TITLE

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

AUTHORS

Irene M. Kaplow[1,3,*], 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, 5000 Forbes Avenue, Pittsburgh, PA 15213

*Corresponding authors

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

Morgan E. Wirthlin: mwirthlin@cmu.edu

Alyssa J. Lawler: alawler@andrew.cmu.edu

Ashley R. Brown: arb1@andrew.cmu.edu

Michael Kleyman: mkleyman@andrew.cmu.edu

Andreas R. Pfenning: apfenning@cmu.edu

ABSTRACT

Many phenotypes have evolved through gene expression, meaning that differences between species are caused in part by differences in enhancers. Here, we demonstrate that we can accurately predict differences between species in open chromatin status at putative enhancers using machine learning models trained on genome sequence across species. We present a new set of criteria that we designed to explicitly demonstrate if models are useful for studying open chromatin regions whose orthologs are not open in every species. Our approach and evaluation metrics can be applied to any tissue or cell type with open chromatin data available from multiple species.

KEYWORDS

Gene expression evolution, open chromatin prediction, machine learning

BACKGROUND

The molecular biology mechanisms underlying the incredible phenotypic diversity across mammals are largely unknown. To study these mechanisms, many consortia, including the Vertebrate Genomes Project, the Genome 10K Project [1], the Bat 1K Project [2], and the Zoonomia Project [3], are sequencing, assembling, and aligning [4] genomes from hundreds of mammals, including endangered species and species that live in remote parts of the world. Using these data, we can investigate mammalian

38 phenotypic diversity by comparing the DNA sequences of species whose most recent common ancestors
39 lived tens of millions of years ago. A large component of phenotypic evolution is mediated by differences
40 in *cis*-regulatory elements, the vast majority of which are enhancers that control gene expression [5-7].
41 Consistent with that understanding of evolution, many complex phenotypes, including vocal learning [8],
42 domestication [9, 10], longevity [11-13], brain size [14], vision [15, 16], echolocation [17], and monogamy
43 [18], are associated with differential gene expression between species. This understanding is further
44 supported by recent studies of transcription factor (TF) binding across species that identify TF binding
45 differences that could be underlying differences in the regulatory activity of enhancers [19-21]. Therefore,
46 to elucidate the ways in which complex phenotypes have evolved, new methods are required that link
47 genome sequence differences at *cis*-regulatory elements to differences in enhancer function.

48 Much of our knowledge of enhancers comes from regulatory genomics measurements that are
49 associated with enhancer activity, especially the ATAC-Seq and DNase hypersensitivity assays for open
50 chromatin and chromatin immunoprecipitation sequencing (ChIP-Seq) for the histone modifications
51 H3K27ac and H3K4me1 [22-25]. These studies have demonstrated that enhancers, relative to genes, are
52 substantially more tissue- or cell type-specific [26] and generally less conserved across species [27, 28].
53 Thus, identifying enhancers through direct experimentation or through comparative genomic annotation
54 are both challenging. To overcome these challenges, multiple recent studies have described machine
55 learning models that use DNA sequences underlying likely enhancers from a small number of mammals
56 to predict whether DNA sequences are likely to be enhancers in other mammals. These studies' success
57 suggests that the *trans*-regulatory environment involved in transcriptional regulation is highly conserved
58 across mammals [29-31].

59 These studies have used models that do not require substantial prior knowledge of important
60 sequence features associated with enhancer activity because many of these sequence features have not
61 yet been discovered. For instance, the presence of known TF motifs only partially explains whether a

62 region is an enhancer [32]. Beyond TF motif presence or absence, enhancer activity is also influenced by
63 many other factors, including TF co-binding events in which TFs do not bind to their full motifs,
64 nucleosome positioning, and DNA shape [32-34]. In one recent study, support vector machines (SVMs)
65 and convolutional neural networks (CNNs) [35] – methods that do not require explicit DNA sequence
66 featurization – were able to predict which 3kb windows have the enhancer-associated histone
67 modification H3K27ac in brain, liver, and limb tissue of human, macaque, and mouse. Importantly, the
68 study found that models trained in one mammal achieved high accuracy in another mammal in the same
69 clade and on another mammal in a different clade, suggesting that the regulatory code in all three of these
70 tissues is highly conserved across mammals [30]. Two other studies have obtained similar results using
71 another proxy for enhancer activity: open chromatin regions (OCRs). One study found that training CNNs
72 on OCRs from multiple mammals had better performance than training CNNs on OCRs from a single
73 mammal, albeit using 131,072bp sequences as input. The boost in power from incorporating multiple
74 species generalized to predicting TF binding strength from ChIP-seq data and gene expression from RNA-
75 seq data [31]. An additional study found that a combined CNN-recurrent neural network [36, 37] trained
76 on sequences underlying 500bp OCRs from melanoma cell lines in one species can accurately predict
77 melanoma cell line open chromatin in other species at a wide range of genetic distances from the training
78 species, including in parts of the genome with low sequence conservation between the training and
79 evaluation species. The study identified an enhancer near the dog melanoma gene *APPL2* that is active in
80 dog melanocytes, but its human ortholog is not active in melanocytes. The study found that this species-
81 specific difference in open chromatin was accurately predicted (in melanocytes), demonstrating the value
82 in accurately predicting differences in open chromatin between orthologous regions [29].

