TITLE

Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin

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

Background: Evolutionary conservation is an invaluable tool for inferring functional significance in the genome, including regions that are crucial across many species and those that have undergone convergent evolution. Computational methods to test for sequence conservation are dominated by algorithms that examine the ability of one or more nucleotides to align across large evolutionary distances. While these nucleotide alignment-based approaches have proven powerful for protein-coding genes and some non-coding elements, they fail to capture conservation at many enhancers, distal regulatory elements that control spatio-temporal patterns of gene expression. The function of enhancers is governed by a complex, often tissue- and cell type-specific, code that links combinations of transcription factor binding sites and other regulation-related sequence patterns to regulatory activity. Thus, function of orthologous enhancer regions can be conserved across large evolutionary distances, even when nucleotide turnover is high.

Results: We present a new machine learning-based approach for evaluating enhancer conservation that leverages the combinatorial sequence code of enhancer activity rather than relying on the alignment of individual nucleotides. We first train a convolutional neural network model that is able to predict tissue-specific open chromatin, a proxy for enhancer activity, across mammals. Then, we apply that model to distinguish instances where the genome sequence would predict conserved function versus a loss regulatory activity in that tissue. We present criteria for systematically evaluating model performance for this task and use them to demonstrate that our models accurately predict tissue-specific conservation and divergence in open chromatin between primate and rodent species, vastly out-performing leading nucleotide alignment-based approaches. We then apply 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.

Conclusion: 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.

KEYWORDS

Gene expression evolution, open chromatin prediction, machine learning, enhancers

BACKGROUND

The study of conservation has had a tremendous impact in multiple areas of mammalian biology. When a new genome is sequenced, conservation is applied to provide high-quality annotations of candidate exons, introns, promoters, and other likely functional genomic regions [1]. Regions of the human genome conserved across other primates or mammals show a stronger enrichment for disease-associated loci than any other evaluated category of regions [2]. Conversely, regions of the human genome that display accelerated evolution have been implicated in human-specific adaptation [3]. In endangered species, molecular conservation has been applied predict which regions of low heterozygosity may be impacting fitness [4]. Conservation has also been applied to find regions of the genome associated with the evolution of complex phenotype across mammals and vertebrates more broadly, including the loss of limb function [5], loss of eyesight [6], and longevity [7]. The powers of these studies are still growing, with many consortia, including the Vertebrate Genomes Project [8], the Genome 10K Project [9], the Bat 1K Project [10], and the Zoonomia Project [4], sequencing, assembling, and aligning [11] genomes from

59 hundreds of mammals, including endangered species and species that live in remote parts of the world.
60 Using these data, we can investigate conservation by comparing the DNA sequences of species whose
61 most recent common ancestors lived tens of millions of years ago.

62 The methods used to infer conservation across species, including those applied to many of the
63 challenges described above, typically rely on nucleotide-level constraint. They generally begin by aligning
64 the nucleotides of multiple genomes together [12, 13]. Building off of those alignments, PhyloP calculates
65 nucleotide-level constraint relative a neutral model of evolution, which can be aggregated across broader
66 regions to identify signatures of conservation or acceleration [3, 13, 14]. To study the diversity of
67 mammalian phenotypes, nucleotide alignments are often modeled in the context of a tree structure to
68 look for signatures of positive or negative selection [15-17]. Once nucleotide-level selection has been
69 inferred, additional techniques have been applied to link those patterns of selection to convergent
70 evolution, instances where a specific phenotype has evolved independently in multiple lineages [6, 17,
71 18].

72 As new genomic resources have become available and computational techniques have advanced,
73 it has become clear that a large component of phenotypic evolution is mediated by differences in *cis*-
74 regulatory elements, many of which are enhancers that control gene expression [19-21]. Within the
75 human population, enhancers show a strong enrichment for disease-associated genetic variants [22, 23].
76 Across species, nucleotide-level selection in enhancers has been associated the loss of eyesight,
77 hindlimbs, and external testes [6, 18, 24, 25]. Numerous other complex phenotypes have been linked to
78 gene expression differences across species, including vocal learning [26], domestication [27, 28], longevity
79 [29-31], brain size [32], echolocation [33], and monogamy [34]. While nucleotide-level selection in
80 enhancers is being applied to study the evolution of some of these phenotypes [35], recent studies of
81 enhancers across and species suggest a model of evolution in which nucleotide-level conservation of
82 enhancers can be low in spite of enhancers maintaining their function [36, 37].

83 Much of our knowledge of enhancers comes from regulatory genomics measurements that are
84 associated with enhancer activity, especially the Assay for Transposase-Accessible Chromatin using
85 Sequencing (ATAC-Seq) and the DNase hypersensitivity assay for open chromatin and chromatin
86 immunoprecipitation sequencing (ChIP-Seq) for the histone modifications H3K27ac and H3K4me1 [38-41].
87 Studies involving these assays have demonstrated that enhancers, relative to genes, are substantially
88 more tissue- or cell type-specific and generally less conserved across species [42-45]. Within a given cell
89 type or tissue, a combinatorial code of transcription factor (TF) binding motifs and other sequence
90 patterns determines the ultimate regulatory activity [46-48]. In a striking example of this, a recent study
91 found that an Islet enhancer's developmental function remains remarkably conserved across the 700
92 million years of evolution between mammals and sponges by maintaining a similar set of TF motifs despite
93 negligible detectable conservation at the nucleotide level [36]. This understanding is further supported
94 by studies of TF binding across species that display a large turnover in individual binding sites [49-51],
95 even though gene expression is often highly conserved, and genes with highly conserved expression often
96 have conservation in the TFs whose motifs occur within their candidate enhancers [52]. Therefore, to
97 study conservation of enhancers, new methods are required that link genome sequence differences at
98 candidate enhancers to differences in enhancer function.

99 Advances in the application of machine learning techniques to regulatory genomics open up the
100 possibility that conservation could by assayed at the level of the regulatory code, rather that the alignment
101 of individual nucleotides. In one recent study, support vector machines (SVMs) and convolutional neural
102 networks (CNNs) were able to predict which 3kb windows have the enhancer-associated histone
103 modification H3K27ac in brain, liver, and limb tissue of human, macaque, and mouse [53]. Importantly,
104 the study found that models trained in one mammal achieved high accuracy in another mammal in the
105 same clade and in another mammal in a different clade, suggesting that the regulatory code in all three
106 of these tissues is highly conserved across mammals [54]. Another study obtained a similar result for

107 regions associated with H3K27ac [55], and two other studies have obtained similar results using another
108 proxy for enhancer activity – open chromatin regions (OCRs) [56, 57]. One of these studies found that
109 training CNNs on OCRs from multiple mammals had better performance than training CNNs on OCRs from
110 a single mammal, albeit using 131,072bp sequences as input. The boost in power from incorporating
111 multiple species generalized to predicting TF binding strength from ChIP-seq data and gene expression
112 from RNA-seq data [56]. An additional study found that a combined CNN-recurrent neural network [58,
113 59] trained on sequences underlying 500bp OCRs from melanoma cell lines in one species can accurately
114 predict melanoma cell line open chromatin in other species at a wide range of genetic distances from the
115 training species, including in parts of the genome with low sequence conservation between the training
116 and evaluation species [57].

117 While these studies represent major advances in cross-species enhancer prediction, they have yet
118 to comprehensively demonstrate an ability to identify sequence differences between species that are
119 associated with differences in regulatory genomic measurements of enhancer activity, which is crucial for
120 their application as a conservation metric (**Table 1**). In fact, an additional study trained SVMs to predict
121 liver enhancers using dinucleotide-shuffled candidate enhancers as negatives. While the overall
122 performance was good, human enhancers whose orthologs are active in Old World monkeys but not New
123 World monkeys were predicted to have consistent activity across all primates, showing that models with
124 good overall performance do not always work well on enhancer orthologs whose activity differs between
125 species [60].

126 To predict and evaluate candidate *cis*-regulatory element differences across species, we chose to
127 focus on OCRs, which are only a proxy of regulatory activity, but whose high resolution has the potential
128 to better identify specific genome sequence differences associated with putative regulatory activity. We
129 leveraged new, controlled open chromatin experiments conducted by our laboratory [61] and trained a
130 set of new models to predict OCR differences across species (**Figure 1**). We evaluated model performance

131 using new criteria that we developed for this task, which focus on the ability to predict similarities and
132 differences across species and tissues rather than the large number of regions that are consistently open
133 or closed (**Supplemental Figure 1a**). We also developed a novel method for associating previously
134 identified candidate enhancer sets with predicted lineage-specific open chromatin and lack of open
135 chromatin, in which we clustered OCRs based on their predicted open chromatin and identified clusters
136 overlapping candidate enhancer sets more than expected by chance. Our methods for evaluating
137 approaches to OCR ortholog open chromatin status prediction and for identifying predicted lineage-
138 specific OCRs associated with candidate enhancer sets can be applied to any tissue or cell type with open
139 chromatin data from multiple species. This allows for reliable annotation of open chromatin across the
140 hundreds of new genomes that are being sequenced, providing a valuable resource for research
141 communities studying gene regulation in non-traditional model organisms, especially where direct
142 measurements are not feasible. We anticipate that this work will encourage researchers to develop and
143 properly evaluate new models for predicting OCR ortholog open chromatin across species and help reveal
144 potential functional roles of lineage-specific enhancers, enabling us to uncover transcriptional regulatory
145 mechanisms underlying the evolution of mammalian phenotypic diversity.

146

147

148 **RESULTS**

149

150 <u>Dataset Construction for Evaluating Approaches for OCR Ortholog Open Chromatin Prediction</u>

151 To demonstrate the ability to predict OCR differences across species, we created a dataset of brain OCRs
152 and their orthologs across dozens of species. To do this, we gathered open chromatin data generated by

153 ATAC-seq [41, 62] or DNase hypersensitivity [63] from two brain regions – cortex and striatum – in four
154 species: *Homo sapiens* [38, 64, 65], *Macaca Mulatta* [61], *Mus musculus* [66], and *Rattus norvegicus* [61].
155 We then defined our OCRs to be the 250bp in each direction of summits of non-exonic cortex open
156 chromatin peaks that (1) overlap striatum open chromatin peaks, (2) are less than 1kb (to exclude super-
157 enhancers), and (3) are at least 20kb from the nearest transcription start site (TSS) so that so they would
158 not overlap promoters. We also identified the orthologs of each OCR in each of the other species collected
159 by the Zoonomia Project [4, 67] (**Figure 1**). Rather than focus on longer regions, which have been shown
160 to work well in some previous studies [54, 56], we used 500bp regions because of the ability shorter
161 regions provide to focus on the impact of local sequence differences, because of the relative ease of
162 obtaining orthologs of shorter regions in genomes with short scaffolds, and because predictions of
163 enhancer activity for shorter regions are easier to experimentally validate.

164 We compared performance for two measures of conservation scores – PhastCons [13] and PhyloP
165 [14] – and machine learning models trained using five different negative sets that are similar to those used
166 for related tasks: (1) flanking regions [68], (2) OCRs from other tissues [56, 57], (3) about ten times as
167 many G/C- and repeat-matched regions as positives [54], (4) about twice as many G/C- and repeat-
168 matched regions as positives, and (5) ten dinucleotide-shuffled versions of each positive [69]
169 (**Supplemental Figure 1b**). We additionally created a sixth, novel negative set to force the model to learn
170 signatures of OCRs whose orthologs' open chromatin status differs between species – sequences with
171 closed chromatin in brain of a given species whose orthologs in another species are brain OCRs (we called
172 this "non-OCR orthologs of OCRs," which are the white regions in left part of **Figure 1**) – and included this
173 negative set in our comparison (**Supplemental Figure 1b**). For a modeling approach, we chose CNNs [53,
174 70] because they can model complex combinatorial relationships between sequence patterns, and
175 changes in a single TF motif often do not cause changes in open chromatin [48]; because they do not
176 require an explicit featurization of the data, and many sequence patterns involved in brain open

177 chromatin remain to be discovered; and because they can make predictions quickly relative to SVMs, the
178 other leading approach for related tasks [54]. We did the comparison using conservation scores
179 generated using mouse reference-based alignments [12, 71] and models trained on only mouse sequences
180 so that we could evaluate their performance on both closely and distantly related species not used for
181 training the machine learning models [54].

182

183 Optimizing and Evaluating Methods for Achieving Lineage-Specific OCR Accuracy

184 We evaluated the machine learning model trained on each negative set on its corresponding test
185 set. We found that all models worked well, with every model achieving an AUC > 0.85 and an AUPRC >
186 0.7 (**Supplemental Figure 1c**). This performance is especially impressive given that the ratios of the
187 number of negatives to the number of positives ranged from approximately 1.2:1 (non-OCR orthologs of
188 OCRs) to approximately 20:1 (OCRs in other tissues). The best-performing model was the model with the
189 negative set consisting of dinucleotide-shuffled brain OCRs (**Supplemental Figure 1c**). However, in this
190 comparison, each model was evaluated on a different negative set, so this evaluation may not be
191 indicative of how useful each model would be in answering questions about gene expression evolution.

192 We therefore also evaluated each model's and the conservation scores' lineage-specific OCR
193 accuracy (**Supplemental Figure 1a**), which we did by evaluating models and computing conservation
194 scores for the OCR orthologs whose brain open chromatin status differs between species, an evaluation
195 not done in any previous studies (**Table 1**). We did this by evaluating whether a model trained in one
196 species could make accurate predictions of brain open chromatin status on sequences obtained from
197 another species [54]. However, unlike other previous work that evaluated overall performance [54], we
198 evaluated performance on the subset of regions in another species whose brain open chromatin status
199 differs from that of their orthologs in the training species, as this is necessary for showing that models can

200 accurately predict conservation – differences and similarities in brain open chromatin status between
201 species. We therefore evaluated our models trained in mouse on macaque brain OCRs whose mouse
202 orthologs are closed and on macaque brain closed chromatin regions whose mouse orthologs are open in
203 brain and compared our results to conservation scores. We found that conservation scores tended to be
204 slightly lower for both of these sets of regions than they for OCRs whose orthologs in both mouse and
205 macaque are open, but they could not clearly distinguish between whether an OCR's ortholog was open
206 in mouse or in macaque (**Figure 2a**). On the other hand, all models achieved decent performance (AUC >
207 0.65, AUPRC > 0.55) (**Supplemental Figure 1g**), though the model trained on the dinucleotide-shuffled
208 brain OCR negatives predicted that almost half of the macaque non-OCRs whose mouse orthologs are
209 OCRs are open in brain (**Figure 2a**), causing it to have the worst performance (**Supplemental Figure 1g**).
210 We also evaluated our models on human (**Supplemental Figure 1h, Supplemental Figure 2**) and rat
211 (**Supplemental Figure 1i, Supplemental Figure 3a**) regions with differing brain open chromatin statuses
212 from their mouse orthologs and found that the models generally did not work as well for such regions but
213 still obtained decent performance and that the relative performance of conservation scores and different
214 negative sets was similar to what it was for the macaque regions. We did additional evaluations to further
215 show that overall model performance is not indicative of lineage-specific OCR accuracy, that re-calibrating
216 the models with the positive training set and the training set from our novel negative set did not
217 substantially change the relative performances of the models trained on different negative sets, and that
218 conservation scores are significantly inferior to our best machine learning models at predicting open
219 chromatin conservation across lineages (**Supplemental Notes, Supplemental Tables 1-9, Supplemental**
220 **Figure 4**). These results demonstrate the necessity of evaluating models on OCRs whose orthologs' open
221 chromatin statuses differ between species.

