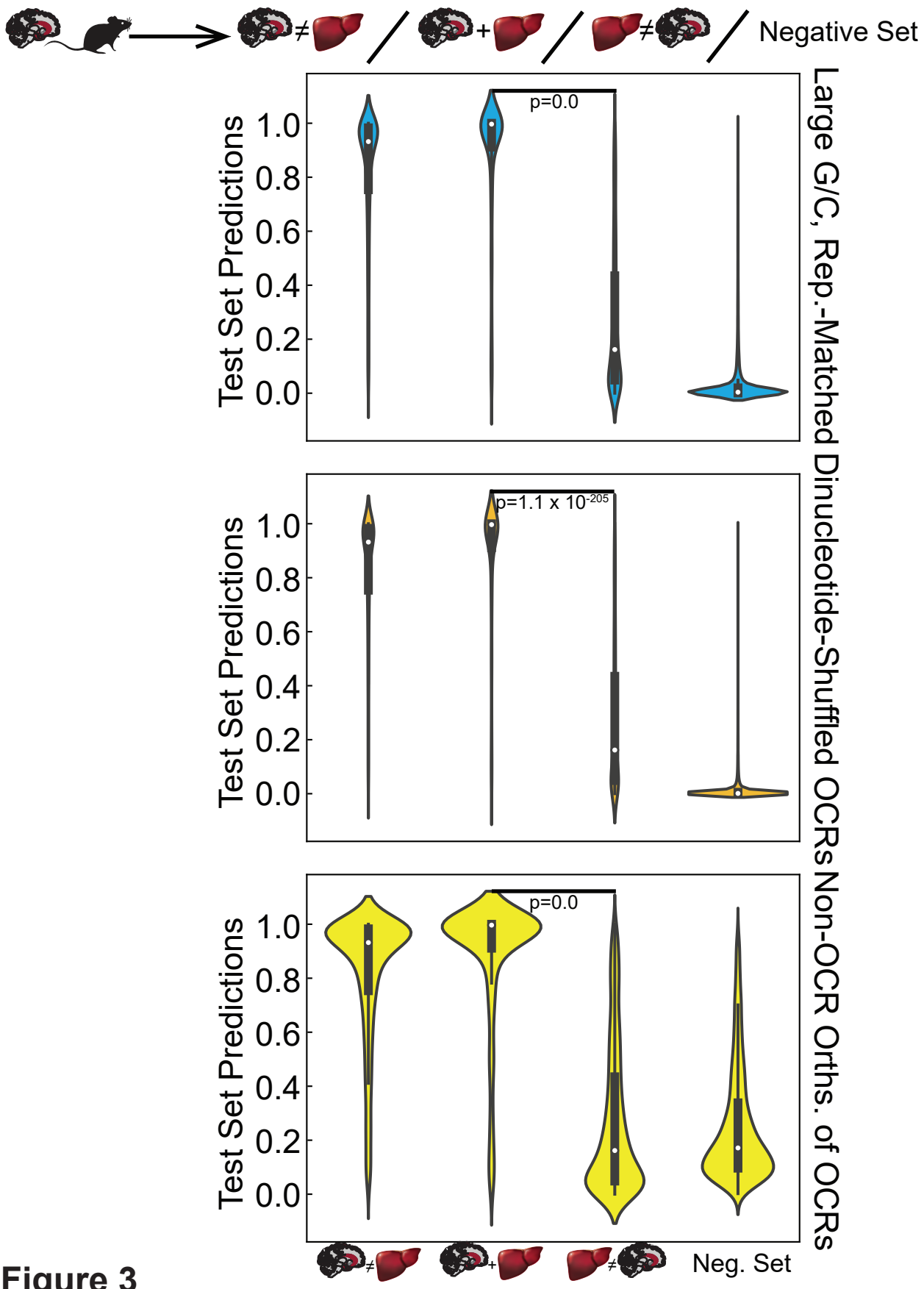

Negative Set
Large G/C, Rep.-Matched Dinucleotide-Shuffled OCRs
Non-OCR Orths. of OCRs
Test Set Predictions
1.0
0.8
0.6
0.4
0.2
0.0
p=0.0
p=1.1 x 10-205
p=0.0
Neg. Set