83 While these studies represent major advances in cross-species enhancer prediction, they have yet
84 to demonstrate an ability to identify sequence differences between species that are associated with
85 differences in regulatory genomic measurements of enhancer activity. To study gene expression

86 evolution and, as an earlier study suggested, potentially gain insight into disease [29], it is necessary to
87 accurately identify enhancers in a tissue of interest in one species whose orthologous sequences in
88 another species are not enhancers in that tissue because these enhancers are likely candidates for causing
89 gene expression differences between species. Instead of doing this, previous studies trained a model to
90 predict whether a region is a putative enhancer in comparison to a negative set consisting of random G/C-
91 and repeat-matched regions [30] or enhancers in other cell types [29] in one species and then used the
92 model to make predictions on enhancers and the same type of negative set in another. In fact, no study
93 has performed a systematic, genome-wide evaluation of predictions of enhancer activity of enhancer
94 orthologs with differences in activity across species, so it is unclear whether the models from any of the
95 previous studies can accurately make such predictions. An additional study trained SVMs to predict liver
96 enhancers using dinucleotide-shuffled enhancers as negatives. While the overall performance was good,
97 human enhancers whose orthologs are active in Old World Monkeys but not New World Monkeys were
98 predicted to have consistent activity across all primates, showing that models with good overall
99 performance do not always work well on enhancer orthologs whose activity differs between species [38].

100 Some of the previous studies' methods are also limited because they require long input
101 sequences, with some requiring input sequences > 100kb long [31]. In order for these methods to make
102 predictions for an enhancer ortholog, the enhancer ortholog needs to be on a sufficiently long scaffold.
103 Since many of the new genomes assembled with short reads consist primarily of scaffolds that are less
104 than 100kb long, and some have many scaffolds that are less than 3kb long, some existing methods could
105 not be used for predicting the activity of most enhancer orthologs in these genomes [3, 4, 39].
106 Furthermore, even in genomes with chromosome-level assemblies, identifying orthologs of long
107 sequences is often infeasible due to genome rearrangements, often undetected, that have happened
108 during evolution [40].

109 To predict and evaluate candidate *cis*-regulatory element differences across species, we chose to
110 focus on OCRs, which are only a proxy of regulatory activity, but whose high resolution has the potential
111 to better identify specific genome sequence differences associated with putative regulatory activity. We
112 leveraged new, controlled open chromatin experiments [41], trained a set of new models to predict OCR
113 differences across species, and developed new criteria for evaluating models for this task. First, such
114 models need to achieve high "Lineage-specific OCR accuracy": they need to accurately predict whether
115 differences in sequence between species are associated with differences in open chromatin status, which
116 we evaluate by evaluating performance on species- and clade-specific open chromatin and lack of open
117 chromatin. The subset of OCR orthologs with lineage-specific open chromatin patterns are strong
118 candidates for enhancers involved in the evolution of gene expression. Second, such models need to
119 achieve high "Tissue-specific OCR accuracy": they need to learn both sequence patterns that are specific
120 to the tissue in which the OCRs were identified and non-tissue specific sequence patterns shared across
121 OCRs from multiple tissues. Third, in keeping with the principle of evolutionary parsimony, such models'
122 predictions should have "Phylogeny-matching correlations": their predictions across large numbers of
123 species should approximately match what would be expected based on the species' phylogeny – mean
124 predictions should decrease and standard deviation of predictions should increase with distance to the
125 species with the OCRs. In addition to developing benchmarks, we designed a novel negative set to
126 explicitly to achieve high lineage-specific OCR accuracy: non-OCRs in a tissue whose orthologs in another
127 species are OCRs in that tissue. We used these benchmarks to evaluate how well our novel negative set
128 and multiple previously suggested negative sets work for predicting brain and liver OCR ortholog open
129 chromatin status. For this evaluation, we used 500bp sequences from mouse, human, macaque, and rat
130 brain and mouse, macaque, and rat liver open chromatin peaks that we generated or obtained from [22,
131 41-43] so that we could easily identify and evaluate performance on enhancer orthologs in different
132 species. We found that good performance on chromosomes held out from training does not guarantee