222

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

224 In addition to evaluating conservation score and machine learning model performance on OCR
225 orthologs whose open chromatin status differs between species, we also determined if conservation
226 scores and our machine learning models achieved high tissue-specific OCR accuracy (**Supplemental Figure**
227 **1a**). While previous studies evaluated model performance in different tissues or cell lines [54, 56, 57] and
228 showed that models trained in the evaluation tissue tend to be more accurate than models trained in
229 other tissues [54], no study has directly evaluated model performance on tissue-specific candidate
230 enhancers or candidate non-enhancers (**Table 1**). To determine if our models learned sequence patterns
231 associated with only brain-specific open chromatin, we evaluated the conservation scores and our
232 models' predictions for the subset of brain OCRs that do not overlap liver OCRs and the subset of brain
233 OCRs that do overlap liver OCRs. Test set predictions from all models for both of these subsets of the
234 positive set were usually close to one (**Figure 2a, Supplemental Figures 2-3, Supplemental Figure 5a**). To
235 determine if our models learned only sequence patterns that are indicative of general open chromatin,
236 we also evaluated our models' predictions for the liver OCRs that do not overlap brain OCRs. We
237 compared this to the predictions on the negative set. We found that predictions on both of these sets
238 tended to be close to zero. However, the liver, non-brain open chromatin status predictions from the
239 model trained with dinucleotide-shuffled OCR negatives tended to be more evenly distributed between
240 zero and one than the liver, non-brain open chromatin status predictions for the models trained with the
241 other negative sets (**Figure 2a, Supplemental Figures 2-3, Supplemental Figure 5a**). In addition, unlike
242 the machine learning model predictions, the conservation scores for brain, non-liver OCRs; OCRs in brain
243 and liver; and liver, non-brain OCRs are similar to each other (**Figure 2a, Supplemental Figures 2-3**),
244 demonstrating that conservation scores, in contrast to most of our machine learning models, provide little
245 information about the tissue-specificity of open chromatin. We also did a comparison of the
246 performances of models trained on different negative sets in which we limited the positive set to brain
247 OCRs that overlap liver OCRs and defined the negative set as liver OCRs that do not overlap brain OCRs.

248 We found that all models worked well (AUC > 0.75, AUPRC > 0.6) on mouse as well as on macaque and
249 rat, which were not used in training, with the model trained on dinucleotide-shuffled brain OCR negatives
250 providing the worst performance (**Supplemental Figure 5b**). The difference in performance may be a
251 result of differences between the TF motifs learned by the models (**Supplemental Notes, Supplemental**
252 **Figure 6**). In addition, we found that the models that tended to work better on positives were more
253 accurate for shared brain and liver OCRs than for liver, non-brain OCRs and that the models that tended
254 to work better on brain closed chromatin orthologs of brain OCRs were more accurate for liver, non-brain
255 OCRs than for shared brain and liver OCRs (**Supplemental Tables 10-11**, **Supplemental Figure 5c**).
256 Calibration with the positive training set and the training set from our novel negative set had a similar
257 effect to calibration for clade-specific open and closed chromatin regions (**Supplemental Tables 10-12**,
258 **Supplemental Figure 5c**). These results show that conservation scores cannot be used directly to
259 understand the tissue-specificity of open chromatin and that models with good overall performance are
260 not always able to comprehensively learn open chromatin sequence patterns that are specific to the tissue
261 in which they were trained.

262

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

264 We also determined whether our models' predictions have phylogeny-matching correlations in a
265 way that does not require open chromatin data from multiple species. To do this, we obtained the
266 orthologs of the mouse brain OCRs in all of the fifty-six Glires species in the Zoonomia Project [4, 67], used
267 our machine learning models to predict the brain open chromatin statuses of these orthologs, computed
268 the mean brain open chromatin statuses across all brain OCR orthologs in each species, and computed
269 the correlation between mean predicted brain open chromatin status and evolutionary distance from
270 mouse. Although there are OCR orthologs, such as species-specific OCRs and OCRs with convergently

271 evolved open chromatin [6], whose open chromatin conservation across species is not associated with
272 phylogenetic distance, we think that such OCRs in most tissues are a minority due to the principle of
273 evolutionary parsimony and a previous study of enhancer activity across multiple species [44]. As
274 anticipated, all models showed a strong negative correlation between mean predicted brain open
275 chromatin status and divergence from mouse (**Supplemental Figure 7a**). Nevertheless, there is still more
276 open chromatin at these brain OCR orthologs than would be expected from brain non-OCRs, even in the
277 most distantly related Glires species, because all mean predictions are greater than the mean predictions
278 for the negative test sets (**Supplemental Figure 7a**). We also expected there to be a strong positive
279 correlation between the standard deviation of open chromatin status and divergence from mouse
280 because most brain OCR orthologs in species closely related to mouse are active in brain, while the brain
281 open chromatin status of brain OCR orthologs in species that are more distantly related should vary. We
282 found this expected positive correlation for the predictions from machine learning models trained on all
283 negative sets (**Supplemental Figure 7b**).

284

285

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

288 Based on other research showing that training models with data from multiple species can
289 improve OCR prediction accuracy [56], we trained additional machine learning models using open
290 chromatin data from multiple species. For each of brain and liver, we used the open chromatin data from
291 all of the species that we had collected as positives (four species for brain, three species for liver) and the
292 orthologs of all these OCRs in the other species for which we had data that did not overlap brain or liver
293 open chromatin, respectively, as negatives. Before training multi-species models, we trained mouse-only

294 models for liver and found that they achieved good performance for all of our criteria (**Figure 2b,**
295 **Supplemental Notes, Supplemental Figure 8, Supplemental Table 13**). We then trained brain and liver
296 multi-species models and found that they achieved high lineage-specific and tissue-specific OCR accuracy,
297 where performance was generally better than the performance for any of the models trained on only
298 mouse sequences (**Figure 2, Figures 3a-b**). In addition, we determined if the multi-species brain and liver
299 models' predictions had phylogeny-matching correlations by using them to predict the OCR ortholog open
300 chromatin status of mouse brain and liver OCRs, respectively, across Glires and found strong negative
301 correlations between divergence from mouse and mean OCR ortholog open chromatin status predictions
302 (**Figures 3c-d**). We also found strong positive correlations between divergence from mouse and standard
303 deviations of OCR ortholog open chromatin status predictions (**Supplemental Figures 9a-b**). For both
304 these models and our mouse-only models, our phylogeny-matching correlation results could not be fully
305 explained by genome quality (**Supplemental Notes, Supplemental Figure 10, Supplemental Tables 14-**
306 **15**). Finally, we evaluated our mouse-only and multi-species liver models on Laurasiatheria-specific liver
307 OCRs and liver non-OCRs [72], as no Laurasiatheria were used in training either model, and found that the
308 multi-species liver model had better performance than the mouse-only liver model (**Figure 3e**). Like our
309 mouse-only models, our multi-species brain and liver models also learned TF motifs for TFs that are known
310 to be involved in brain and liver, respectively (**Supplemental Notes, Supplemental Figures 9c-d**), and their
311 predictions were significantly more associated with open chromatin conservation than conservation
312 scores were (**Figure 2, Supplemental Notes, Supplemental Table 16**).

313 Some of the OCR orthologs for which our multi-species brain model correctly predicted brain open
314 chromatin conservation in spite of low mean sequence conservation or for which our models correctly
315 predicted lack of brain open chromatin conservation in spite of high mean sequence conservation are
316 near genes that have been shown to play important roles in the brain. For example, there is a region on
317 mouse chromosome 2 – part of our test set – that has low mean sequence conservation according to

318 PhastCons [13] and PhyloP [14] but high brain experimentally identified and predicted open chromatin
319 conservation between mouse and macaque (**Figure 4a**) and whose mouse and macaque orthologs are
320 located near the gene *Stx16*. This gene is involved in vesicle trafficking in most tissues, including the brain
321 [73], and may play a role in Alzheimer's disease [74]; in fact, its role in axon regeneration is conserved
322 between mammals and *C. elegans* [75]. Although this region near *Stx16* has generally low sequence
323 conservation, running TomTom [76] on the 22bp sequence with high conservation revealed a
324 subsequence that is similar to the *Fos* motif, which is also found in the macaque ortholog. Since our
325 machine learning model used sequence similarity to the *Fos* motif in making predictions (**Supplemental**
326 **Figure 9c**), the machine learning model was likely able to automatically determine that it should use this
327 sequence in making its prediction. In addition, there is a region on mouse chromosome 2 that has high
328 mean sequence conservation but low experimentally identified and predicted open chromatin
329 conservation between mouse and macaque (**Figure 4b**) and whose mouse and macaque orthologs are
330 located near the gene *Lnpk*. This gene is an endoplasmic reticulum junction stabilizer [77] that has been
331 shown to play a role in brain and limb development [78], and mutations in this gene have been associated
332 with neurodevelopmental disorders [79]. It is possible that this region of the genome near *Lnpk* has a
333 high degree of sequence conservation because it has functions in other tissues. These results
334 demonstrate the potential benefits of using OCR ortholog open chromatin status predictions instead of
335 mean conservation scores for studying the evolution of the expression of important genes in a tissue of
336 interest.

337 In addition, some of the accurate multi-species model liver open chromatin conservation
338 predictions disagreed with the mean conservation scores. For instance, there is a region on mouse
339 chromosome 2 that has high experimentally identified and predicted liver open chromatin conservation
340 but low sequence conservation (**Figure 4c**) and whose mouse and macaque orthologs are located near
341 *Rxra*. This gene is a TF involved in regulating lipid metabolism [80-82], TF-MoDISco identified a sequence

342 similar to its motif as being important in our liver models (**Supplemental Figure 8f, Supplemental Figure**
343 **9d**), and its liver expression is stable across fifteen mammals [83]. Although this region near *Rxra* has
344 generally low sequence conservation, the 15bp segment with high conservation is similar to the motif for
345 *Ctcf* according to TomTom [76], and that motif is also found in the macaque ortholog. Since our machine
346 learning model used sequence similarity to the *Ctcf* motif in making predictions (**Supplemental Figure 9d**),
347 the machine learning model was likely able to automatically determine that it should use this sequence in
348 making its prediction. There is also a region on mouse chromosome 1, which is part of our test set, whose
349 mouse ortholog is an OCR and macaque ortholog is not an OCR according to our data and predictions in
350 spite of being highly conserved (**Figure 4d**) and whose mouse and macaque orthologs are near *Fn1*. This
351 gene has been implicated in liver fibrosis [84-86], and a multi-species liver RNA-seq study found that it
352 has higher expression in mouse liver relative to livers of other mammals and birds and lower expression
353 in primate livers relative to livers of other mammals and birds [87]. For both of these OCRs, the H3K27ac
354 signal conservation in the same regions [37] is similar to the open chromatin status conservation,
355 suggesting that the open chromatin status conservation is indicative of enhancer activity conservation
356 (**Figures 4c-d**). These results suggest that using predicted OCR ortholog open chromatin status instead of
357 conservation has the potential to be beneficial for understanding gene expression evolution in multiple
358 tissues.

359

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

361 Since our models can accurately predict lineage-specific OCRs, we evaluated whether lineage-
362 specific brain and liver OCRs were associated with brain and liver functions. First, we used publicly
363 available RNA-seq data from the livers of thirteen placental mammals with high-quality genomes [83] to
364 identify genes with Rodent-specific expression; we obtained 15 genes with false discovery rate (FDR) <

365 0.05 [88, 89]. We then used our predictions to identify liver OCRs with predicted Rodent-specific activity.
366 One of the genes with Rodent-specific expression that is near OCRs with predicted liver-specific open
367 chromatin is Atp9a (**Supplemental Figure 11**). This gene has high expression in mouse liver according to
368 Mouse Genome Informatics [90], it has little or no RNA or protein expression in human liver according to
369 all of the studies referenced in the Human Protein Atlas [91, 92], and its protein is expressed in the rat
370 canicular membrane [93], a part of hepatocytes involved in bile production. This suggests that OCRs with
371 predicted lineage-specific open chromatin may be regulating genes with lineage-specific expression.

372 To further investigate the roles of predicted lineage-specific OCRs, we clustered each of the brain
373 and liver OCRs, where the features were the predicted activity in each of the Boreoeutheria from the
374 Zoonomia Project [4]. We first used k-means clustering to cluster the OCRs into thousands of small
375 clusters and then used affinity propagation cluster to cluster the smaller clusters into larger clusters. We
376 selected affinity propagation clustering because we did not know how many clusters to expect, and
377 affinity propagation clustering automatically determines the number of clusters. For each of brain and
378 liver, we obtained a little over one hundred clusters with different patterns of predicted open chromatin
379 across species (brain clusters: https://github.com/pfenninglab/OCROrthologPrediction/clusters/brain,
380 liver clusters: https://github.com/pfenninglab/OCROrthologPrediction/liver).

381 We then determined whether each brain cluster that was open in mouse overlapped mouse
382 candidate enhancers associated with neuron firing [94] more than expected by chance (**Supplemental**
383 **Table 17**) and each brain cluster that was open in human overlapped human candidate enhancers
384 associated with neuron activity [95] more than expected by chance (**Supplemental Table 18**).
385 Interestingly, the candidate enhancers from each of these sets intersected clusters with predicted lineage-
386 specific open chromatin or predicted lineage-specific lack of open chromatin more than expected by
387 chance. Specifically, mouse neuron bicuculline (Bic)-specific candidate enhancers, where Bic induces
388 neuron firing, were enriched for overlapping two predicted Murinae-specific brain open chromatin

389 clusters – cluster 43 (**Figure 5**) and cluster 27 (**Supplemental Figure 12a**). We think that these results are
390 unlikely to be explained by the number of usable orthologs or conservation because mouse brain OCRs
391 overlapping Bic-specific candidate enhancers do not have significantly fewer usable orthologs or lower
392 conservation according to PhastCons [13] or PhyloP [14] than mouse brain OCRs in general. In contrast,
393 mouse activity-invariant candidate enhancers were enriched for overlapping the predicted open
394 chromatin in all species cluster (cluster 1), a noisy cluster without a clear pattern of predicted open
395 chromatin (cluster 37), a noisy predicted Yangochioptera-specific brain non-open chromatin cluster
396 (cluster 81), and a noisy predicted Primate-specific brain non-open chromatin cluster (cluster 88).
397 Likewise, human candidate enhancers from GABAergic neurons made from hiPSCs from four genotypes
398 that had increased activity two hours after exposure to potassium chloride (KCl), where KCl induces
399 neuron activity, were enriched for overlapping a predicted Carnivora-, Perissodactyla- and Euarchonta-
400 specific brain open chromatin cluster (cluster 74, **Figure 5**), a predicted Hystricognathi-specific brain non-
401 open chromatin cluster (cluster 11, **Supplemental Figure 12b**), and a predicted Muroidea- and Pecora-
402 specific brain non-open chromatin (cluster 48, **Supplemental Figure 12b**). These results are unlikely to be
403 explained the number of usable orthologs or conservation because human brain OCRs overlapping this
404 set of GABAergic neuron candidate enhancers tended to have more usable orthologs and higher
405 conservation than human brain OCRs overall according to both PhastCons [13] and PhyloP [13]. In
406 contrast, candidate enhancers from the same source that had decreased activity two hours after exposure
407 to KCl were enriched for overlapping the predicted open chromatin in all species cluster (cluster 1) and a
408 predicted Ruminantia-specific brain non-open chromatin cluster (cluster 82). These results suggest that
409 there may be a relationship between enhancer response to neuron activity and whether the enhancer's
410 activity tends to be specific to the lineage in which it was identified.

411 We also investigated whether each liver cluster that was active in mouse overlapped mouse
412 candidate enhancers associated with liver regeneration [96] more than expected by chance

413 (**Supplemental Table 19**). We found that candidate liver enhancers that have increased activity four
414 weeks after hepatocyte repopulation relative to the control were enriched for overlapping a cluster with
415 predicted Murinae-specific open chromatin – cluster 29 – as well as two clusters with predicted Muroidea-
416 specific open chromatin – cluster 36 and cluster 100 (**Supplemental Figure 12c**). We think that these
417 results are unlikely to be explained by the number of usable orthologs or conservation because OCRs
418 overlapping this candidate enhancer set do not have significantly fewer usable orthologs or lower
419 conservation according to PhastCons [13] or PhyloP [14] than liver OCRs overall. We also found that
420 candidate liver enhancers that have increased activity one week after hepatocyte repopulation relative to
421 the control were enriched for overlapping a cluster with noisy predicted Primate-specific liver non-open
422 chromatin (cluster 83). In contrast to these findings, liver candidate enhancers with decreased activity
423 four weeks after hepatocyte repopulation relative to the control were not enriched for overlapping any
424 clusters, and liver candidate enhancers with decreased activity four weeks after hepatocyte repopulation
425 relative to one week after hepatocyte repopulation were enriched for overlapping a cluster without a
426 clear pattern of predicted open chromatin (cluster 39). These results suggest that there may be a
427 relationship between liver regeneration in Murinae and Murinae-specific open chromatin.