Figure 3

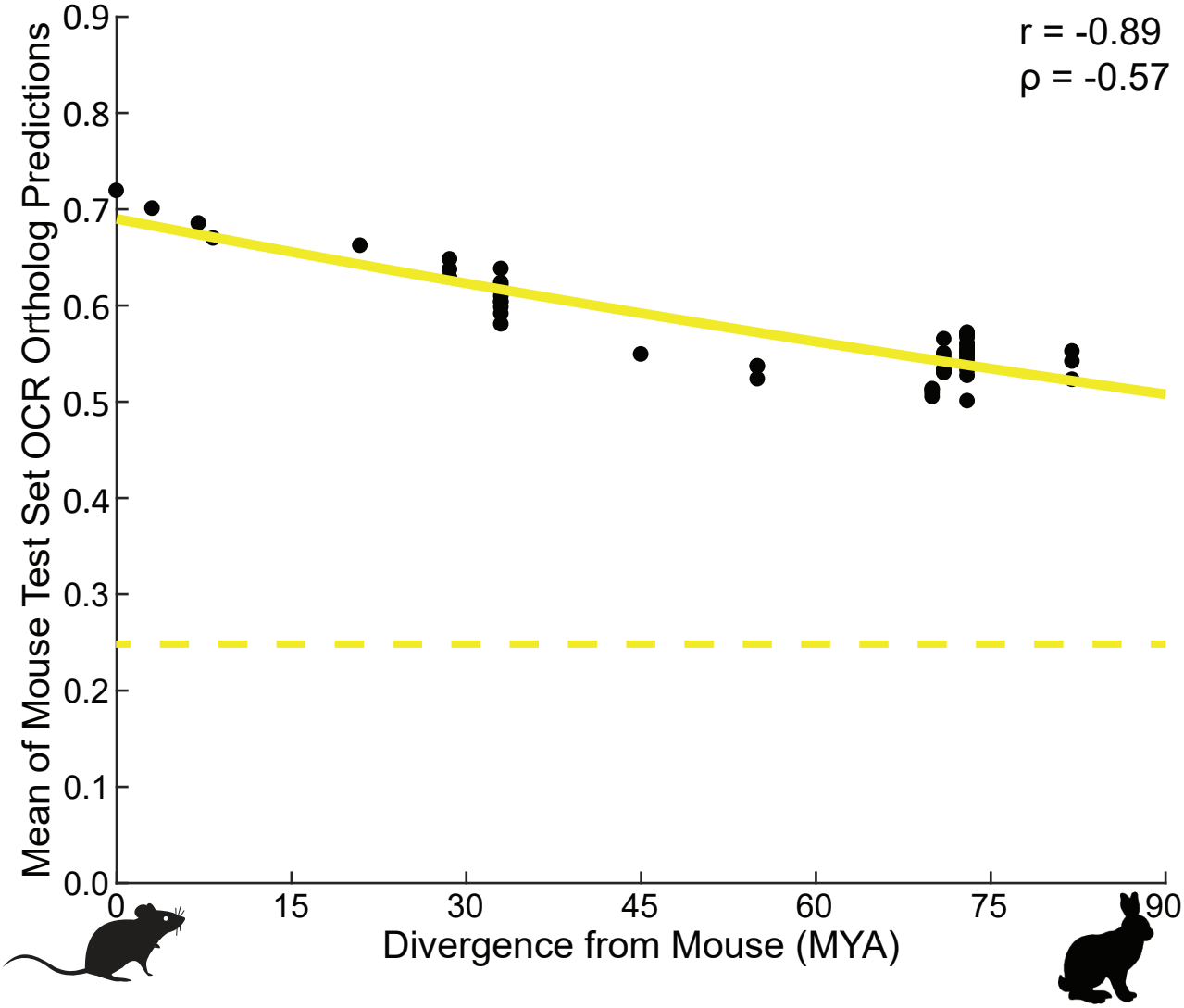

r = -0.89
ρ = -0.57
Mean of Mouse Test Set OCR Ortholog Predictions
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
15
30
45
60
75
90
Divergence from Mouse (MYA)