133 good performance for all of our metrics, that our negative set as well as the larger G/C- and repeat-
134 matched negative set had good performance for most metrics, and that machine learning models tend to
135 more accurately predict whether open chromatin status is conserved than OCRs' mean sequence
136 conservation scores. Our approach to OCR ortholog open chromatin status prediction and our method
137 for evaluating approaches for this problem can be applied to any tissue or cell type with open chromatin
138 data from multiple species. We anticipate that the guidelines we propose will encourage researchers to
139 develop and properly evaluate new models for predicting OCR ortholog open chromatin across species,
140 enabling us to uncover transcriptional regulatory mechanisms underlying the evolution of mammalian
141 phenotypic diversity.

142

143

144 **RESULTS**

145

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

148 To demonstrate the ability to predict OCR differences across species, we created a dataset of brain
149 OCRs and their orthologs across dozens of species. To do this, we gathered open chromatin data
150 generated by ATAC-seq [44, 45] or DNase hypersensitivity [46] from two brain regions – cortex and
151 striatum – in four species: *Homo sapiens* [22, 42, 47], *Macaca Mulatta* [41], *Mus musculus* [48], and *Rattus*
152 *norvegicus* [41]. We then defined our OCRs to be the 250bp in each direction of summits of non-exonic
153 cortex open chromatin peaks that (1) overlap striatum open chromatin peaks, (2) are less than 1kb (not
154 super-enhancers), and (3) are at least 20kb from the nearest transcription start site (TSS) so that so they

155 would not overlap promoters. We also identified the orthologs of each OCR in each of the other species
156 collected by the Zoonomia Project [3, 4] (**Figure 1**). Rather than focus on longer regions, which have been
157 shown to work well in some previous studies [30, 31], we used 500bp regions because of the ability shorter
158 regions provide to focus on the impact of local sequence differences, because of the relative ease of
159 obtaining orthologs of shorter regions in genomes with short scaffolds, and because predictions of
160 enhancer activity for shorter regions are easier to experimentally validate.

161 We compared machine learning model performance for five different negative sets that are
162 similar to those used for related tasks: (1) flanking regions [49], (2) OCRs from other tissues [29, 31], (3)
163 about ten times as many G/C- and repeat-matched regions as positives [30], (4) about twice as many G/C-
164 and repeat-matched regions as positives, and (5) ten dinucleotide-shuffled versions of each positive [50]
165 (**Figure 2a**). We additionally created a sixth, novel negative set to force the model to learn signatures of
166 OCRs whose orthologs' open chromatin status differs between species – sequences with closed chromatin
167 in brain of a given species whose orthologs in another species are brain OCRs (we called this "non-OCR
168 orthologs of OCRs," red-brown regions in **Figure 1**) – and included this negative set in our comparison
169 (**Figure 2a**). For a modeling approach, we chose CNNs [35, 51] because they can model complex
170 combinatorial relationships between sequence patterns, and changes in a single TF motif often do not
171 cause changes in open chromatin [32]; because they do not require an explicit featurization of the data,
172 and many sequence patterns involved in brain open chromatin remain to be discovered; and because they
173 can make predictions quickly relative to SVMs, the other leading approach for related tasks [30]. We did
174 the comparison using models trained on only mouse sequences so that we could evaluate their
175 performance on both closely and distantly related species not used for training [30].

176

177 <u>Optimizing and Evaluating Models for Achieving Lineage-Specific OCR Accuracy</u>

178 We evaluated the machine learning model trained on each negative set on its corresponding test
179 set. We found that all models worked well, with every model achieving an AUC > 0.85 and an AUPRC >
180 0.7 (**Figure 2b**). This performance is especially impressive given that ratios of number of negatives to
181 number of positives ranged from approximately 1.2:1 (non-OCR orthologs of OCRs) to approximately 20:1
182 (OCRs in other tissues). The best-performing model was the model with the negative set consisting of
183 dinucleotide-shuffled brain OCRs (**Figure 2b**). However, in this comparison, each model was evaluated on
184 a different negative set, so this evaluation may not be indicative of how useful each model would be in
185 answering questions about gene expression evolution.