428

429

430 **DISCUSSION**

431

432 <u>Developing a Tissue-Specific Metric of Regulatory Conservation</u>

433 While there are well-established methods for quantifying nucleotide-level conservation based on a
434 sequence alignment [13, 14, 97], these methods have limited utility for the many enhancers with tissue-

435 specific activity [42, 43] or low sequence conservation in spite of high functional conservation [36]. Yet
436 quantifying enhancers' conservation can provide insight into their functional relevance, such as the
437 identification of convergent evolution in gene regulation underlying complex traits [20, 26, 29]. Here, we
438 developed a machine learning-based approach to measure tissue-specific enhancer conservation. First,
439 we trained CNNs to predict OCR orthologs' open chromatin statuses in the tissue from which the OCRs
440 were obtained. Then, we applied our models to predict the conservation of over 100,000 OCRs from each
441 of brain and liver in over 200 mammals.

442 Our approach vastly outperformed nucleotide alignment-based methods of conservation
443 PhastCons [13] and PhyloP [14] at predicting lineage- and tissue-specific open chromatin status. As
444 expected, we found many examples where nucleotide-level conservation is low but the predicted open
445 chromatin in a tissue of interest is conserved [36]. Conversely, we also identified cases where the
446 nucleotide-level conservation is high, but the few differences between species disrupt open chromatin in
447 a tissue of interest. We attribute the success of our method to our CNNs' abilities to learn a conserved
448 regulatory code linking genome sequence to tissue-specific open chromatin [54].

449 Our method builds upon recently published approaches that constructed machine learning
450 models that learned conserved tissue-specific regulatory codes across species [54-57]. Yet a model's
451 ability to predict regulatory genomics features across multiple species does not guarantee that the model
452 can accurately predict conservation or divergence. For example, a study predicting liver open chromatin
453 using SVMs showed overall high accuracy but was not able to accurately predict cases where sequence
454 differences between primates were associated with regulatory differences [60]. To demonstrate our
455 ability to predict regulatory conservation, we developed a new, systematic set of evaluation criteria that
456 evaluates if lineage- and tissue-specific differences can be predicted (**Table 1**). In the future, our criteria
457 could be used to evaluate a wide range of extensions to our approach (**Supplemental Notes**).

458

459 Limitations of Our Approach

460 Although our models excel in many criteria for their stated purpose of predicting lineage- and
461 tissue-specific open chromatin, our models' set-up also has inherent limitations that prevent them from
462 fully fulfilling their purposes. For example, many OCRs' open chromatin statuses are influenced by factors
463 beyond the 250 base pairs in each direction of their summits, and the fixed-size inputs required by CNNs
464 prevents us from modeling some of the long-range interactions that may influence open chromatin [98]
465 (**Supplemental Notes**). In addition, since we predict open chromatin status of only OCR orthologs, we
466 cannot identify OCRs whose orthologs are not OCRs in any species for which we have open chromatin
467 data (**Supplemental Notes**). We also treat open chromatin status as binary, where, in reality, in bulk open
468 chromatin data, open chromatin is a continuous signal (**Supplemental Notes**). Finally, our approach
469 requires high-quality open chromatin from at least two species and is likely to have stronger performance
470 when more are included. Cases where the datasets are not matched for sex, age, time of day, or other
471 factors could influence model performance (**Supplemental Notes**).

472

473 Applications of Inferring Conservation of Tissue-Specific Regulatory Activity

474 In addition to being accurate, our models' predictions revealed that many OCRs may have lineage-
475 specific patterns of open or closed chromatin and that these patterns may be functional. We identified a
476 predicted Rodent-specific liver OCR near Atp9a, a gene with Rodent-specific expression in the liver that
477 may play a role in bile production in rats [93]. In addition, we developed a novel approach to deciphering
478 the putative lineage-specificity of candidate enhancers identified in other studies in which we clustered
479 our OCRs using predicted activity as features and identified clusters with more overlap with these
480 candidate enhancers than expected by chance. Using this approach, we found that mouse candidate

481 enhancers activated during neuron firing [94] are enriched for overlapping clusters with predicted
482 Murinae-specific open chromatin, and human candidate enhancers activated during neuron activity [95]
483 are enriched for overlapping clusters with predicted Euarchonta-specific gains or Muroidea-specific losses
484 of open chromatin, suggesting that there may be Muroidea-specific enhancer signatures of Muroidea
485 neuron activity. Although no study, to our knowledge, has evaluated this for large numbers of species, a
486 few studies have compared neuron activity between mice and one or two Euarchonta and found striking
487 differences, including lower spiking frequencies in fast-spiking mouse cortical neurons relative to human
488 and rhesus macaque [99] and lack of expression of h-Channels in mouse excitatory neurons relative to
489 human [100]. We also found that candidate enhancers activated during liver regeneration [96] are
490 enriched for overlapping predicted Murinae-specific open chromatin. We found clusters with predicted
491 lineage-specific open chromatin for additional linages, predicted lineage-specific losses in open chromatin
492 for additional lineages, and predicted convergent gains and losses in open chromatin; investigating the
493 roles of the open chromatin regions in these clusters could be an exciting path towards identifying
494 mechanisms underlying linage-specific differences between mammalian brains and livers.

495 Our approach has the potential to be applied to numerous other groups of species, cell types, and
496 tissues. It does not require experimentally determined OCR data from more than a few species; it requires
497 only genomes from a large number of species, which are being generated in unprecedented quantities [4,
498 9, 10]. In addition to having many advantages relative to nucleotide-level conservation scores, our use of
499 500bp sequences for training our OCR ortholog open chromatin status prediction models, as compared
500 to other machine learning approaches that require sequences of 3kb [54] or over 100kb [56], enables us
501 to make predictions in species whose assemblies have short scaffolds, including many of the species in
502 the Zoonomia [4] and B10K datasets [101].

503 Our open chromatin conservation predictions can also be used in a forward genomics approach
504 [102] to help identify the mechanisms underlying the evolution of the expression of genes of interest or

505 of phenotypes that have evolved through gene expression. This can be done by identifying OCR orthologs
506 whose changes in predicted open chromatin status correspond to changes in gene expression or
507 phenotypes. Many multi-species gene expression datasets in tissues with data from an assay that can
508 serve as a proxy for enhancers are publicly available [37, 83, 103, 104], and more will likely be generated
509 in the near future [105]. Additional multi-species single-cell RNA-seq datasets are being generated from
510 some of these tissues [106, 107] and will likely soon be supplemented by single-cell ATAC-seq. Many of
511 these tissues have been associated with phenotypes that have evolved through gene expression [26, 29],
512 so we can use the gene expression data along with predictions from models like ours to gain insights how
513 these phenotypes evolved. Such insights may also reveal mechanisms underlying diseases associated with
514 these phenotypes [57].

515

516

517 **CONCLUSIONS**

518

519 The lineage and tissue-specificity of many enhancers limits our ability to quantify the conservation of
520 enhancers through nucleotide-level conservation scores. Therefore, rather than focusing on identifying
521 cases where natural selection is acting on individual nucleotides, we used open chromatin data from
522 tissues of interest to train CNNs for predicting open chromatin and showed that they are able to identify
523 cases where natural selection operates to maintain the combination of TF binding sites that are needed
524 to regulate gene expression. We evaluated the success of our CNNs as well as nucleotide conservation-
525 based and other machine learning-based methods using new criteria that we designed explicitly for
526 evaluating open chromatin conservation prediction. We then used CNNs to predict brain and liver open
527 chromatin conservation across mammals and found that candidate enhancers associated with neuron

528 firing and liver regeneration tended to overlap regions of predicted lineage-specific open or closed
529 chromatin. Our approaches to quantifying enhancer conservation and evaluating methods for this task
530 can be applied to any tissue or cell type with enhancer data available from multiple species. As this is, to
531 our knowledge, the first study to predict open chromatin conservation across more than a few species,
532 we anticipate that our work will serve as a foundation for identifying OCRs whose open chromatin
533 conservation is involved in gene expression evolution, providing insights into transcriptional regulatory
534 mechanisms underlying phenotypes that have evolved through gene expression and their associated
535 diseases.

536

537

538 **METHODS**

539

540 Constructing Positive Sets

541 Data generation and processing is described in **Supplemental Methods**. Mouse experiments were
542 approved by Carnegie Mellon University's Institutional Animal Care and Use Committee. To obtain OCRs
543 in each species, we intersected the IDR "optimal set" reproducible peaks from each of the brain regions
544 and datasets for brain and each of the liver datasets for liver and defined OCRs to be the intersected peaks
545 that are likely to be enhancers. Specifically, for each species, we selected one set of reproducible open
546 chromatin peaks to be the "base peaks," used bedtools intersect with the -wa and -u options to intersect
547 it with each of the other reproducible peak sets in series, and then used bedtools closestBed with the -t
548 first and -d options to identify the "base peaks" that overlapped at least one peak from each other set
549 that were over 20kb from the nearest protein-coding TSS (not promoters), at most 1kb long (not super-

550 enhancers), and non-exonic [108]. The base peaks for human brain were the IDR "optimal set" from
551 NeuN+ cells in the primary motor cortex from GSE96949 [65], for the macaque brain were the IDR
552 "optimal set" from the orofacial motor cortex from data we previously generated [61], for the macaque
553 liver were the IDR reproducible peaks across self-pseudo replicates from the first macaque liver replicate
554 from data we previously generated [61], for mouse brain were the IDR "optimal set" from the cortex from
555 the seven-week-old mouse from data we previously generated [66], for mouse liver were the IDR "optimal
556 set" from our mouse liver ATAC-seq dataset, for the rat brain were the IDR "optimal set" from the primary
557 motor cortex from data we previously generated [61], and for the rat liver were the IDR "optimal set"
558 from the second and third rat liver replicates from data we previously generated [61]. To determine the
559 distance from the nearest protein-coding TSS, we used the GENCODE protein-coding TSS's for human
560 (version 27) and mouse (version M15) [109, 110], the union of the RefSeq rheMac8 protein-coding TSS's
561 [1] and the human GENCODE protein-coding TSSs mapped to rheMac8 using liftOver [111] for macaque,
562 the union of the RefSeq rn6 protein-coding TSS's [1] and the mouse GENCODE protein-coding TSSs
563 mapped to rn6 using liftOver [111] for rat, the union of the RefSeq Btau_5.0.1 TSS's [1] and the human
564 GENCODE protein-coding TSS's mapped to Btau_5.0.1 with halLiftover [112] on the version 1 Zoonomia
565 Cactus alignment [11] for cow, and the union of the susScr11 TSS's mapped to susScr3 with liftOver [111]
566 and the human GENCODE protein-coding TSS's mapped to susScr3 with halLiftover [112] on the version 1
567 Zoonomia Cactus alignment [11] for pig. To identify non-exonic peaks, we used bedtools [108] subtract
568 with option -A to identify peaks that did not overlap protein-coding exons, where human protein-coding
569 exons were obtained from GENCODE (version 27), mouse protein-coding exons were obtained from
570 GENCODE (version M15) [109, 110], macaque protein-coding exons were obtained from RefSeq for
571 rheMac8, and rat protein-coding exons were obtained from RefSeq for rn6 [1], cow protein-coding exons
572 were obtained from RefSeq for Btau_5.0.1 [1], and pig protein-coding exons were obtained from RefSeq
573 for susScr11 [1] and then mapped to susScr3 with liftOver [111]. We defined the peak summit of an OCR

574 to be the peak summit of the corresponding base peak, and we constructed positive examples by taking
575 likely enhancer peaks summits +/- 250bp and their reverse complements. We centered peaks on their
576 summits because previous work has shown that there is a concentration of TF motifs at peak summits
577 [113-115]. By requiring OCRs to be reproducible open chromatin peaks according to IDR, intersecting
578 OCRs across multiple datasets, and filtering OCRs in a conservative way, we limited the number of false
579 positive OCRs being used to train our machine learning models.

580

581 Constructing Negative Set with Non-OCR Orthologs of OCRs

582 We obtained the non-OCR orthologs of OCRs by obtaining the orthologs of OCRs in each species
583 in each other species for which we had data from the same tissue and filtering the orthologs to include
584 only those that did not overlap open chromatin in the same tissue. For example, mouse chr4:127435564-
585 127436049 does not overlap any OCRs in mouse cortex or striatum, but its human ortholog,
586 chr1:34684126-34684689, is an OCR in both cortex and striatum, so this mouse region was a member of
587 our novel negative set. To ensure that we would have enough negatives for training the model, we
588 created a less conservative set of OCRs, which we called "loose OCRs." For each species, tissue
589 combination, the loose OCRs are the base peaks that are non-exonic, at least 20kb from a TSS, and at most
590 1kb long (same criteria used in constructing the positive set) and intersect at least 1 peak from the pooled
591 reads across replicates from each of other datasets that are used for the species, tissue combination; we
592 obtained these loose OCRs using bedtools [108]. We defined the peak summit of a loose OCR to be the
593 peak summit of the corresponding base peak.

594 We identified orthologs of our loose OCRs in each other species with open chromatin data from
595 the same tissue using halLiftover [112] followed by HALPER [113]. We used these tools instead of liftOver
596 [111] because they map regions using Cactus alignments [116], which, unlike the pairwise alignment

597 liftOver chains, contain many species and account for a wide range of structural rearrangements, including
598 inversions. We ran halLiftover with default parameters on our loose OCRs and their peak summits using
599 the Zoonomia version 1 Cactus alignment [11]. We then constructed contiguous loose OCR orthologs
600 from the outputs of halLiftover by running HALPER with parameters -max_frac 2.0, -min_len 50, and -
601 protect_dist 5. Finally, we used bedtools subract with the -A option [108] to remove the loose OCR
602 orthologs in each species that overlapped peaks from the pooled reads across replicates from any of the
603 datasets from the same tissue in that species.

604 For the models trained only on mouse sequences, we used only the mouse orthologs of loose
605 OCRs in non-mouse species. For the multi-species brain models, we used the mouse, human, macaque,
606 and rat orthologs of loose brain OCRs from each of the other species, and for the multi-species liver
607 models, we used the mouse, macaque, and rat orthologs of loose liver OCRs from each of the other
608 species. When constructing negatives of the non-OCR orthologs of OCRs, we used sequence underlying
609 the ortholog of the base peak summit +/- 250bp and that sequence's reverse complement. Our
610 negative:positive training set ratios were approximately 1.16:1 for the brain model trained on only mouse
611 sequences, 0.704:1 for the liver model trained on only mouse sequences, 1.49:1 for the multi-species
612 brain model, and 0.822:1 for the multi-species liver model. Construction of additional negative sets is
613 described in **Supplemental Methods**.