Figure 4

Figure 5

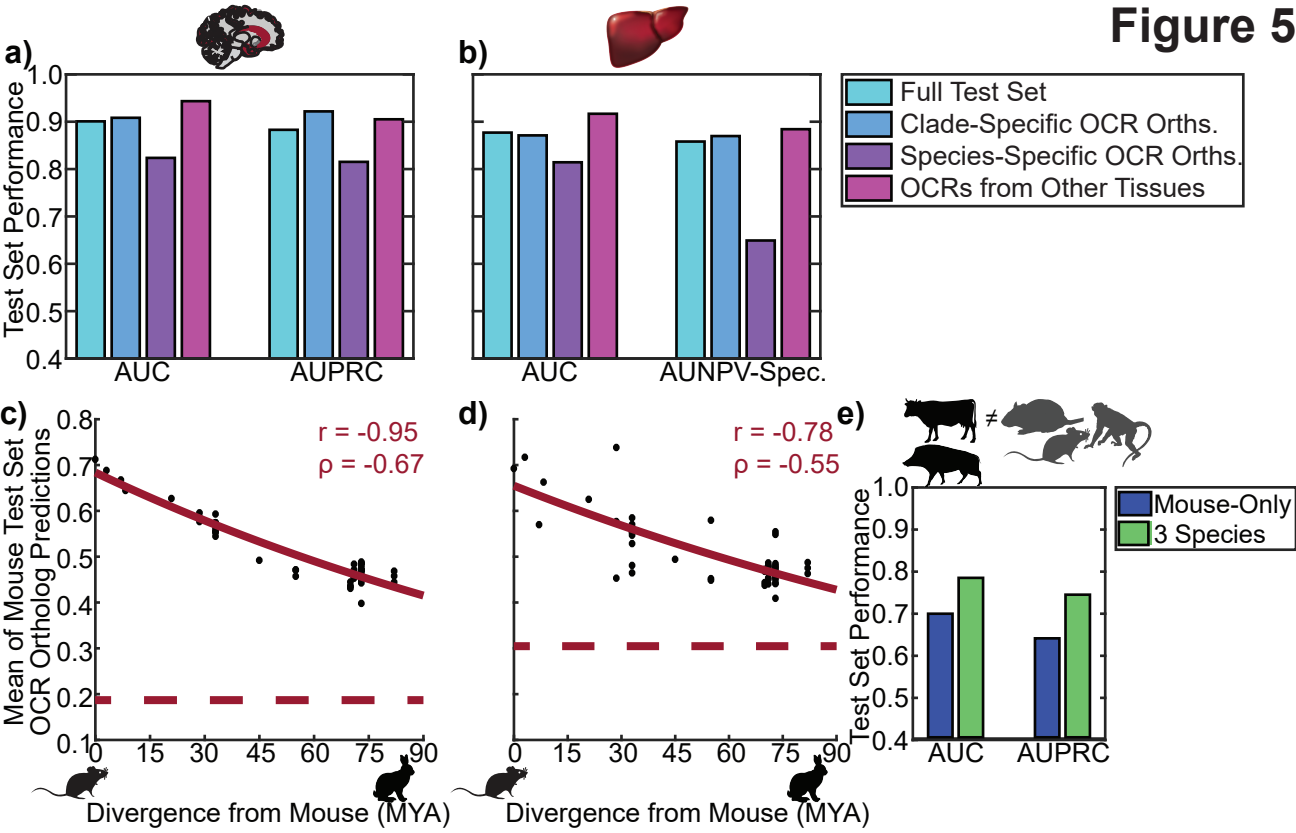