186 We therefore also evaluated each model's lineage-specific OCR accuracy, which we did by
187 evaluating models on the OCR orthologs whose brain open chromatin status differs between species.
188 First, we evaluated each model on the subset of mouse brain OCRs whose orthologs in at least one other
189 species are closed (subset of positive set for all models) and the mouse brain closed chromatin regions
190 whose orthologs in at least one other species are open in brain. Interestingly, although all models
191 performed decently on these genomic regions (AUC > 0.65, AUPRC > 0.55), none of the models worked as
192 well on these genomic regions as they did on the test set corresponding to the regions used in training
193 them, and the best-performing model on its own test set – the model trained with dinucleotide-shuffled
194 negatives – was the worst-performing model on these regions (**Figure 2c**). The best-performing model
195 was the model trained on our novel negative set, a negative set that was designed for this task (**Figure**
196 **2c**). We obtained similar results for the subset of mouse brain open and closed chromatin regions whose
197 rat orthologs have the opposite brain open chromatin status (**Supplemental Figure 1a**), showing that
198 some of these models are capable of accurately predicting differences in brain open chromatin between
199 closely related species. These findings reveal a bias in some existing methods for accurately predicting
200 conserved epigenomic features and show that these biases can be mitigated by focusing the training set
201 on regions whose open chromatin status differs across species.

202 To facilitate open chromatin prediction in species for which open chromatin datasets are not
203 available, we also evaluated whether a model trained in one species could make accurate predictions of
204 brain open chromatin status on sequences obtained from another species [30]. However, unlike other
205 previous work that evaluated overall performance [30], we evaluated performance on the subset of
206 regions in another species whose brain open chromatin status differs from that of their orthologs in the
207 training species, as this is necessary for showing that models can accurately predict differences in brain
208 open chromatin status between species. We therefore evaluated our models trained in mouse on
209 macaque brain OCRs whose mouse orthologs are closed and on macaque brain closed chromatin regions
210 whose mouse orthologs are open in brain. We found that all models achieved decent performance (AUC
211 > 0.65, AUPRC > 0.55), with the model trained on the dinucleotide-shuffled brain OCR negatives providing
212 the worst performance and the model trained with our novel negative set providing the best performance
213 (**Figure 2d**), further demonstrating the necessity of evaluating models on OCRs whose orthologs' open
214 chromatin statuses differ between species. We also evaluated our models on human (**Supplemental**
215 **Figure 1b**) and rat (**Supplemental Figure 1c**) regions with different brain open chromatin statuses from
216 their mouse orthologs and found that the models generally did not work as well for such regions but still
217 obtained decent performance and that the relative performance of different negative sets was similar to
218 what it was for the macaque regions.

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

226 tree shrews). We found that all models obtained decent performance for Glires-specific open and closed
227 chromatin regions (AUC > 0.75, AUNPV-Specificity > 0.7), with the model trained on the dinucleotide-
228 shuffled brain OCR negatives providing the worst performance and the models trained with our negative
229 set and the larger G/C- and repeat-matched negative set providing the best performance (**Supplemental**
230 **Figure 1d**). Because we trained our models on sequences from mouse (clade: Glires), we also evaluated
231 how well our models work on clade-specific brain open and closed chromatin regions from Euarchonta
232 species, as this clade was not used in training. We found that the models obtained slightly worse
233 performance on Euarchonta-specific brain open and closed chromatin regions than they did on Glires-
234 specific brain open and closed chromatin regions, with similar relative performances for the models
235 trained on different negative sets (**Figure 2e**). Since many phenotypes are clade-specific, these results
236 demonstrate the necessity of evaluating models for OCR ortholog open chromatin prediction on clade-
237 specific open and closed chromatin regions for clades used and not used in training.

238 Since, for many applications, we need to make a binary classification as to whether a region is
239 open in brain, we also investigated how well-calibrated our models are. We found that models trained
240 on some negative sets — including flanking regions, OCRs in other tissues, the smaller G/C- and repeat-
241 matched set, and dinucleotide-shuffled brain OCRs — tended to do better on clade-specific OCRs than on
242 clade-specific closed chromatin regions. On the other hand, the models trained with the larger G/C- and
243 repeat-matched set and our novel negative set tended to do better on clade-specific closed chromatin
244 regions than on clade-specific OCRs. We tried re-calibrating all of the models with the positive training
245 set and the training set from our novel negative set. For the models trained on all negative sets except
246 for ours, this led to an increase in specificity and a decrease in sensitivity. In general, the increase in
247 specificity was similar to the decrease in sensitivity (**Supplemental Tables 1-6**), but, for the smaller G/C-
248 and repeat-matched region negative set, the increase in specificity was substantially larger than the

249 decrease in sensitivity (**Supplemental Figure 2**). Thus, while some models were poorly calibrated, re-
250 calibrating models with our negative set usually had limited utility.