614

615 <u>Training Machine Learning Models</u>

616 Construction of training, validation, and test sets is described in **Supplemental Methods**. We used
617 CNNs [70] for our machine learning model because they have achieved state-of-the-art performance in
618 related tasks [58, 117, 118]; they can learn complex combinatorial relationships between sequences,
619 which we know can play an important role in enhancer activity [48]; and they do not require an explicit

620 featurization of the data, enabling them to learn yet-to-be-discovered sequence patterns that are
621 important for enhancer activity. Our inputs were one-hot-encoded DNA sequences [69], and our outputs
622 were the probability of the sequence being an OCR in the tissue for which the model was trained. We
623 used smaller architectures than those used by many previous studies for related tasks [56, 58, 117, 118]
624 because these previous studies were training multi-task models, where our models were single-task
625 models and therefore likely needed to learn substantially fewer sequence signatures. We first tuned
626 hyper-parameters for the brain model trained on only mouse sequences with our novel negative set by
627 comparing the validation set performances of various models. We did not do an exhaustive hyper-
628 parameter search for any model because our goal was to evaluate the feasibility of training models that
629 had satisfactory performance for all of our evaluation criteria and not to train optimally performing
630 machine learning models. Our final architecture was five convolutional layers with 300 filters of width 7
631 and stride 1 in each, followed by a max-pooling layer with width and stride 26, followed by a fully
632 connected layer with 300 units, followed by a fully-connected layer that went into a sigmoid. All
633 convolutional layers had dropout 0.2 and L2 regularization 0.00001. The model was trained using
634 stochastic gradient decent with learning rate 0.001, Nesterov momentum 0.99, and batch size 100, and
635 each class was assigned a weight equal to the fraction of the other class in the training set. The model
636 was trained using the training set until there were three consecutive epochs with no improvement in recall
637 at eighty percent precision (or, if there were more positives than negatives, no improvement in specificity
638 at eighty percent NPV) on the validation set. Before training, weights were initialized to be those from a
639 pre-trained neural network with the same hyper-parameters and the negative set randomly down-
640 sampled to be the size of the positive set (or a positive set randomly down-sampled to be the size of the
641 negative set if the positive set was larger). The weights for the pre-training were initialized using Keras's
642 He normal initializer [119, 120].

643 We then tuned hyper-parameters for the CNNs for the other negative sets, the CNNs for the liver
644 data, and the CNNs for the multi-species models. We began with the architecture described above and
645 adjusted the number of convolutional filters per layer and the learning rate, ultimately selecting the values
646 that provided the best performance on the validation set. For the model with flanking regions negatives,
647 we used 250 convolutional filters per layer and learning rate 0.001. For the model with the OCRs from
648 other tissues negatives, we used 250 convolutional filters per layer and learning rate 0.0005. For the
649 model with the larger number of random G/C- and repeat-matched negatives, the multi-species brain
650 model, and both liver models, we used 350 convolutional filters per layer and learning rate 0.001. For the
651 models with the smaller number of random G/C- and repeat-matched negatives and the dinucleotide-
652 shuffled negatives, we used 300 convolutional filters per layer and learning rate 0.001. All models were
653 implemented and trained using Keras [120, 121] version 1.2.2 with the Theano backend [122] and
654 evaluated using Scikit-learn [123] and PRROC [124].

655

656 Evaluating Machine Learning Models

657 *Evaluating Models' Lineage-Specific OCR Accuracy*

658 OCR Orthologs with Differing Open Chromatin Statuses between Two Species:

659 To obtain OCR orthologs that are open in one species but not in another, we used as positives the
660 sequences and reverse complements of sequences underlying OCRs from one species whose orthologs in
661 the other species have closed chromatin and as negatives sequences and reverse complements of
662 sequences underlying non-OCRs whose orthologs in at least one other species are OCRs. More specifically,
663 for evaluating mouse OCRs whose open chromatin status differs in other species, we used halLiftover
664 [112] followed by HALPER [113] with the same parameters we used previously to identify mouse OCR
665 orthologs in human, macaque, and rat for brain and macaque and rat for liver that did not overlap any of

666 the peaks from the pooled replicates from any of the datasets that we used for identifying OCRs. This
667 gave us 1,570 positive test set examples for evaluating brain models and 3,738 positive test set examples
668 for evaluating liver models. For evaluating mouse non-OCRs whose open chromatin status differs in other
669 species, we used the same approach that we took to construct our novel negative set except that we
670 identified non-OCR orthologs of OCRs instead of loose OCRs. This gave us 2,062 negative test set examples
671 for evaluating brain models and 4,080 negative test set examples for evaluating liver models
672 (**Supplemental Figure 1d, Supplemental Figure 8b**). For evaluating mouse OCRs whose rat orthologs are
673 not OCRs, we identified mouse OCRs whose rat orthologs do not overlap pooled peaks across replicates
674 for any rat OCRs from any of the datasets used to define OCRs; this gave us 674 positive test set examples
675 for evaluating brain models and 2,482 positive test set examples for evaluating liver models. For
676 evaluating mouse closed chromatin regions whose orthologs are open in rat, we used the subset of our
677 novel negative set that came from rat OCR orthologs (not rat loose OCR orthologs); this gave us 890
678 negative test set examples for evaluating brain models and 2,050 negative test set examples for evaluating
679 liver models (**Supplemental Figure 1e, Supplemental Figure 8b**). For evaluating macaque OCRs whose
680 open chromatin status differs in mouse, we identified macaque OCRs whose mouse orthologs do not
681 overlap peaks from the pooled replicates of any of the datasets we used for identifying mouse OCRs; this
682 gave us 734 positive test set examples for evaluating brain models and 2,384 positive test set examples
683 for evaluating liver models. For identifying macaque non-OCRs whose OCR status differs in mouse, we
684 identified mouse OCR orthologs in macaque that do not overlap peaks from the pooled replicates from
685 any of the datasets we used for identifying macaque OCRs; this gave us 788 negative test set examples
686 for evaluating brain models and 2,228 negative test set examples for evaluating liver models
687 (**Supplemental Figure 1g, Supplemental Figure 8b**). We obtained human and rat OCRs and non-OCRs
688 whose OCR status is different in mouse using same process that we used for macaque. We obtained 416
689 positive and 896 negative test set examples for evaluating brain models (**Supplemental Figure 1h**). For

690 evaluating rat OCRs and non-OCRs whose mouse orthologs have different OCR statuses, we obtained 990
691 positive and 676 negative test set examples for evaluating brain models and 2,050 positive and 2,482
692 negative test set examples for evaluating liver models (**Supplemental Figure 1i, Supplemental Figure 8b**).

693 Clade- and Species-Specific OCRs:

694 We defined a clade-specific OCR in a tissue as an OCR whose ortholog is open in that tissue in
695 every species within a clade for which we have data and closed in every other species for which we have
696 data, and we defined a species-specific OCR as an OCR whose ortholog is open a species for which we
697 have data and is closed in the most closely related species for which we have data. Laurasiatheria data
698 was used only for comparing mouse-only to multi-species liver models (**Figure 3e**). More specifically, we
699 identified clade-active OCRs in each clade – OCRs in a "base species" whose ortholog in the other species
700 in that clade (if there was another) overlaps an open chromatin peak from the pooled reads across
701 replicates in all of the datasets we used in that tissue from that species. We did not require the OCR
702 ortholog in the non-base species to overlap a reproducible open chromatin peak so that we could have at
703 least one hundred test set examples for each evaluation. We chose the "base species" to be the species
704 in each clade with the highest-quality genomes – mouse for Glires for brain and liver, human for
705 Euarchonta for brain, macaque for Euarchonta for liver, and cow for Laurasiatheria for liver. We then
706 identified the subset of clade-active peaks from the base species whose orthologs in all species in the
707 other clade do not overlap any open chromatin peaks from the pooled replicates from any of the datasets
708 we used to identify OCRs; these were our clade-specific OCRs. To obtain clade-specific non-OCRs for a
709 clade, we identified orthologs of clade-active OCRs from the other clade in the base species in the clade
710 whose orthologs in all species in the clade did not overlap open chromatin peaks from pooled replicates
711 in any of the datasets we used for identifying OCRs. The sequences of clade-specific OCRs and non-OCRs
712 used for evaluating the models were those from the base species and their reverse complements. This
713 gave us 230 Glires-specific test set brain OCRs, 134 Glires-specific test set brain non-OCRs (**Supplemental**

714 **Figure 1f, Supplemental Figure 4a, Figure 3a**), 134 Euarchonta-specific test set brain OCRs, 230
715 Euarchonta-specific test set brain non-OCRs (**Supplemental Figure 1j, Supplemental Figure 4b, Figure 3a**),
716 1,024 Glires-specific test set liver OCRs, 1,826 Glires-specific test set liver non-OCRs (**Supplemental Figure**
717 **8b, Figure 3b**), 1,826 Euarchonta-specific test set liver OCRs, 1,024 Euarchonta-specific test set liver non-
718 OCRs (**Supplemental Figure 8b, Figure 3b**), 77 Laurasiatheria-specific test set liver OCRs, and 86
719 Laurasiatheria-specific test set liver non-OCRs (**Figure 3e**). When evaluating the multi-species models, we
720 combined the clade-specific OCRs and non-OCRs from Euarchonta and Glires.

721 When evaluating species-specific OCRs and non-OCRs, we identified mouse OCRs and non-OCRs
722 whose orthologs' open chromatin status differs in rat, rat OCRs and non-OCRs whose orthologs' open
723 chromatin status differs in mouse, human OCRs and non-OCRs whose orthologs' open chromatin status
724 differs in macaque, and macaque OCRs and non-OCRs whose orthologs' open chromatin status differs in
725 human. We obtained 66 human-specific brain OCRs, 188 human-specific brain non-OCRs, 188 macaque-
726 specific brain OCRs, and 66 macaque-specific brain non-OCRs (**Figure 3a**). We did not include macaque-
727 specific OCRs and non-OCRs when evaluating the multi-species liver model because we did not have liver
728 open chromatin from any other Euarchonta species. We combined the species-specific OCRs and non-
729 OCRs from different species when evaluating the multi-species models (**Figures 3a-b**).

730 *Evaluating Models' Tissue-Specific OCR Accuracy*

731 To evaluate the performance of models trained in one tissue on OCRs from another tissue, we
732 defined our positives to be OCRs that are shared between the two tissues (shows our models were not
733 learning only the sequences involved in tissues-specific open chromatin), and we defined our negatives
734 to be OCRs in the evaluation tissue that do not overlap OCRs in the training tissue (shows our models were
735 not learning only sequences involved in general open chromatin). More specifically, we used bedtools
736 intersectBed with options -wa and -u [108] to identify OCRs from our training tissue that overlap OCRs

737 from the evaluation tissue. For brain models, we obtained 1,040 positives for mouse, 846 positives for
738 macaque, and 1,770 positives for rat (**Supplemental Figure 5**). For liver models, we obtained 2,012
739 positives for mouse, 946 positives for macaque, and 1,130 positives for rat (**Supplemental Figure 8c**). We
740 used bedtools subtractBed with option -A [108] to identify liver OCRs that do not overlap any open
741 chromatin from pooled replicates in any of the datasets that were used to generate the brain OCRs and
742 brain OCRs that do not overlap open chromatin from pooled replicates in any of the datasets that were
743 used to generate liver OCRs. For brain, we obtained 3,382 negatives for mouse, 1,898 negatives for
744 macaque, and 3,518 negatives for rat (**Supplemental Figure 5**). For liver, we obtained 2,212 negatives for
745 mouse, 1,428 negatives for macaque, and 2,942 negatives for rat (**Supplemental Figure 8c**). We combined
746 the data from all three species for evaluating the multi-species models (**Figures 3a-b**).

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

753 *Evaluating if Models' Predictions Had Phylogeny-Matching Correlations*

754 To evaluate the relationship between OCR ortholog open chromatin status and phylogenetic
755 distance, we identified test set mouse OCR orthologs in all of the fifty-six Glires species from Zoonomia,
756 predicted the open chromatin statuses of those orthologs, and computed the correlation between those
757 predictions and the species' phylogenetic divergences from mouse. This provides us with an approximate
758 measure of how predicted OCR ortholog open chromatin statuses change over evolution. We identified
759 the test set mouse OCR orthologs and OCR summit orthologs in Glires using halLiftover [112] with the

760 Zoonomia version 1 Cactus alignment [11, 116]; we used brain OCRs for evaluating brain OCR models and
761 liver OCRs for evaluating liver OCR models. We next constructed contiguous orthologs from the outputs
762 of halLiftover using HALPER [113] with parameters -max_frac 2.0, -min_len 50, and -protect_dist 5. We
763 constructed inputs for our models from the contiguous OCR orthologs by using bedtools fastaFromBed
764 [108] with fasta files downloaded from NCBI [4, 125] and the UCSC Genome Browser [126] to obtain the
765 sequences underlying their summit orthologs +/- 250bp. We constructed the reverse complements of
766 sequences, used our models to predict each sequence and its reverse complement's open chromatin
767 status, and averaged the predictions between the forward and reverse strands. We then removed all
768 predictions from OCRs with orthologs in less than one quarter of species. After that, for each model, we
769 computed the mean OCR ortholog open chromatin status prediction and the standard deviation of
770 predictions across all remaining OCR orthologs in each species. We finally computed the Pearson and
771 Spearman correlations between these means and standard deviations of predictions and the millions of
772 years since divergence from mouse, which we obtained from TimeTree [127]. We did this for brain OCR
773 orthologs using brain models trained on mouse sequences from each negative set and the multi-species
774 brain model as well as for liver OCR orthologs using the liver model trained on only mouse sequences and
775 the multi-species liver model.

776

777 Comparing Predictions to Mean Conservation Scores

778 We compared the predictions to mean conservation scores by identifying OCR orthologs with
779 conserved and non-conserved open chromatin status between species, computing the mean conservation
780 scores of those OCR orthologs, and comparing those scores to the predicted open chromatin status of
781 those OCR orthologs. We defined an OCR ortholog with conserved open chromatin status between mouse
782 and another species as a mouse OCR whose ortholog in the other species overlaps an OCR in the same

783 tissue. For mouse brain test set OCRs, we identified 441 OCRs with conserved open chromatin status in
784 macaque, 195 OCRs with conserved open chromatin status in human, and 670 OCRs with conserved open
785 chromatin status in rat. For mouse liver test set OCRs, we identified 689 OCRs with conserved open
786 chromatin status in macaque and 580 OCRs with conserved open chromatin status in rat. We defined an
787 OCR ortholog with non-conserved open chromatin status between mouse and another species as a mouse
788 OCR whose ortholog in another species does not overlap any OCR from the pooled replicates from any
789 dataset used in defining an OCR in that tissue in that species. For mouse brain test set OCRs, we identified
790 394 OCR orthologs with non-conserved open chromatin status in macaque, 448 OCR orthologs with non-
791 conserved open chromatin status in human, and 338 OCR orthologs with non-conserved open chromatin
792 status in rat. For mouse liver test set OCRs, we identified 1,114 OCR orthologs with non-conserved open
793 chromatin status in macaque and 1,241 OCR orthologs with non-conserved open chromatin status in rat.
794 We think that the differences in numbers of OCR orthologs with conserved and non-conserved open
795 chromatin status between species is due not only to differences in evolutionary relatedness but also to
796 differences between species in numbers of datasets used to define OCRs and differences in sequencing
797 depths of those datasets [38, 61, 65].

798 We used our models to predict the OCR ortholog open chromatin status for the open chromatin
799 status-conserved and open chromatin status non-conserved OCR orthologs in the non-mouse species and
800 compared it to the conservations scores of the mouse OCRs. We computed mean conservation scores of
801 the mouse OCRs by calculating the mean PhastCons [13] and PhyloP [14] scores at the peak summits +/-
802 250bp. We evaluated if the distributions of the predictions and each type of conservation score differed
803 between the open chromatin status-conserved and open chromatin status non-conserved orthologs using
804 a Wilcoxon rank-sum test. We did a Bonferroni correction by multiplying all p-values by 20 (2 conservation
805 score comparisons and 2 model predictions comparisons – models trained on only mouse sequences and
806 multi-species models – for 5 species, tissue pairs).