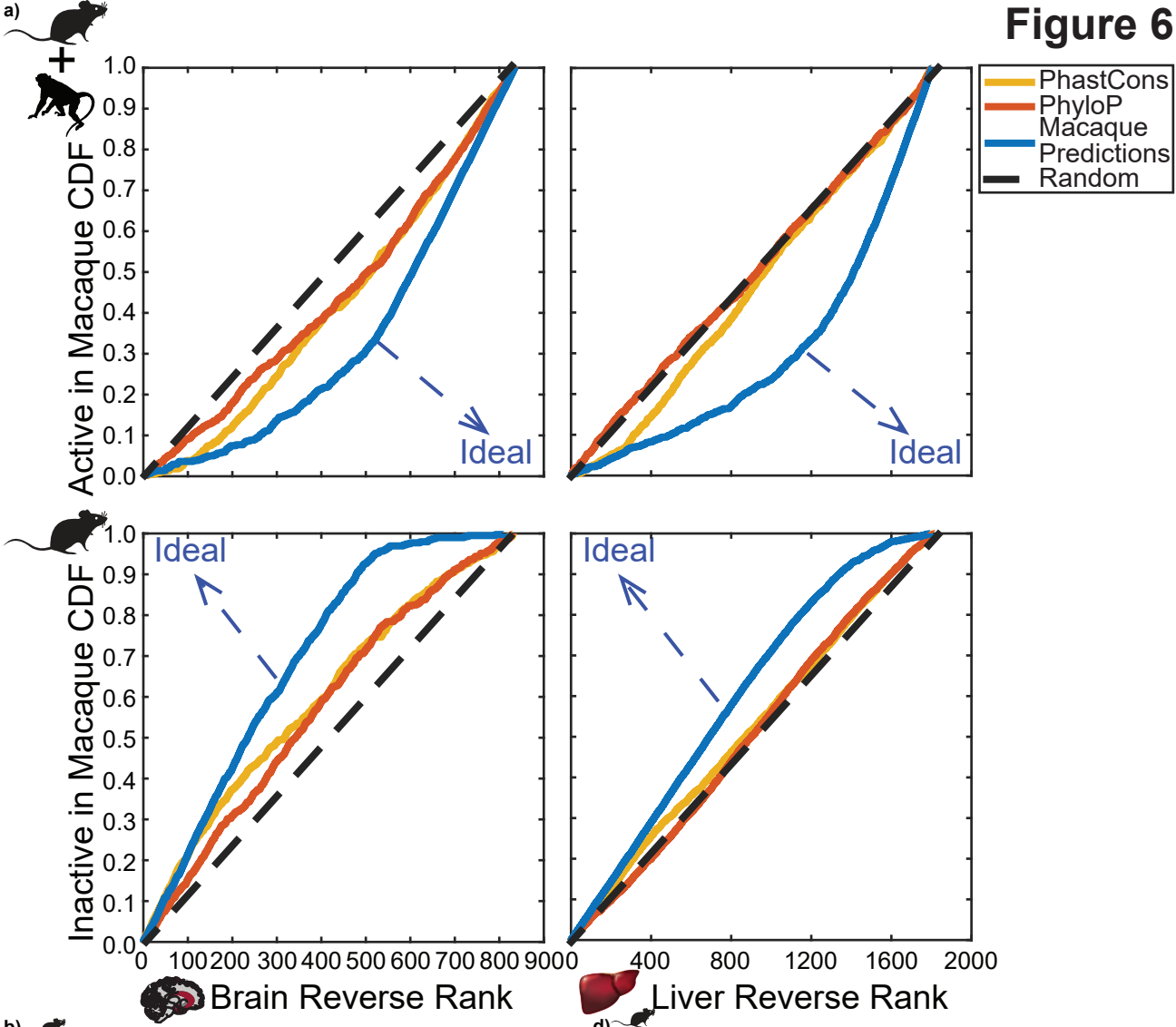
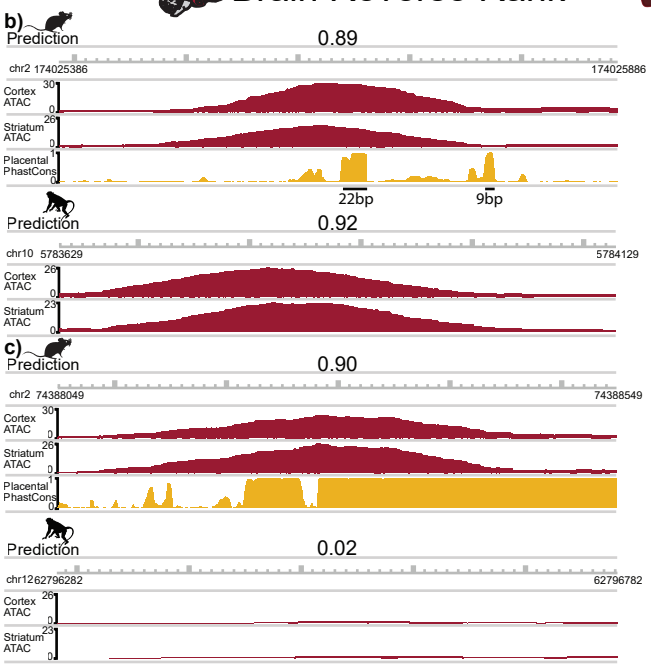