251

252 Best Overall Model Performance Does Not Guarantee Tissue-Specific OCR Accuracy

253 In addition to evaluating model performance on OCR orthologs whose open chromatin status
254 differs between species, we also determined if the models achieved high tissue-specific OCR accuracy. To
255 determine if our models learned sequence patterns associated with only brain-specific open chromatin,
256 we evaluated our models' predictions for the subset of brain OCRs that do not overlap liver OCRs and the
257 subset of brain OCRs that overlap liver OCRs. Test set predictions from all models for both of these subsets
258 of the positive set were usually close to one (**Figure 3**). To determine if our models learned only sequence
259 patterns that are indicative of general open chromatin, we also evaluated our models' predictions for the
260 liver OCRs that do not overlap brain OCRs. We compared this to the predictions on the negative set. We
261 found that predictions on both of these sets tended to be close to zero. However, the liver, non-brain
262 open chromatin status predictions from the model trained with dinucleotide-shuffled OCR negatives
263 tended to be more evenly distributed between zero and one than the liver, non-brain open chromatin
264 status predictions for the models trained with the other negative sets (**Figure 3**). We also did a comparison
265 of the performances of models trained on different negative sets in which we limited the positive set to
266 brain OCRs that overlap liver OCRs and defined the negative set as liver OCRs that do not overlap brain
267 OCRs. We found that all models worked well (AUC > 0.75, AUPRC > 0.6) on mouse as well as on macaque
268 and rat, which were not used in training, with the model trained on dinucleotide-shuffled brain OCR
269 negatives providing the worst performance (**Supplemental Figure 3a**). In addition, we found that the
270 models that tended to work better on positives were more accurate for shared brain and liver OCRs than
271 for liver, non-brain OCRs and that the models that tended to work better on brain closed chromatin

272 orthologs of brain OCRs were more accurate for liver, non-brain OCRs than for shared brain and liver OCRs
273 (**Supplemental Tables 8-9**, **Supplemental Figure 3b**). Calibration with the positive training set and the
274 training set from our novel negative set had a similar effect to calibration for clade-specific open and
275 closed chromatin regions (**Supplemental Tables 8-9**, **Supplemental Figure 3b**). These results show that
276 models with good overall performance are not always able to comprehensively learn open chromatin
277 sequence patterns that are specific to the tissue in which they were trained.

278

279 Predictions from Models of OCR Ortholog Open Chromatin Status Have Phylogeny-Matching Correlations

280 We also determined whether our models' predictions have phylogeny-matching correlations in a
281 way that does not require open chromatin data from multiple species. To do this, we obtained the
282 orthologs of the mouse brain OCRs in all of the fifty-six Glires clade species in the Zoonomia Project [3, 4],
283 used our machine learning models to predict the brain open chromatin status of these orthologs,
284 computed the mean brain open chromatin status across all brain OCR orthologs in each species, and
285 computed the correlation between mean predicted brain open chromatin status and evolutionary
286 distance from mouse. As we expected based on the principle of evolutionary parsimony, all models
287 showed a strong negative correlation between mean predicted brain open chromatin status and
288 divergence from mouse (**Figure 4**, **Supplemental Figure 4a**). Nevertheless, there is still more open
289 chromatin at these brain OCR orthologs than would be expected from brain non-OCRs, even in the most
290 distantly related Glires species, because all mean predictions are greater than the mean predictions for
291 the negative test sets (**Figure 4, Supplemental Figure 4a**). We also expected there to be a strong positive
292 correlation between the standard deviation of open chromatin status and divergence from mouse
293 because most brain OCR orthologs in species closely related to mouse are active in brain, while the brain

294 open chromatin status of brain OCR orthologs in species that are more distantly related should vary. We
295 found this expected positive correlation for all negative sets (**Supplemental Figure 4b**).