807 We then evaluated whether the predictions were more effective than the mean conservation
808 scores at differentiating between open chromatin status-conserved and open chromatin status non-
809 conserved OCR orthologs. We first averaged the predictions of the sequence underlying the non-mouse
810 OCR ortholog's summit +/-250bp and its reverse complement so that each OCR ortholog would have a
811 single prediction value. We next combined our open chromatin status-conserved and open chromatin
812 status non-conserved OCR orthologs and ranked them according to each of PhastCons score, PhyloP score,
813 and OCR ortholog open chromatin status prediction. We then did a Wilcoxon sign-rank test to compare
814 the ranking distributions of the open chromatin status-conserved OCR orthologs between the OCR
815 ortholog open chromatin status predictions and each type of conservation score. We also did this for the
816 ranking distributions of the open chromatin status non-conserved OCR orthologs. We did this for
817 predictions made by the models trained using only mouse sequences and by the multi-species models.
818 Finally, we did a Bonferroni correction by multiplying all p-values by 40 (2 conservation score comparisons
819 for each of open chromatin status-conserved and open chromatin status non-conserved OCR orthologs in
820 2 tissues for 5 species, tissue pairs). Processing of liver H3K27ac ChIP-seq data for comparing predicted
821 open chromatin conservation and conservation scores to conservation of H3K27ac ChIP-seq is described
822 in **Supplemental Methods**.

823

824 <u>Clustering OCRs</u>:

825 Identifying genes with Rodent-specific expression OCRs with predicted Rodent-specific open
826 chromatin is described in **Supplemental Methods**. To generate OCR clusters for OCRs from a tissue, we
827 mapped the OCRs from each species across all of the Boreoeutheria from Zoonomia [4] except for *Manis*
828 *tricuspis*, filtered the OCRs, and clustered the OCRs. Specifically, we mapped OCRs from each species with
829 OCRs using halLiftover with the Zoonomia version 1 Cactus [11] and constructed contiguous orthologs

830 from halLiftover outputs using HALPER [113] with settings -max_frac 2.0, -min_len 50, and -protect_dist
831 5. We then used the model corresponding to the tissue for which the OCR was generated to make
832 predictions for the OCR orthologs and OCR orthologs' reverse complements in all of the Boreoeutheria
833 that we mapped to except for *Galeopterus variegatus*, *Hippopotamus amphibius*, *Monodon monoceros*,
834 and *Platanista gangetica*. For each OCR ortholog, we set the prediction to be the average between the
835 prediction for the ortholog and the prediction for its reverse complement.

836 We filtered OCRs to ensure that we did not have redundant OCRs and to ensure that we had
837 usable OCR orthologs in enough species to use predictions in each species as features for clusters. First,
838 for brain, we removed all human OCRs whose mouse ortholog overlapped a mouse brain OCR, all macaque
839 OCRs whose mouse ortholog overlapped a mouse brain OCR or whose human ortholog overlapped a
840 human brain OCR, and all rat OCRs whose mouse ortholog overlapped a mouse brain OCR, whose human
841 ortholog overlapped a human brain OCR, or whose macaque ortholog overlapped a macaque brain OCR.
842 Likewise, for liver, we removed all macaque OCRs whose mouse ortholog overlapped a mouse liver OCR
843 and all rat OCRs whose mouse ortholog overlapped a mouse liver OCR or whose macaque ortholog
844 overlapped a macaque liver OCR. After that, we removed all remaining OCRs that did not have a usable
845 ortholog in at least half of all species or at least one quarter of each of Euarchonta, Glires, and
846 Laurasiatheria.

847 We clustered the remaining OCRs using the prediction of each species's OCR ortholog's open
848 chromatin status as a feature and treating species without usable orthologs as missing data. We first used
849 k-means clustering with k = 9,000 (brain) or 12,000 (liver) to cluster the OCRs into small clusters; cluster
850 centroid values for each species were defined as the mean of all OCR ortholog activity predictions for that
851 species in the cluster. We then used affinity propagation clustering [128], implemented in scikit-learn
852 [123], with preference = -0.6 to cluster the outputs of the k-means clustering, which we defined as the
853 centroids of the small clusters, into larger clusters. In both clustering steps, we defined the distance

854 between two OCRs as 1 minus the cosine similarity between the vectors of enhancer activity predictions
855 in the species for which both had usable orthologs. We ultimately obtained 102 brain clusters and 103
856 liver clusters.

857

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

859 To identify enhancer sets with more overlap with a cluster than expected by chance, we first
860 obtained the enhancer sets from the supplemental information of the relevant manuscripts. For the
861 mouse neuron firing candidate enhancers, we used the regions in Tables 11, S12 and S13 from a recent
862 study of enhancers activated during mouse neuron firing [94] and used liftOver [111] to map the
863 coordinates from mm9 to mm10. For the human GABAergic neuron activity candidate enhancers, we
864 used the regions in Supp Data 12 from a recent study of enhancers activated during human neuron
865 activation [95] and used liftOver to map the coordinates from hg19 to hg38 [111]. For the mouse liver
866 regeneration candidate enhancers, we used various subsets of regions from Supplemental Table 1 from a
867 recent study of enhancers activated during liver regeneration [96]; for each category of regions in
868 **Supplemental Table 19**, we required the FDR to be less than 0.05 and the log2 fold-change to be greater
869 than 1 (for ↑) or less than 1 (for ↓). We then used LOLA [129] to run a hypergeometric test to evaluate
870 the statistical significance of the overlap of each enhancer set with each cluster from the relevant tissue
871 that is open in the relevant species; our query was our enhancer set, our database was our list of bed files
872 with the relevant cluster locations, our universe was our list of OCRs in the relevant species, tissue
873 combination, and we set "redefineUserSets" to TRUE. We did a Bonferroni correction by multiplying all
874 p-values by 391, which was the total number of tests.

875 We wanted to determine if our enrichment for overlaps with clusters with lineage-specific open
876 chromatin or lack of open chromatin could be partially explained by OCRs overlapping enhancer sets

877 having fewer usable orthologs or lower conservation. To evaluate this, we used bedtools [108] to identify
878 the OCRs from each relevant tissue, species combination overlapping each enhancer set. We next
879 computed the number of usable orthologs, the average PhastCons [13] score for the peak summit +/-
880 250bp, and the average PhyloP [14] score for the peak summit +/- 250bp for all OCRs from the relevant
881 tissue, species combination as well as for the OCRs overlapping each enhancer set. Then, for each
882 enhancer set, we used a Wilcoxon rank-sum test to compare the numbers of usable orthologs and the
883 PhastCons and PhyloP scores between the OCRs overlapping the enhancer set and the full set of OCRs.
884 We considered the difference to not be statistically significant if the nominal p-value was greater than or
885 equal to 0.05.

886

887

888 **DECLARATIONS**

889 **Ethics approval and consent to participate:** All human data is publicly available, and all other animal data
890 is either publicly available or was collected in experiments approved by Carnegie Mellon University's
891 Institutional Animal Care and Use Committee.

892 **Consent for publication:** Not applicable.

893 **Availability of data and materials:** Human DNase hypersensitivity data analyzed in this study was
894 downloaded from the ENCODE portal (https://www.encodeproject.org/) [38, 64], and human ATAC-seq
895 data analyzed in this study was downloaded from Gene Expression Omnibus accession GSE96949 [65].
896 Macaque and rat ATAC-seq data from data we previously generated [61] and mouse liver ATAC-seq data
897 in this manuscript can be accessed in Gene Expression Omnibus accession GSE159815 with token
898 gzkbqwyobvmvfsd. Mouse brain data analyzed in this study can be accessed in Gene Expression Omnibus

accession GSE161374 with token cropkwsgnnyxhgh. Publicly available mouse liver ATAC-seq data was downloaded from China National Gene Bank accession CNP0000198 [130]. Other mouse ATAC-seq data was downloaded from China National Gene Bank accession CNP0000198 [130] and the ENCODE portal (https://www.encodeproject.org/) [131]. Cow and pig ATAC-seq data was obtained from the authors of [72]. Mouse and macaque H3K27ac ChIP-seq data was downloaded from ArrayExpress accession E-MTAB-2633 [37]. 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.

Competing interests: The authors declare that they have no competing interests.

Funding: The Carnegie Mellon University Computational Biology Department Lane Fellowship supported I.M.K. The Alfred P. Sloan Foundation Research Fellowship supported I.M.K., M.E.W., and A.R.P. The NIH NIDA DP1DA046585 supported I.M.K., D.E.S., M.E.W., A.J.L., A.R.B., M.K., and A.R.P. The NSF Graduate Research Fellowship Program under grants DGE1252522 and DGE1745016 supported A.J.L. The Carnegie Mellon Neuroscience Institute Presidential Fellowship supported M.K.

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 criteria 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.

921 **Acknowledgements:** We would like to thank the other members of the Pfenning Lab for useful discussions
922 and suggestions and J. Ma and Y. Zhang for feedback on the manuscript. We would also like to thank the
923 members of the Paten Lab and the Zoonomia Project for assistance with using Cactus multi-species
924 alignments and providing us with early access to these alignments. This work used the Extreme Science
925 and Engineering Discovery Environment (XSEDE), through the Pittsburgh Supercomputing Center Bridges
926 Compute Cluster, which is supported by National Science Foundation grant number TG-MCB190067.

927

928

929 **FIGURE CAPTIONS**

930

931 **Figure 1: OCR Ortholog Open Chromatin Status Prediction Framework Overview**

932 We trained a convolutional neural network (CNN) for predicting brain open chromatin using sequences
933 underlying brain open chromatin region (OCR) orthologs in a small number of species and used the CNN
934 to predict brain OCR ortholog open chromatin status across all of the species in the Zoonomia Project.
935 We used the sequences underlying the orthologs for which we have brain open chromatin data to train a
936 CNN for predicting open chromatin. Then, we used the CNN to predict the probability of brain open
937 chromatin for all brain OCR orthologs; predictions are illustrated on the right.

938

939 **Figure 2: Violin Plots for Lineage-Specific and Tissue-Specific OCR Accuracy Evaluation in Macaque**

940 **a)** Comparison of PhastCons [13] and PhyloP [14] scores to three different machine learning models'
941 predictions for brain OCRs with conserved open chromatin across mouse and macaque, macaque brain

942 OCRs whose mouse orthologs are closed in brain, macaque brain non-OCRs whose mouse orthologs are
943 open in brain, macaque brain OCRs that are closed in liver, macaque brain OCRs that are open in liver
944 (centered on brain peak summits), and macaque liver OCRs that are closed in brain.

945 **b)** Comparison of PhastCons [13] and PhyloP [14] scores to two different machine learning models'
946 predictions for liver OCRs with conserved open chromatin across mouse and macaque, macaque liver
947 OCRs whose mouse orthologs are closed in liver, macaque liver non-OCRs whose mouse orthologs are
948 open in liver, macaque liver OCRs that are closed in brain, macaque liver OCRs that are open in brain
949 (centered on liver peak summits), and macaque brain OCRs that are closed in liver.

950 +'s indicate that values should be large, and -'s indicate that values should be small. Conservation scores
951 were generated from the mm10-based placental mammals alignment [12, 71] and averaged over 500bp
952 centered on peak summits, where mouse peak summits were used for OCRs conserved between mouse
953 and macaque and OCRs in mouse whose macaque orthologs are closed, and mouse orthologs of macaque
954 peak summits were used for other evaluations. All machine learning model predictions were made using
955 macaque sequences, where the macaque sequences for OCRs conserved between mouse and macaque
956 and OCRs in mouse whose macaque orthologs are closed were centered on macaque orthologs of mouse
957 peak summits, and macaque peak summits were used for other evaluations. Animal silhouettes were
958 obtained from PhyloPic [132]. *'s indicate the species from which sequences were obtained for making
959 predictions. Dinuc.-shuf. stands for dinucleotide-shuffled, and orths. stands for orthologs.

960

961 **Figure 3: Multi-Species Model Performance**

962 **a)** Performance of multi-species brain model on full test set, subset of test set consisting of clade-specific
963 brain OCRs and non-OCRs, subset of test set consisting of species-specific brain OCRs and non-OCRs, and

964 subset of positive test set consisting of OCRs shared between brain and liver as well as subset of liver OCRs
965 from test set genomic regions that do not overlap brain OCRs.

966 **b)** Performance of multi-species liver model on full test set, subset of test set consisting of clade-specific
967 liver OCRs and non-OCRs, subset of test set consisting of species-specific liver OCRs and non-OCRs, and
968 subset of positive test set consisting of OCRs shared between liver and brain as well as subset of brain
969 OCRs from test set genomic regions that do not overlap liver OCRs. We reported area under the negative
970 predictive value (NPV)-specificity (Spec.) curve instead of the AUPRC because these test sets have more
971 positives than negatives.

972 **c)** Divergence from mouse versus mean multi-species brain model predictions across mouse brain OCR
973 orthologs in Glires.

974 **d)** Divergence from mouse versus mean multi-species liver model predictions across mouse liver OCR
975 orthologs in Glires.

976 **e)** Performance of mouse-only liver model versus multi-species liver model on Laurasiatheria-specific liver
977 enhancers and liver non-enhancers whose mouse orthologs are in the test set.

978 Animal silhouettes were obtained from PhyloPic [132]. AUC stands for area under the receiver operating
979 characteristic curve, AUPRC stands for area under the precision-recall curve, and MYA stands for millions
980 of years ago. The red curves are the best fit exponential function of the form $y = ae^{bx}$. The red dotted
981 lines are the average prediction across test set negatives.

982

983 **Figure 4: Examples of Mean Conservation Score and Open Chromatin Status Prediction versus Open**
984 **Chromatin Conservation**

985 **a)** 7-week-old mouse cortex and striatum and macaque orofacial motor cortex ("Cortex") and putamen
986 ("Striatum") open chromatin signal for a mouse brain OCR that is 50,328 bp away from the *Stx16*
987 transcription start site (TSS). Experimentally identified and predicted brain open chromatin statuses are
988 conserved even though mean mouse PhastCons score is low.

989 **b)** 7-week-old mouse cortex and striatum and macaque orofacial motor cortex ("Cortex") and putamen
990 ("Striatum") open chromatin signal for a mouse brain OCR that is 144,474 bp away from the *Lnpk* TSS.
991 Experimentally identified and predicted brain open chromatin statuses are not conserved even though
992 mean mouse PhastCons score is high.

993 **c)** Our mouse liver and macaque liver open chromatin signal for a mouse liver OCR that is 24,814 bp away
994 from the *Rxra* TSS. Experimentally identified and predicted liver open chromatin statuses are conserved
995 even though mean mouse PhastCons score is low.

996 **d)** Our mouse liver and macaque liver open chromatin signal for a mouse liver OCR that is 154,404 bp
997 away from the *Fn1* TSS. Experimentally identified and predicted liver open chromatin statuses are not
998 conserved even though mean mouse PhastCons score is high.

999 Animal silhouettes were obtained from PhyloPic [132]. Regions are mouse cortex open chromatin peak
1000 summits +/-250bp and their macaque orthologs, signals are from pooled reads across biological replicates,
1001 and liver H3K27ac ChIP-seq data comes from E-MTAB-2633 [37].

1002

1003 **Figure 5: Examples of Brain OCR Clusters with Predicted Lineage-Specific Open Chromatin Associated**
1004 **with Neuron Activity**

1005 We clustered the brain OCRs, where the features were the brain predictions in each Boreoeutheria species
1006 from Zoonomia, and then identified clusters whose regions had significant overlap with regions associated

1007 with mouse neuron firing and human neuron activity. Mouse neuron firing enhancers had significant
1008 overlap with a predicted Murinae-specific OCR cluster (cluster 43), and human neuron activity enhancers
1009 had significant overlap with a predicted predicted Euarchonta and Carnivora-specific non-OCR cluster
1010 (cluster 74). Animal silhouettes were obtained from PhyloPic [132].