a)
b)
Test Set Performance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
AUC
AUPRC
AUNPV-Spec.
Full Test Set
Clade-Specific OCR Orths.
Species-Specific OCR Orths.
OCRs from Other Tissues
c)
d)
e)
Mean of Mouse Test Set OCR Ortholog Predictions
r = -0.95
ρ = -0.67
r = -0.78
ρ = -0.55
0
15
30
45
60
75
90
Divergence from Mouse (MYA)
≠
Mouse-Only
3 Species

Figure 6


a)
Active in Macaque CDF
Inactive in Macaque CDF
Ideal
Ideal
Ideal
Ideal
PhastCons
PhyloP
Macaque
Predictions
Random
Brain Reverse Rank
Liver Reverse Rank


b)
Prediction
0.89
chr2 174025386
174025886
Cortex ATAC
Striatum ATAC
Placental PhastCons
22bp
9bp
Prediction
0.92
chr10 5783629
5784129
Cortex ATAC
Striatum ATAC
c)
Prediction
0.90
chr2 74388049
74388549
Cortex ATAC
Striatum ATAC
Placental PhastCons
Prediction
0.02
chr12 62796282
62796782
Cortex ATAC
Striatum ATAC

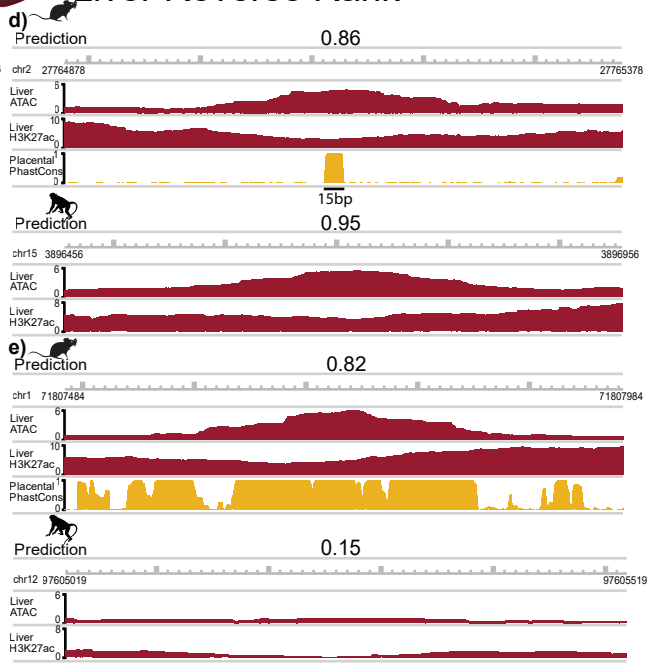
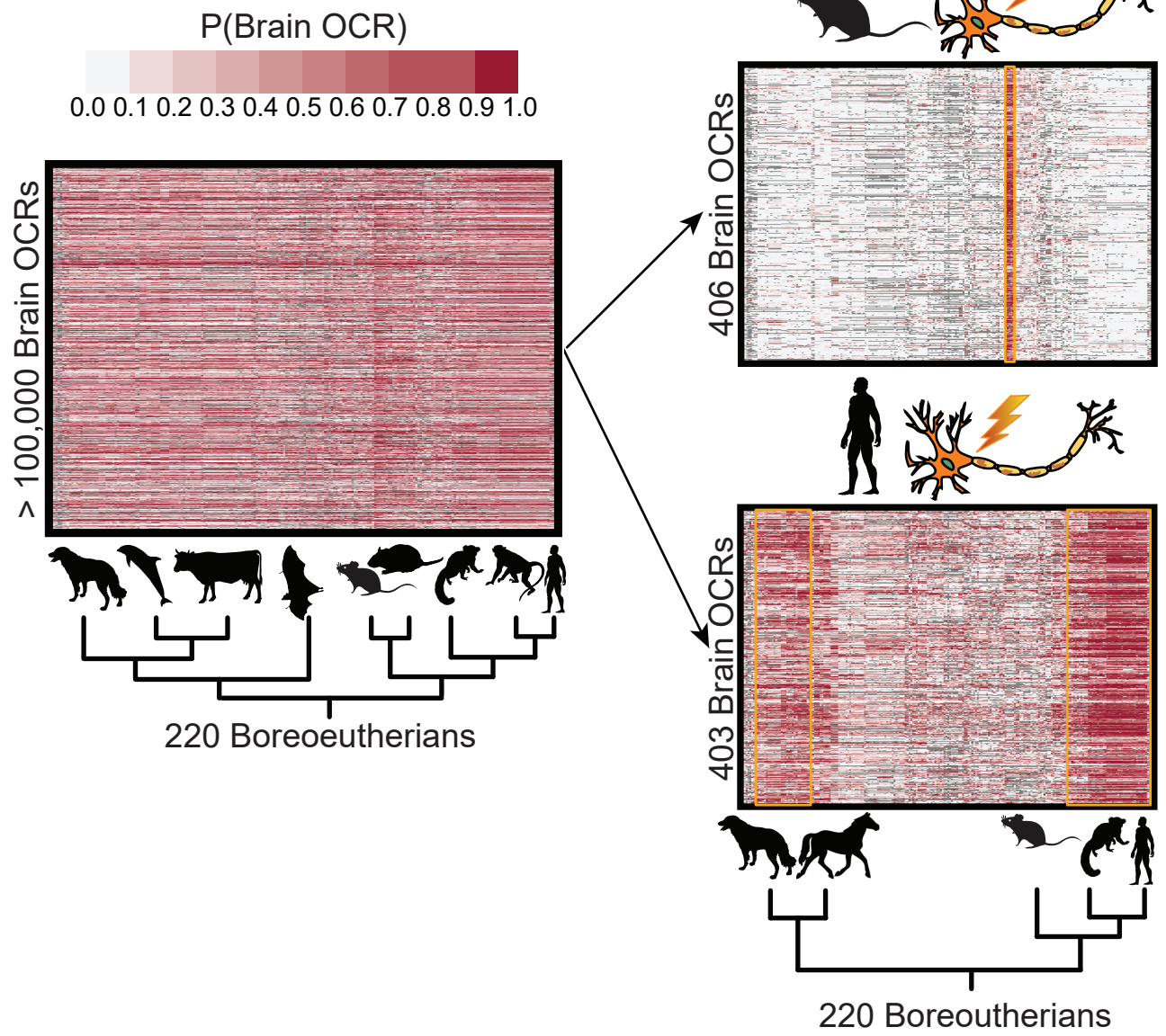

d)
Prediction
0.86
chr2 27764878
27765378
Liver ATAC
Liver H3K27ac
Placental PhastCons
15bp
Prediction
0.95
chr15 3896456
3896956
Liver ATAC
Liver H3K27ac
e)
Prediction
0.82
chr1 71807484
71807984
Liver ATAC
Liver H3K27ac
Placental PhastCons
Prediction
0.15
chr12 97605019
97605519
Liver ATAC
Liver H3K27ac

Figure 7


P(Brain OCR)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> 100,000 Brain OCRs
220 Boreoeutherians
406 Brain OCRs
403 Brain OCRs
220 Boreoutherians