296

297 Machine Learning Models Learned Motifs of Important Brain Transcription Factors

298 To determine what sequence patterns our models were prioritizing, we ran DeepLIFT with the
299 rescale rule [52] followed by TF-MoDISco [53] on the positive examples from the validation set from each
300 model and compared the results to known motifs. All of the models seemed to have learned motifs of
301 TFs that are known to play important roles in the brain, including Ctcf [54, 55], *Fos* [56-58], *Egr2* [59, 60],
302 and *Rfx4* [61-63] (**Supplemental Figure 5**). All of the models except for those trained with flanking region
303 negatives (**Supplemental Figure 5a**) and those trained with dinucleotide-shuffled brain OCR negatives
304 (**Supplemental Figure 5e**) seemed to also have learned the motif of *Mef2c*, a TF with multiple roles in the
305 brain [64-66] (**Supplemental Figure 5**). The model with OCRs in other tissues as negatives seemed to have
306 learned the depletion of motifs of multiple TFs whose human orthologs are not expressed in the brain,
307 including *Hnf4g*, *Nr5a1*, *Elf3*, and *Foxd2*, and the model with model with the larger number of G/C- and
308 repeat-matched negatives seemed to have learned a depletion of the motif for *Nr2f6*, which has very low
309 expression in the brain [67, 68] (**Supplemental Figures 5b-c**). The model with the dinucleotide-shuffled
310 brain OCR negatives seemed to have learned the motif for *Bcl6* [69, 70], but this motif consists almost
311 exclusively of G's, so it might be indicative of many consecutive G's being more common in brain OCRs
312 than in shuffled brain OCRs (**Supplemental Figure 5e**). The model trained with our novel negative set also
313 seemed to have learned the motif for *Dbp*, which has been implicated in circadian rhythms [71, 72]; two
314 slightly different *Rfx* motifs (also learned by the model with the smaller number of G/C- and repeat-
315 matched negatives), which is not surprising because multiple *Rfx* TFs play important roles in the brain [61,

316 63, 73]; and a depletion of the motif for *Thra* (**Supplemental Figures 5d, f**). It is possible that these
317 apparent differences in motifs learned by the models caused their differences in performance.

318

319 Approach to Evaluating Machine Learning Models for OCR Ortholog Open Chromatin Status Prediction
320 Can Be Applied to Any Tissue or Cell Type with Open Chromatin Data from Multiple Species

321 Although we prototyped our approach to evaluating machine learning models for predicting open
322 chromatin status of OCR orthologs in the brain, this approach can be applied to any tissue or cell type with
323 open chromatin data from multiple species. We therefore applied it to another tissue, the liver, and found
324 that our novel approach to negative set construction also worked well for most metrics. Our positive set
325 for liver was 250bp in each direction of peak summits of our mouse liver ATAC-seq peaks that overlapped
326 liver ATAC-seq peaks from [43]. We obtained negatives by mapping rat and macaque liver ATAC-seq data
327 from [41] to mouse and identifying the mouse orthologs that did not overlap mouse liver ATAC-seq peaks.
328 We found that the model achieved high lineage-specific and tissue-specific accuracy (AUC > 0.7, AUPRC >
329 0.65, **Supplemental Figures 6a-b**). We also determined if our predictions had phylogeny-matching
330 correlations by obtaining orthologs of the mouse liver OCRs in all of the species from the Zoonomia project
331 [3, 4] and predicting their open chromatin statuses. As with brain, we found a strong negative correlation
332 between the predicted mean liver OCR ortholog open chromatin status in Glires species and those species'
333 divergence from mouse (**Supplemental Figure 6c**) and a strong positive correlation between standard
334 deviation of predicted liver OCR ortholog open chromatin status in Glires species and those species'
335 divergence from mouse (**Supplemental Figure 6d**). In addition, we interpreted the model using DeepLIFT
336 with the rescale rule [52] followed by TF-MoDISco [53] and found that the model seemed to have learned
337 motifs of multiple known liver TFs, including *Ctcf* [54, 74], *Ppara* [75-77], and *Cebpa* [78, 79], as well as a
338 depletion the motif for *Wt1*, which is not expressed in liver [67, 68] (**Supplemental Figure 6e**). This

339 illustrates that our novel approach to constructing a negative set for open chromatin status prediction of
340 OCR orthologs works well in multiple tissues.

341

342 Machine Learning Models Predict OCR Orthologs' Open Chromatin Status More Accurately than Mean
343 Conservation Scores