1011

1012

1013 **TABLES**

1014 **Table 1: Comparison of Evaluation Criteria Used to Evaluate Models in Candidate Enhancer Activity**
1015 **Conservation Prediction Papers**

Evaluation Criterium	Chen *et al*., 2018	Huh *et al*., 2018	Kelley, 2020	Minnoye *et al*., 2020	This Paper
Evaluation on Genomic Regions not Used in Training	✓	✓	✓	✓	✓
Evaluation on Species not Used in Training	✓	✓	✓	×	✓
Comparison of Model Trained on One Species to Model Trained on Multiple Species	×	✓	✓	×	✓
Evaluation of Lineage-Specific Enhancer Accuracy	×	×	×	×	✓
Evaluation of Tissue-Specific Enhancer Accuracy	×	×	×	×	✓
Evaluation of Phylogeny-Matching Correlations	×	×	×	×	✓
Comparison to Conservation Scores	×	×	×	×	✓
Identification of Biologically Relevant Enhancers with Correctly Predicted Lineage-Specificity	×	×	×	✓	✓

1016

1017

1018 **REFERENCES**

1019 1. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-
1020 White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status,**
1021 **taxonomic expansion, and functional annotation**. *Nucleic Acids Res* 2016, **44**(D1):D733-745.
1022 2. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM,
1023 Consortium SWGotPG: **LD Score regression distinguishes confounding from polygenicity in**
1024 **genome-wide association studies**. *Nat Genet* 2015, **47**(3):291-295.
1025 3. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B,
1026 Onodera C, Siepel A *et al*: **An RNA gene expressed during cortical development evolved rapidly**
1027 **in humans**. *Nature* 2006, **443**(7108):167-172.
1028 4. Zoonomia Consortium: **A comparative genomics multitool for scientific discovery and**
1029 **conservation**. *Nature* 2020, **587**(7833):240-245.

1030 5. Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissieres V, Pickle CS,
1031 Plajzer-Frick I, Lee EA *et al*: **Progressive Loss of Function in a Limb Enhancer during Snake**
1032 **Evolution**. *Cell* 2016, **167**(3):633-642.e11.
1033 6. Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL:
1034 **Subterranean mammals show convergent regression in ocular genes and enhancers, along**
1035 **with adaptation to tunneling**. *Elife* 2017, **6**:e25884.
1036 7. Kowalczyk A, Partha R, Clark NL, Chikina M: **Pan-mammalian analysis of molecular constraints**
1037 **underlying extended lifespan**. *Elife* 2020, **9**:e51089.
1038 8. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W,
1039 Fungtammasan A, Gedman GL *et al*: **Towards complete and error-free genome assemblies of all**
1040 **vertebrate species**. *Nature* 2021, **592**(7856):737-746.
1041 9. Koepfli KP, Paten B, Genome 10K Community of Scientists, O'Brien SJ: **The Genome 10K Project:**
1042 **a way forward**. *Annu Rev Anim Biosci* 2015, **3**:57-111.
1043 10. Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, Consortium BK: **Bat Biology,**
1044 **Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat**
1045 **Species**. *Annu Rev Anim Biosci* 2018, **6**:23-46.
1046 11. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J
1047 *et al*: **Progressive Cactus is a multiple-genome aligner for the thousand-genome era**. *Nature*
1048 2020, **587**(7833):246-251.
1049 12. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K,
1050 Clawson H, Green ED *et al*: **Aligning multiple genomic sequences with the threaded blockset**
1051 **aligner**. *Genome Res* 2004, **14**(4):708-715.
1052 13. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J,
1053 Hillier LDW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm,**
1054 **and yeast genomes**. *Genome Research* 2005, **15**(8):1034-1050.
1055 14. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on**
1056 **mammalian phylogenies.** *Genome research* 2010, **20**(1):110-121.
1057 15. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood**. *Comput*
1058 *Appl Biosci* 1997, **13**(5):555-556.
1059 16. Partha R, Kowalczyk A, Clark NL, Chikina M: **Robust Method for Detecting Convergent Shifts in**
1060 **Evolutionary Rates**. *Mol Biol Evol* 2019, **36**(8):1817-1830.
1061 17. Pond SL, Frost SD, Muse SV: **HyPhy: hypothesis testing using phylogenies**. *Bioinformatics* 2005,
1062 **21**(5):676-679.
1063 18. Prudent X, Parra G, Schwede P, Roscito JG, Hiller M: **Controlling for Phylogenetic Relatedness**
1064 **and Evolutionary Rates Improves the Discovery of Associations Between Species' Phenotypic**
1065 **and Genomic Differences**. *Mol Biol Evol* 2016, **33**(8):2135-2150.
1066 19. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, London D, Song L, Lee
1067 BK, Iyer VR *et al*: **Extensive evolutionary changes in regulatory element activity during human**
1068 **origins are associated with altered gene expression and positive selection**. *PLoS Genet* 2012,
1069 **8**(6):e1002789.
1070 20. King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees**. *Science* 1975,
1071 **188**(4184):107-116.
1072 21. Wray GA: **The evolutionary significance of cis-regulatory mutations**. *Nat Rev Genet* 2007,
1073 **8**(3):206-216.
1074 22. Corradin O, Scacheri PC: **Enhancer variants: evaluating functions in common disease**. *Genome*
1075 *Med* 2014, **6**(10):85.

1076 23. Tak YG, Farnham PJ: **Making sense of GWAS: using epigenomics and genome engineering to**
1077 **understand the functional relevance of SNPs in non-coding regions of the human genome**.
1078 *Epigenetics Chromatin* 2015, **8**:57.
1079 24. Langer BE, Roscito JG, Hiller M: **REforge Associates Transcription Factor Binding Site Divergence**
1080 **in Regulatory Elements with Phenotypic Differences between Species**. *Mol Biol Evol* 2018,
1081 **35**(12):3027-3040.
1082 25. Berger MJ, Wenger AM, Guturu H, Bejerano G: **Independent erosion of conserved transcription**
1083 **factor binding sites points to shared hindlimb, vision and external testes loss in different**
1084 **mammals**. *Nucleic Acids Res* 2018, **46**(18):9299-9308.
1085 26. Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell
1086 PV, Ganapathy G *et al*: **Convergent transcriptional specializations in the brains of humans and**
1087 **song-learning birds**. *Science* 2014, **346**(6215):1256846.
1088 27. Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, Blanco-Aguiar JA,
1089 Plyusnina IZ, Trut L, Villafuerte R *et al*: **A comparison of brain gene expression levels in**
1090 **domesticated and wild animals**. *PLoS Genet* 2012, **8**(9):e1002962.
1091 28. Trut L, Oskina I, Kharlamova A: **Animal evolution during domestication: the domesticated fox**
1092 **as a model**. *Bioessays* 2009, **31**(3):349-360.
1093 29. Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov AV, Yim SH, Buffenstein R, Lee SR, Chang KT,
1094 Rhee H *et al*: **Gene expression defines natural changes in mammalian lifespan**. *Aging Cell* 2015,
1095 **14**(3):352-365.
1096 30. Ma S, Gladyshev VN: **Molecular signatures of longevity: Insights from cross-species**
1097 **comparative studies**. *Semin Cell Dev Biol* 2017, **70**:190-203.
1098 31. Fraser HB, Khaitovich P, Plotkin JB, Pääbo S, Eisen MB: **Aging and gene expression in the**
1099 **primate brain**. *PLoS Biol* 2005, **3**(9):e274.
1100 32. McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke
1101 DB, Schaar BT *et al*: **Human-specific loss of regulatory DNA and the evolution of human-**
1102 **specific traits**. *Nature* 2011, **471**(7337):216-219.
1103 33. Shen YY, Liang L, Li GS, Murphy RW, Zhang YP: **Parallel evolution of auditory genes for**
1104 **echolocation in bats and toothed whales**. *PLoS Genet* 2012, **8**(6):e1002788.
1105 34. Young RL, Ferkin MH, Ockendon-Powell NF, Orr VN, Phelps SM, Pogány Á, Richards-Zawacki CL,
1106 Summers K, Székely T, Trainor BC *et al*: **Conserved transcriptomic profiles underpin monogamy**
1107 **across vertebrates**. *Proc Natl Acad Sci U S A* 2019, **116**(4):1331-1336.
1108 35. Marcovitz A, Jia R, Bejerano G: **"Reverse Genomics" Predicts Function of Human Conserved**
1109 **Noncoding Elements**. *Mol Biol Evol* 2016, **33**(5):1358-1369.
1110 36. Wong ES, Zheng D, Tan SZ, Bower NL, Garside V, Vanwalleghem G, Gaiti F, Scott E, Hogan BM,
1111 Kikuchi K *et al*: **Deep conservation of the enhancer regulatory code in animals**. *Science* 2020,
1112 **370**(6517).
1113 37. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT,
1114 Jasinska AJ *et al*: **Enhancer evolution across 20 mammalian species**. *Cell* 2015, **160**(3):554-566.
1115 38. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis
1116 AB, Wang H, Vernot B *et al*: **The accessible chromatin landscape of the human genome**. *Nature*
1117 2012, **489**(7414):75-82.
1118 39. Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E *et*
1119 *al*: **Global reference mapping of human transcription factor footprints**. *Nature* 2020,
1120 **583**(7818):729-736.
1121 40. Vermunt MW, Tan SC, Castelijns B, Geeven G, Reinink P, de Bruijn E, Kondova I, Persengiev S,
1122 Bontrop R, Cuppen E *et al*: **Epigenomic annotation of gene regulatory alterations during**
1123 **evolution of the primate brain**. *Nat Neurosci* 2016, **19**(3):494-503.

1124 41. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ: **Transposition of native chromatin**
1125 **for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and**
1126 **nucleosome position.** *Nature methods* 2013, **10**:1213-1218.
1127 42. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford
1128 GE, Dekker J *et al*: **Defining functional DNA elements in the human genome**. *Proc Natl Acad Sci*
1129 *U S A* 2014, **111**(17):6131-6138.
1130 43. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM,
1131 Bilmes JA, Birney E *et al*: **Integrative annotation of chromatin elements from ENCODE data**.
1132 *Nucleic Acids Res* 2013, **41**(2):827-841.
1133 44. Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond AM, Ramachanderan R, Harewood L,
1134 Odom DT, Flicek P: **LINE retrotransposons characterize mammalian tissue-specific and**
1135 **evolutionarily dynamic regulatory regions**. *Genome Biol* 2021, **22**(1):62.
1136 45. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z,
1137 Wang J, Ziller MJ *et al*: **Integrative analysis of 111 reference human epigenomes**. *Nature* 2015,
1138 **518**(7539):317-330.
1139 46. Farley EK, Olson KM, Levine MS: **Regulatory Principles Governing Tissue Specificity of**
1140 **Developmental Enhancers**. *Cold Spring Harb Symp Quant Biol* 2015, **80**:27-32.
1141 47. Jindal GA, Farley EK: **Enhancer grammar in development, evolution, and disease: dependencies**
1142 **and interplay**. *Dev Cell* 2021, **56**(5):575-587.
1143 48. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R: **Absence of a simple code:**
1144 **How transcription factors read the genome**. *Trends in Biochemical Sciences* 2014, **39**(9)381-399.
1145 49. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S,
1146 Martinez-Jimenez CP, Mackay S *et al*: **Five-vertebrate ChIP-seq reveals the evolutionary**
1147 **dynamics of transcription factor binding**. *Science* 2010, **328**(5981):1036-1040.
1148 50. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves n, Kutter C, Brown GD, Marshall A,
1149 Flicek P, Odom DT: **Waves of retrotransposon expansion remodel genome organization and**
1150 **CTCF binding in multiple mammalian lineages**. *Cell* 2012, **148**:335-348.
1151 51. Schwalie PC, Ward MC, Cain CE, Faure AJ, Gilad Y, Odom DT, Flicek P: **Co-binding by YY1**
1152 **identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate**
1153 **genomes.** *Genome Biology* 2013, **14**:1-15.
1154 52. Alam T, Agrawal S, Severin J, Young RS, Andersson R, Arner E, Hasegawa A, Lizio M, Ramilowski
1155 JA, Abugessaisa I *et al*: **Comparative transcriptomics of primary cells in vertebrates**. *Genome*
1156 *Res* 2020, **30**(7):951-961.
1157 53. LeCun Y, Jackel LD, Boser B, Denker JS, Graf HP, Guyon I, Henderson D, Howard RE, Hubbard W:
1158 **Handwritten digit recognition: applications of neural network chips and automatic learning**.
1159 *IEEE Communications Magazine* 1989, **27**(11):41-46.
1160 54. Chen L, Fish AE, Capra JA: **Prediction of gene regulatory enhancers across species reveals**
1161 **evolutionarily conserved sequence properties**. *PLoS Comput Biol* 2018, **14**(10):e1006484.
1162 55. Huh I, Mendizabal I, Park T, Yi SV: **Functional conservation of sequence determinants at rapidly**
1163 **evolving regulatory regions across mammals**. *PLoS Comput Biol* 2018, **14**(10):e1006451.
1164 56. Kelley DR: **Cross-species regulatory sequence activity prediction**. *PLoS Comput Biol* 2020,
1165 **16**(7):e1008050.
1166 57. Minnoye L, Taskiran II, Mauduit D, Fazio M, Van Aerschot L, Hulselmans G, Christiaens V,
1167 Makhzami S, Seltenhammer M, Karras P *et al*: **Cross-species analysis of enhancer logic using**
1168 **deep learning**. *Genome Res* 2020.
1169 58. Quang D, Xie X: **DanQ: a hybrid convolutional and recurrent deep neural network for**
1170 **quantifying the function of DNA sequences**. *Nucleic Acids Res* 2016, **44**(11):e107.

1171 59. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J: **A novel connectionist**
1172 **system for unconstrained handwriting recognition**. *IEEE Trans Pattern Anal Mach Intell* 2009,
1173 **31**(5):855-868.
1174 60. Klein JC, Keith A, Agarwal V, Durham T, Shendure J: **Functional characterization of enhancer**
1175 **evolution in the primate lineage**. *Genome Biol* 2018, **19**(1):99.
1176 61. Wirthlin M, Kaplow, Irene M., Lawler, Alyssa J., He, J., Phan., BaDoi, N., Brown, Ashley R.,
1177 Stauffer, William R., Pfenning, Andreas R.: **The Regulatory Evolution of the Primate Fine-Motor**
1178 **System**. *bioRχiv* 2020.
1179 62. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ: **ATAC-seq: A Method for Assaying Chromatin**
1180 **Accessibility Genome-Wide**. *Curr Protoc Mol Biol* 2015, **109**:21.29.21-21.29.29.
1181 63. John S, Sabo PJ, Canfield TK, Lee K, Vong S, Weaver M, Wang H, Vierstra J, Reynolds AP,
1182 Thurman RE *et al*: **Genome-scale mapping of DNase I hypersensitivity**. *Curr Protoc Mol Biol*
1183 2013, **Chapter 27**:Unit 21.27.
1184 64. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human**
1185 **genome.** *Nature* 2012, **489**(7414):57-74.
1186 65. Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru MD, Reach SM, Motl J, Ehrlich ME, Hurd YL,
1187 Roussos P: **An atlas of chromatin accessibility in the adult human brain**. *Genome Res* 2018,
1188 **28**(8):1243-1252.
1189 66. Srinivasan C, Phan BN, Lawler AJ, Ramamurthy E, Kleyman M, Brown AR, Kaplow IM, Wirthlin
1190 ME, Pfenning AR: **Addiction-associated genetic variants implicate brain cell type- and region-**
1191 **specific cis-regulatory elements in addiction neurobiology**. *Journal of Neuroscience* In press.
1192 67. Armstrong J, Hickey G, Diekhans M, Deran A, Fang Q, Xie D, Feng S, Stiller J, Genereux D,
1193 Johnson J *et al*: **Progressive alignment with Cactus: a multiple-genome aligner for the**
1194 **thousand-genome era**. *Nature* 2020, **587**(7833):246-251.
1195 68. Arvey A, Agius P, Noble WS, Leslie C: **Sequence and chromatin determinants of cell-type-**
1196 **specific transcription factor binding**. *Genome Res* 2012, **22**(9):1723-1734.
1197 69. Alipanahi B, Delong A, Weirauch MT, Frey BJ: **Predicting the sequence specificities of DNA- and**
1198 **RNA-binding proteins by deep learning**. *Nat Biotechnol* 2015, **33**(8):831-838.
1199 70. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**(7553):436-444.
1200 71. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D,
1201 Gonzalez JN, Guruvadoo L *et al*: **The UCSC Genome Browser database: 2017 update**. *Nucleic*
1202 *Acids Res* 2017, **45**(D1):D626-D634.
1203 72. Halstead MM, Kern C, Saelao P, Wang Y, Chanthavixay G, Medrano JF, Van Eenennaam AL, Korf
1204 I, Tuggle CK, Ernst CW *et al*: **A comparative analysis of chromatin accessibility in cattle, pig, and**
1205 **mouse tissues**. *BMC Genomics* 2020, **21**(1):698.
1206 73. Chua CEL, Tang BL: **Syntaxin 16 is enriched in neuronal dendrites and may have a role in**
1207 **neurite outgrowth**. *Molecular Membrane Biology* 2008, **25**(1):35-45.
1208 74. Ray M, Zhang W: **Analysis of Alzheimer's disease severity across brain regions by topological**
1209 **analysis of gene co-expression networks**. *BMC Systems Biology* 2010, **4**(1):1-11.
1210 75. Chen L, Liu Z, Zhou B, Wei C, Zhou Y, Rosenfeld MG, Fu X-D, Chisholm AD, Jin Y: **CELF RNA**
1211 **binding proteins promote axon regeneration in C. elegans and mammals through alternative**
1212 **splicing of Syntaxins**. *eLife* 2016, **5**:e16072.
1213 76. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W: **Quantifying similarity between motifs**.
1214 *Genome Biology* 2007, **8**(2):R24.
1215 77. Mookherjee D, Majumder P, Mukherjee R, Chatterjee D, Kaul Z, Das SD, Sougrat R, Chakrabarti
1216 S, Chakrabarti O: **Cytosolic aggregates in presence of non-translocated proteins perturb**
1217 **endoplasmic reticulum structure and dynamics.** *Traffic* 2019, **20**(12):943-960.

1218 78. Spitz F, Gonzalez F, Duboule D: **A Global Control Region Defines a Chromosomal Regulatory**
1219 **Landscape Containing the HoxD Cluster**. *Cell* 2003, **113**(3):405-417.
1220 79. Breuss MW, An N, Song Q, Nguyen T, Stanley V, James KN, Musaev D, Chai G, Wirth SA,
1221 Anzenberg P *et al*: **Mutations in LNPK, Encoding the Endoplasmic Reticulum Junction Stabilizer**
1222 **Lunapark, Cause a Recessive Neurodevelopmental Syndrome: The American Journal of Human**
1223 **Genetics**. *American Journal of Human Genetics* 2018, **103**(2):296-304.
1224 80. Zhang R, Wang Y, Li R, Chen G: **Transcriptional Factors Mediating Retinoic Acid Signals in the**
1225 **Control of Energy Metabolism**. *Int J Mol Sci* 2015, **16**(6):14210-14244.
1226 81. Stossi F, Dandekar RD, Johnson H, Lavere P, Foulds CE, Mancini MG, Mancini MA: **Tributyltin**
1227 **chloride (TBT) induces RXRA down-regulation and lipid accumulation in human liver cells**. *PLoS*
1228 *One* 2019, **14**(11):e0224405.
1229 82. Xue Y, Guo C, Hu F, Zhu W, Mao S: **PPARA/RXRA signalling regulates the fate of hepatic non-**
1230 **esterified fatty acids in a sheep model of maternal undernutrition**. *Biochim Biophys Acta Mol*
1231 *Cell Biol Lipids* 2020, **1865**(2):158548.
1232 83. Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P: **Complexity and conservation of regulatory**
1233 **landscapes underlie evolutionary resilience of mammalian gene expression**. *Nat Ecol Evol*
1234 2018, **2**(1):152-163.
1235 84. Govaere O, Cockell S, Van Haele M, Wouters J, Van Delm W, Van den Eynde K, Bianchi A, van
1236 Eijsden R, Van Steenbergen W, Monbaliu D *et al*: **High-throughput sequencing identifies**
1237 **aetiology-dependent differences in ductular reaction in human chronic liver disease**. *J Pathol*
1238 2019, **248**(1):66-76.
1239 85. Chen G, Wang R, Chen H, Wu L, Ge RS, Wang Y: **Gossypol ameliorates liver fibrosis in diabetic**
1240 **rats induced by high-fat diet and streptozocin**. *Life Sci* 2016, **149**:58-64.
1241 86. Hernandez-Gea V, Friedman SL: **Pathogenesis of liver fibrosis**. *Annu Rev Pathol* 2011, **6**:425-456.
1242 87. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-
1243 Petri A, Kircher M *et al*: **The evolution of gene expression levels in mammalian organs**. *Nature*
1244 2011, **478**(7369):343-348.
1245 88. Benjamini Y and Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful**
1246 **Approach to Multiple Testing.** *Journal of the Royal Statistical Society* 1995, **57**(1):289-300.
1247 89. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential**
1248 **expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Res* 2015,
1249 **43**(7):e47.
1250 90. Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, Baldarelli RM, Beal JS, Campbell J,
1251 Corbani LE *et al*: **The mouse Gene Expression Database (GXD): 2019 update**. *Nucleic Acids Res*
1252 2019, **47**(D1):D774-D779.
1253 91. **The Human Protein Atlas.** https://www.proteinatlas.org/.
1254 92. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf
1255 C, Sjostedt E, Asplund A *et al*: **Tissue-based map of the human proteome**. *Science* 2015,
1256 **347**:1260419-1260419.
1257 93. Chaubey PM, Hofstetter L, Roschitzki B, Stieger B: **Proteomic Analysis of the Rat Canalicular**
1258 **Membrane Reveals Expression of a Complex System of P4-ATPases in Liver**. *PLoS One* 2016,
1259 **11**(6):e0158033.
1260 94. Beagan JA, Pastuzyn ED, Fernandez LR, Guo MH, Feng K, Titus KR, Chandrashekar H, Shepherd
1261 JD, Phillips-Cremins JE: **Three-dimensional genome restructuring across timescales of activity-**
1262 **induced neuronal gene expression**. *Nat Neurosci* 2020, **23**(6):707-717.
1263 95. Boulting GL, Durresi E, Ataman B, Sherman MA, Mei K, Harmin DA, Carter AC, Hochbaum DR,
1264 Granger AJ, Engreitz JM *et al*: **Activity-dependent regulome of human GABAergic neurons**

1265 **reveals new patterns of gene regulation and neurological disease heritability**. *Nat Neurosci*
1266 2021, **24**(3):437-448.
1267 96. Wang AW, Wang YJ, Zahm AM, Morgan AR, Wangensteen KJ, Kaestner KH: **The Dynamic**
1268 **Chromatin Architecture of the Regenerating Liver**. *Cell Mol Gastroenterol Hepatol* 2020,
1269 **9**(1):121-143.
1270 97. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: **Identifying a high fraction**
1271 **of the human genome to be under selective constraint using GERP++**. *PLoS Comput Biol* 2010,
1272 **6**(12):e1001025.
1273 98. Long HK, Prescott SL, Wysocka J: **Ever-Changing Landscapes: Transcriptional Enhancers in**
1274 **Development and Evolution**. *Cell* 2016, **167**(5):1170-1187.
1275 99. Wang B, Ke W, Guang J, Chen G, Yin L, Deng S, He Q, Liu Y, He T, Zheng R *et al*: **Firing Frequency**
1276 **Maxima of Fast-Spiking Neurons in Human, Monkey, and Mouse Neocortex**. *Front Cell*
1277 *Neurosci* 2016, **10**:239.
1278 100. Kalmbach BE, Buchin A, Long B, Close J, Nandi A, Miller JA, Bakken TE, Hodge RD, Chong P, de
1279 Frates R *et al*: **h-Channels Contribute to Divergent Intrinsic Membrane Properties of**
1280 **Supragranular Pyramidal Neurons in Human versus Mouse Cerebral Cortex**. *Neuron* 2018,
1281 **100**(5):1194-1208.e5.
1282 101. Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC *et al*:
1283 **Dense sampling of bird diversity increases power of comparative genomics**. *Nature* 2020,
1284 **587**(7833):252-257.
1285 102. Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G: **A "forward genomics"**
1286 **approach links genotype to phenotype using independent phenotypic losses among related**
1287 **species**. *Cell Rep* 2012, **2**(4):817-823.
1288 103. Sudmant PH, Alexis MS, Burge CB: **Meta-analysis of RNA-seq expression data across species,**
1289 **tissues and studies**. *Genome Biol* 2015, **16**:287.
1290 104. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK,
1291 Narayanan AK *et al*.: **The Encyclopedia of DNA elements (ENCODE): data portal update.** *Nucleic*
1292 *Acids Res* 2018, **46**(D1):D794-D801.
1293 105. Giuffra E, Tuggle CK, FAANG Consoritum: **Functional Annotation of Animal Genomes (FAANG):**
1294 **Current Achievements and Roadmap**. *Annu Rev Anim Biosci* 2019, **7**:65-88.
1295 106. Tosches MA, Yamawaki TM, Naumann RK, Jacobi AA, Tushev G, Laurent G: **Evolution of pallium,**
1296 **hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles**. *Science*
1297 2018, **360**(6391):881-888.
1298 107. Zhu Y, Sousa AMM, Gao T, Skarica M, Li M, Santpere G, Esteller-Cucala P, Juan D, Ferrández-
1299 Peral L, Gulden FO *et al*: **Spatiotemporal transcriptomic divergence across human and**
1300 **macaque brain development**. *Science* 2018, **362**(6420).
1301 108. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.**
1302 *Bioinformatics* 2010, **26**(6):841-842.
1303 109. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D,
1304 Zadissa A, Searle S *et al*: **GENCODE: The reference human genome annotation for the ENCODE**
1305 **project**. *Genome Research* 2012, **22**(9):1760-1774.
1306 110. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright
1307 J, Armstrong J *et al*: **GENCODE reference annotation for the human and mouse genomes**.
1308 *Nucleic Acids Res* 2019, **47**(D1):D766-D773.
1309 111. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication,**
1310 **deletion, and rearrangement in the mouse and human genomes**. *Proc Natl Acad Sci U S A* 2003,
1311 **100**(20):11484-11489.

112. Hickey G, Paten B, Earl D, Zerbino D, Haussler D: **HAL: a hierarchical format for storing and analyzing multiple genome alignments**. *Bioinformatics* 2013, **29**(10):1341-1342.
113. Zhang X, Kaplow IM, Wirthlin M, Park TY, Pfenning AR: **HALPER facilitates the identification of regulatory element orthologs across species**. *Bioinformatics* 2020, **36**(15):4339-4340.
114. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W *et al*: **Model-based Analysis of ChIP-Seq (MACS)**. *Genome Biology* 2008, **9**(9):R137.
115. Bailey TL, Machanick P: **Inferring direct DNA binding from ChIP-seq**. *Nucleic Acids Res* 2012, **40**(17):e128.
116. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D: **Cactus: Algorithms for genome multiple sequence alignment**. *Genome Res* 2011, **21**(9):1512-1528.
117. Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning-based sequence model**. *Nat Methods* 2015, **12**(10):931-934.
118. Kelley DR, Snoek J, Rinn JL: **Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks**. *Genome Research* 2016, **26**:990-999.
119. He K, Zhang X, Ren S, Sun J: **Delving deep into rectifiers: Surpassing human-level performance on imagenet classification**. *Proceedings of the IEEE International Conference on Computer Vision* 2016, **11**:1026-1034.
120. Chollet F: **Keras**. https://keras.io/.
121. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ *et al*: **Array programming with NumPy.** *Nature* 2020, **585**:357–362
122. Team TTD, Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F, Bayer J, Belikov A *et al*: **Theano: A Python framework for fast computation of mathematical expressions**. *arXiv* 2016.
123. Pedregosa F, Varoquaux G: **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research* 2011, **12**(85):2825-2830.
124. Grau J, Grosse I, Keilwagen J: **PRROC: Computing and visualizing Precision-recall and receiver operating characteristic curves in R**. *Bioinformatics* 2015, **31**(15):2595-2597.
125. National Library of Medicine (US) NCBI Assembly. Bethesda, MD. https://www.ncbi.nlm.nih.gov/assembly.
126. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The Human Genome Browser at UCSC**. *Genome Research* 2002, **12**(6)996-1006.
127. Kumar S, Stecher G, Suleski M, Hedges SB: **TimeTree: A Resource for Timelines, Timetrees, and Divergence Times**. *Mol Biol Evol* 2017, **34**(7):1812-1819.
128. Frey BJ, Dueck D: **Clustering by passing messages between data points**. *Science* 2007, **315**(5814):972-976.
129. Sheffield NC, Bock C: **LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor**. *Bioinformatics* 2015, **32**(4):587-589.
130. Liu C, Wang M, Wei X, Wu L, Xu J, Dai X, Xia J, Cheng M, Yuan Y, Zhang P *et al*: **An ATAC-seq atlas of chromatin accessibility in mouse tissues**. *Sci Data* 2019, **6**(1):65.
131. Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T *et al*: **An encyclopedia of mouse DNA elements (Mouse ENCODE)**. *Genome Biol* 2012, **13**(8):418.
132. **PhyloPic.** http://phylopic.org/.

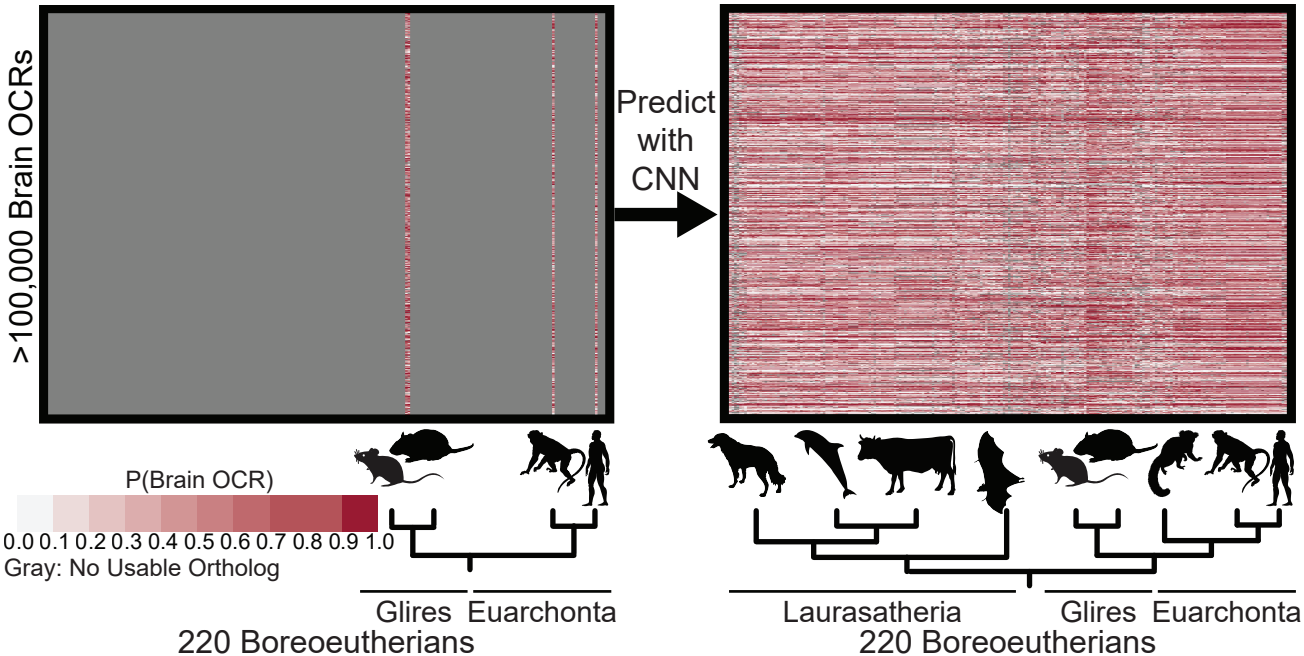

>100,000 Brain OCRs
Predict with CNN
P(Brain OCR)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Gray: No Usable Ortholog
Glires
Euarchonta
220 Boreoeutherians
Laurasatheria
Glires
Euarchonta
220 Boreoeutherians