344 Since OCRs whose open chromatin status is conserved across species tend to have higher
345 sequence conservation than those whose open chromatin status is not conserved [80], we compared the
346 accuracy of our machine learning models in predicting OCR open chromatin status conservation to the
347 accuracy of using mean conservation scores. To do this, we identified test set mouse brain and liver OCRs
348 whose macaque orthologs do and do not overlap OCRs in brain and liver, respectively, and computed the
349 mean conservation scores of these OCRs [81, 82] as well as the predictions on test set macaque orthologs
350 of the machine learning models trained in the corresponding tissue with our novel negative set. We found
351 that mean conservation scores and model predictions tended to be higher for the macaque orthologs for
352 which open chromatin status was conserved than those for which open chromatin status was not
353 conserved (**Supplemental Tables 10-11**). For each tissue, we then ranked the macaque OCR orthologs
354 based on their mean conservation scores and their model predictions, with the highest rank
355 corresponding to the highest score or open chromatin status prediction. For the open chromatin status-
356 conserved OCRs in each tissue, we used a Wilcoxon signed-rank test to evaluate whether these OCRs
357 tended to have higher ranks for our predictions than they do for mean conservation scores; we found that
358 the ranks were significantly higher for our predictions (**Tables 1-2, Figure 5**). For each tissue, we also used
359 a Wilcoxon signed-rank test to evaluate whether the OCR orthologs without open chromatin tended to
360 have lower ranks for our predictions than they do for mean conservation scores, and we found that the
361 ranks were significantly lower for our predictions (**Tables 1-2, Figure 5**). We repeated this for human and

362 rat orthologs of mouse brain OCRs with conserved and non-conserved open chromatin statuses and for
363 rat orthologs of mouse liver OCRs with conserved and non-conserved OCR statuses and obtained similar
364 results (**Supplemental Figure 7, Supplemental Tables 10-14**). This shows that using machine learning
365 models for predicting open chromatin status conservation of OCR orthologs can be more accurate than
366 using mean sequence conservation scores.

367

368 Machine Learning Models Trained Using Data from Multiple Species Can Accurately Predict OCR
369 Orthologs' Open Chromatin Statuses

370 Based on other research showing that training models with data from multiple species can
371 improve OCR prediction accuracy [31], we trained additional machine learning models using open
372 chromatin data from multiple species. For each of brain and liver, we used the open chromatin data from
373 all of the species that we had collected as positives (four species for brain, three species for liver) and the
374 orthologs of all these OCRs in the other species for which we had data that did not overlap brain or liver
375 open chromatin, respectively, as negatives. We found that the brain and liver multi-species models
376 achieved high lineage-specific and tissue-specific accuracy, where performance was generally better than
377 the performance for any of the models trained on only mouse sequences (**Figures 6a-b**). In addition, we
378 determined if the multi-species brain and liver models' predictions had phylogeny-matching correlations
379 by using them to predict the OCR ortholog open chromatin status of mouse brain and liver OCRs,
380 respectively, across Glires and found strong negative correlations between divergence from mouse and
381 mean OCR ortholog open chromatin status predictions (**Figures 6c-d**). We also found strong positive
382 correlations between divergence from mouse and standard deviations of OCR ortholog open chromatin
383 status predictions (**Supplemental Figures 8a-b**). When interpreting the multi-species brain model, in
384 addition to the motifs that we found for the model trained on only mouse sequences, we found a

385 depletion of the motifs for *Nr1i3* and *Pit1*, both of which are not expressed in human brain (**Supplemental**
386 **Figure 8c**) [67, 68]. When interpreting the multi-species liver model, in addition to the motifs that we
387 found for the model trained on mouse sequences, we also found the motifs for additional TFs that are
388 known to be involved in the liver, including *Hnf4a* [83-85], *Foxk1* [86, 87], *Ets2* [88, 89], *Sp1* [90, 91],
389 *Onecut1* [92, 93], *Bcl6* [94, 95], and *Nfe2l2* [96, 97], as well as a depletion of the motif for *Dbx1*, which is
390 not expressed in human liver [67, 68], and a depletion of the motif for *Zfp637* (**Supplemental Figure 8d**).
391 Overall, these results suggest that machine learning models trained on data from multiple species can
392 accurately predict open chromatin statuses of OCR orthologs.

393

394 Machine Learning Models Trained with Data from Multiple Species Make More Accurate Predictions than
395 Mean Conservation Scores

396 We also compared the test set predictions of our multi-species models to those made by mean
397 conservation scores. First, we found that our model predictions for orthologs in other species of mouse
398 brain and liver OCRs whose OCR status is conserved tends to be higher than for orthologs in other species
399 of mouse brain and liver OCRs whose OCR status is not conserved (**Supplemental Tables 10-11**). Then,
400 for each tissue, we ranked the macaque OCR orthologs based on their model predictions, with the highest
401 rank corresponding to the highest score or open chromatin status prediction. For the OCR status-
402 conserved OCRs, we evaluated whether these OCRs tended to have higher ranks for our multi-species
403 model predictions than they did for mean conservation scores; we found that the ranks were significantly
404 higher for our predictions (**Tables 3-4, Figure 7a**). For each tissue, we also evaluated whether the OCR
405 orthologs without open chromatin tend to have lower ranks for our multi-species model predictions than
406 they do for mean conservation scores, and we found that the ranks were significantly lower for our
407 predictions (**Tables 3-4, Figure 7a**). We also did this for human and rat brain orthologs of mouse brain

408 OCRs with conserved and non-conserved open chromatin statuses and for rat liver orthologs of mouse
409 liver OCRs with conserved and non-conserved open chromatin statuses and obtained similar results
410 (**Supplemental Figure 9, Supplemental Tables 10-11, Supplemental Tables 15-17**).