Figure 1

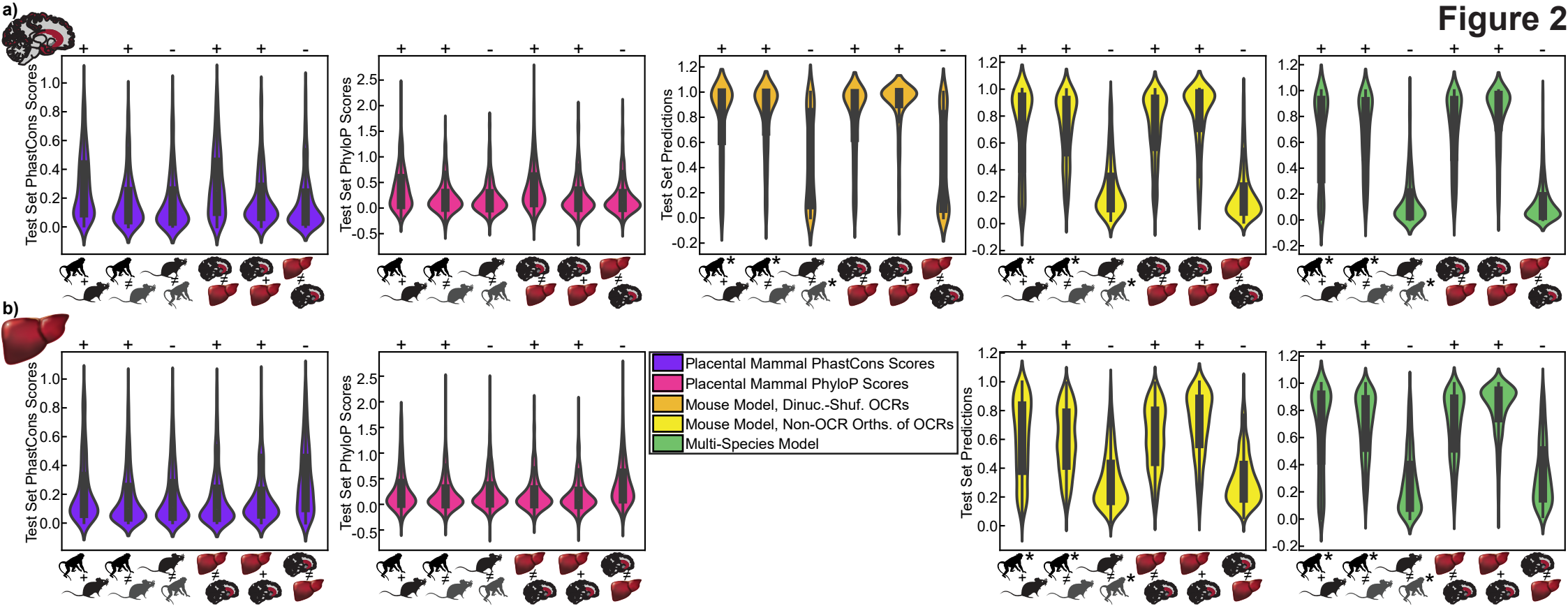


Figure 2
a)
b)
Test Set PhastCons Scores
Test Set PhyloP Scores
Test Set Predictions
Placental Mammal PhastCons Scores
Placental Mammal PhyloP Scores
Mouse Model, Dinuc.-Shuf. OCRs
Mouse Model, Non-OCR Orths. of OCRs
Multi-Species Model

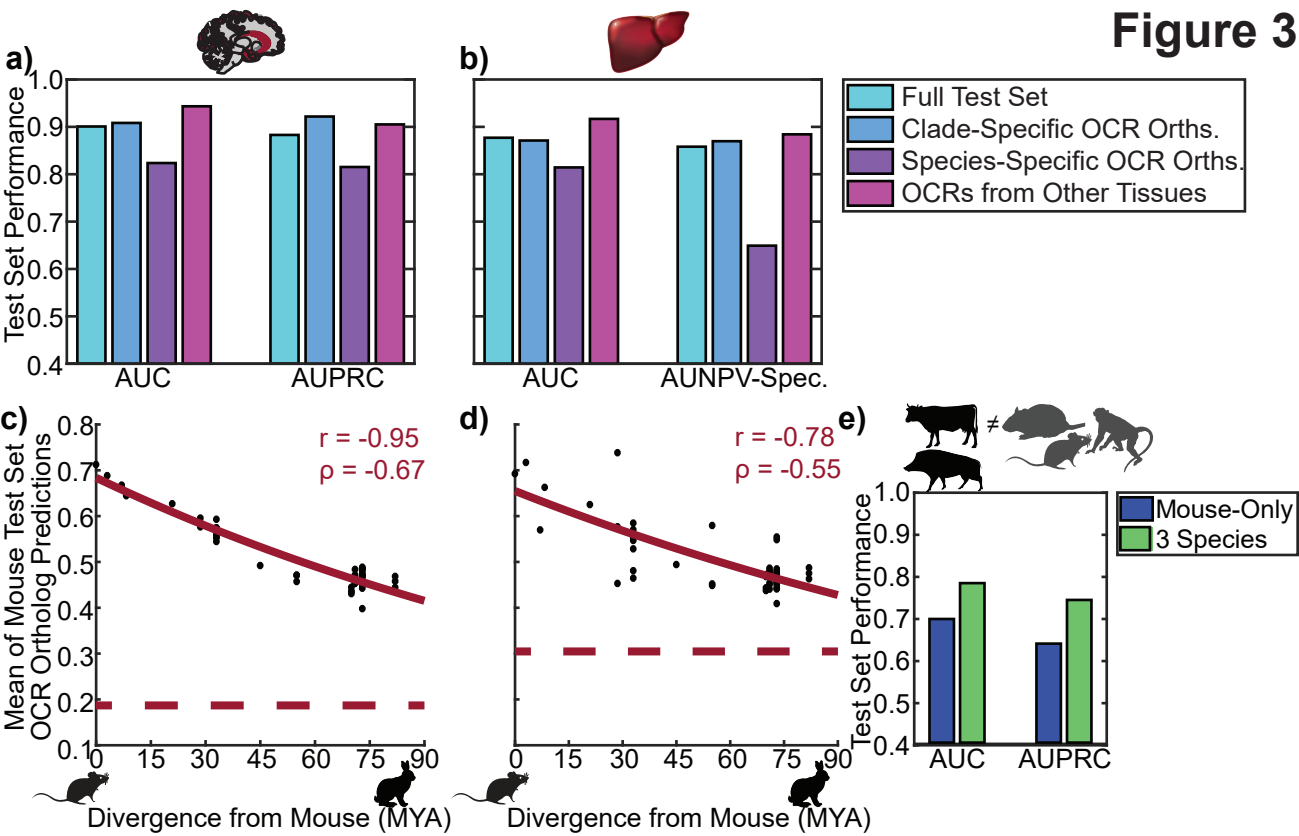


Figure 3
a)
b)
Test Set Performance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
AUC
AUPRC
AUC
AUNPV-Spec.
Full Test Set
Clade-Specific OCR Orths.
Species-Specific OCR Orths.
OCRs from Other Tissues
c)
d)
e)
Mean of Mouse Test Set OCR Ortholog Predictions
r = -0.95
ρ = -0.67
r = -0.78
ρ = -0.55
0
15
30
45
60
75
90
Divergence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

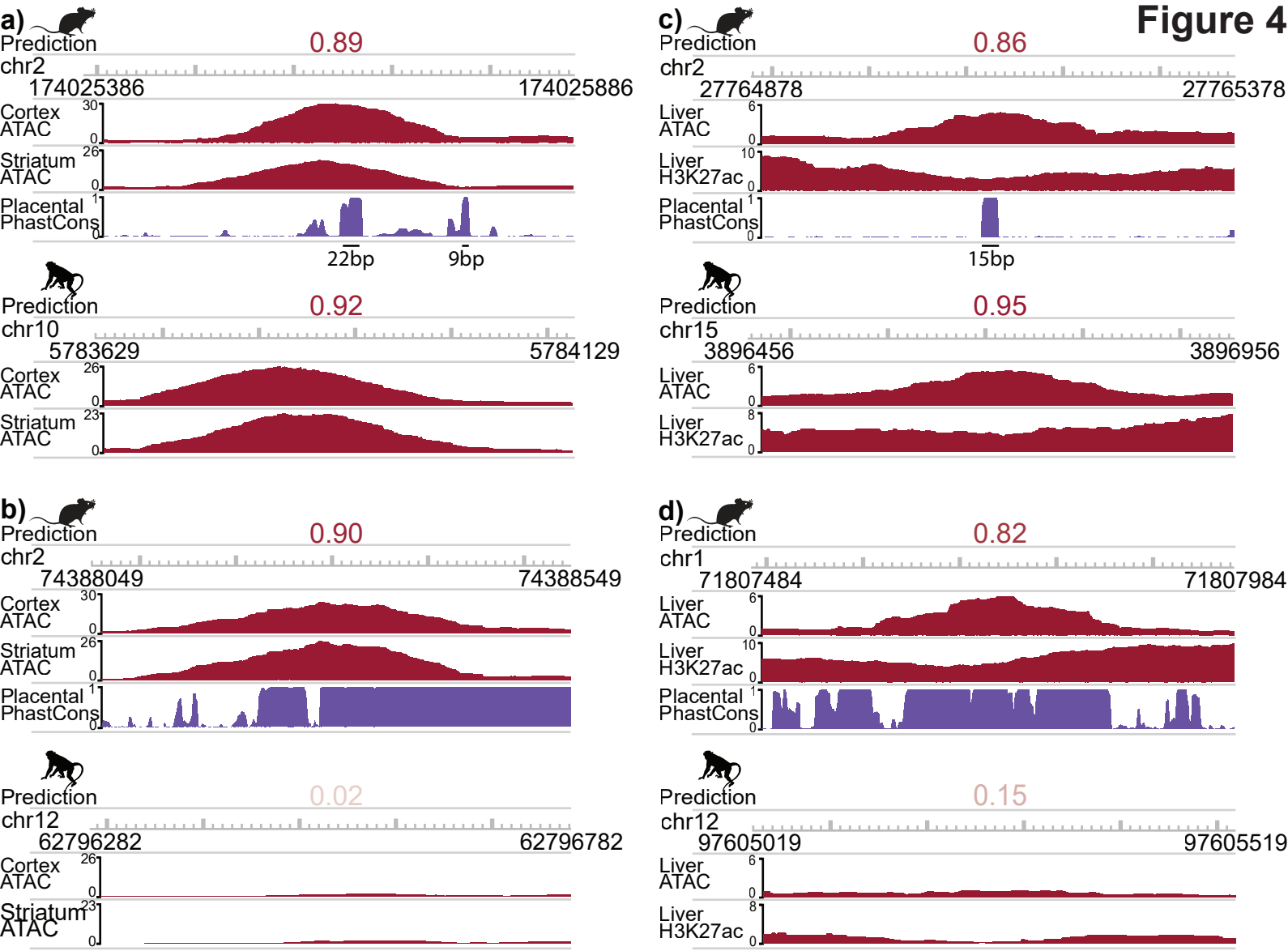
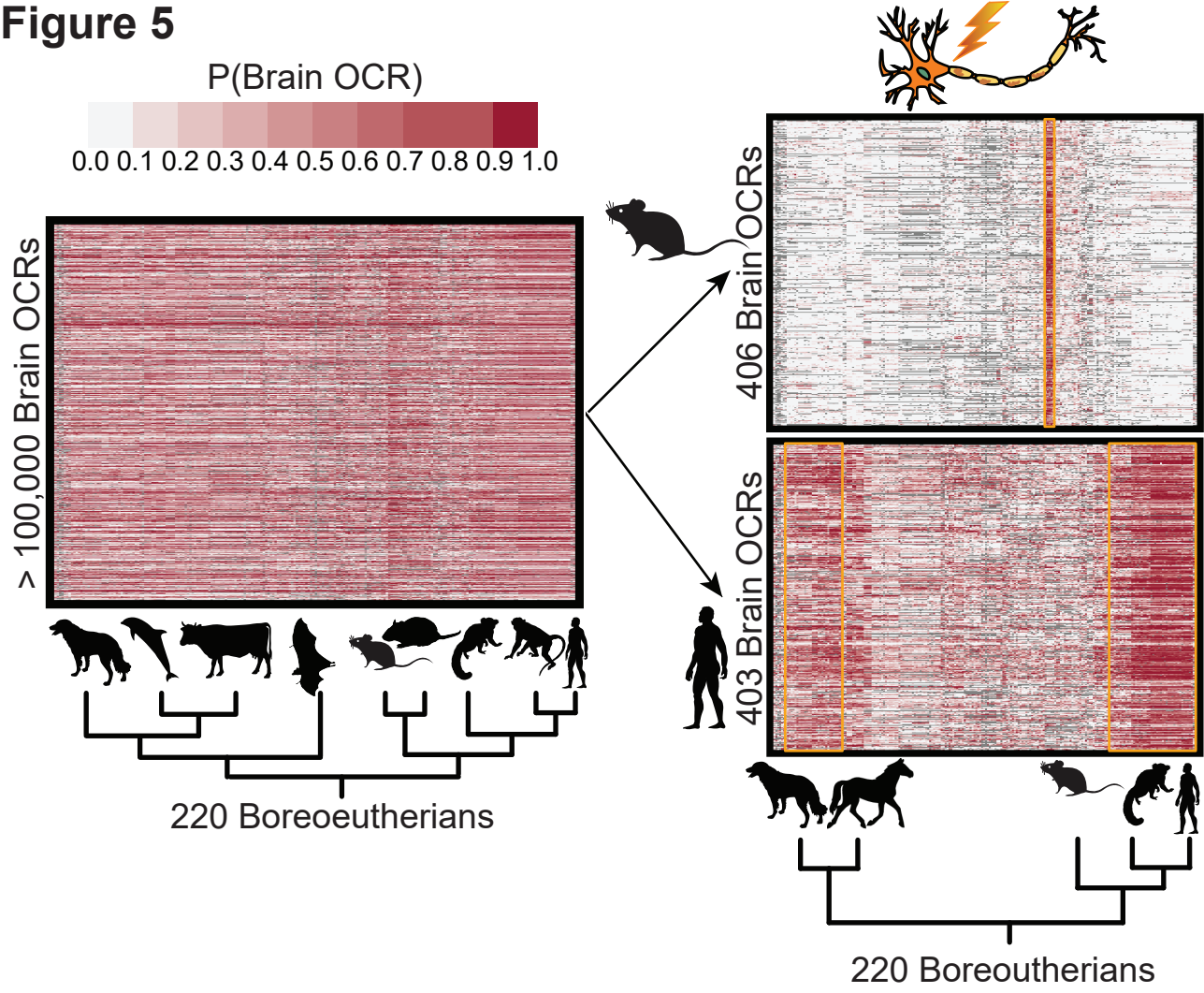


Figure 4
a)
Prediction
0.89
chr2
174025386
174025886
Cortex ATAC
Striatum ATAC
Placental PhastCons
22bp
9bp
Prediction
0.92
chr10
5783629
5784129
Cortex ATAC
Striatum ATAC
b)
Prediction
0.90
chr2
74388049
74388549
Cortex ATAC
Striatum ATAC
Placental PhastCons
Prediction
0.02
chr12
62796282
62796782
Cortex ATAC
Striatum ATAC
c)
Prediction
0.86
chr2
27764878
27765378
Liver ATAC
Liver H3K27ac
Placental PhastCons
15bp
Prediction
0.95
chr15
3896456
3896956
Liver ATAC
Liver H3K27ac
d)
Prediction
0.82
chr1
71807484
71807984
Liver ATAC
Liver H3K27ac
Placental PhastCons
Prediction
0.15
chr12
97605019
97605519
Liver ATAC
Liver H3K27ac

Figure 5


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