411 Some of the OCR orthologs for which our models correctly predicted brain open chromatin
412 conservation in spite of low mean sequence conservation or for which our models correctly predicted lack
413 of brain open chromatin conservation in spite of high mean sequence conservation are near genes that
414 have been shown to play important roles in the brain. For example, there is a region on mouse
415 chromosome 2 – part of our test set – that has low mean sequence conservation according to PhastCons
416 [82] and PhyloP [81] but high brain experimentally identified and predicted open chromatin conservation
417 between mouse and macaque (**Figure 7b**) and whose mouse and macaque orthologs are located near the
418 gene *Stx16*. *Stx16* is involved in vesicle trafficking in most tissues, including the brain [98], and may play
419 a role in Alzheimer's disease [99]; in fact, its role in axon regeneration is conserved between mammals
420 and *C. elegans* [100]. Although this region near *Stx16* has generally low conservation, running TomTom
421 [101] on the 22bp sequence with high conservation revealed a subsequence that is similar to the *Fos*
422 motif, which is also found in the macaque ortholog. Since our machine learning model used sequence
423 similarity to the *Fos* motif in making predictions (**Supplemental Figure 8c**), the machine learning model
424 was likely able to automatically determine that it should use this sequence in making its prediction. In
425 addition, there is a region on mouse chromosome 2 that has high mean sequence conservation but low
426 experimentally identified and predicted open chromatin conservation between mouse and macaque
427 (**Figure 7c**) and whose mouse and macaque orthologs are located near the gene *Lnpk*. *Lnpk* in an
428 endoplasmic reticulum junction stabilizer [102] that has been shown to play a role in brain and limb
429 development [103], and mutations in *Lnpk* have been associated with neurodevelopmental disorders
430 [104]. It is possible that this region of the genome near *Lnpk* has a high degree of sequence conservation
431 because it has functions in other tissues. These results demonstrate the potential benefits of using OCR

432 ortholog open chromatin status predictions instead of mean conservation scores for studying the
433 evolution of the expression of important genes in a tissue of interest.

434 In addition, some of the accurate liver open chromatin conservation predictions disagreed with
435 the mean conservation scores. For instance, there is a region on mouse chromosome 2 that has high
436 experimentally identified and predicted liver open chromatin conservation but low sequence
437 conservation (**Figure 7d**) and whose mouse and macaque orthologs are located near *Rxra*. *Rxra* is a TF
438 involved in regulating in lipid metabolism [105-107], TF-MoDISco identified a sequence similar to its motif
439 as being important in our liver models (**Supplemental Figure 6d, Supplemental Figure 9b**), and its liver
440 expression is stable across fifteen mammals [108]. Although this region near *Rxra* has generally low
441 conservation, the 15bp segment with high conservation is similar to the motif for *Ctcf* according to
442 TomTom [101], and that motif is also found in the macaque ortholog. Since our machine learning model
443 used sequence similarity to the Ctcf motif in making predictions (**Supplemental Figure 8d**), the machine
444 learning model was likely able to automatically determine that it should use this sequence in making its
445 prediction. There is also a region on mouse chromosome 1, which is part of our test set, whose mouse
446 ortholog is an OCR and macaque ortholog is not an OCR according to our data and predictions in spite of
447 being highly conserved (**Figure 7e**) and whose mouse and macaque orthologs are near *Fn1*. *Fn1* has been
448 implicated in liver fibrosis [109-111], and a multi-species liver RNA-seq study found that it has higher
449 expression in mouse liver relative to livers of other mammals and birds and lower expression in primate
450 livers relative to livers of other mammals and birds [112]. For both of these OCRs, the H3K27ac signal
451 conservation in the same regions [80] is similar to the open chromatin status conservation, suggesting
452 that the open chromatin status conservation is indicative of enhancer activity conservation (**Figures 7d-**
453 **e**). These results suggest that using predicted OCR ortholog open chromatin status instead of
454 conservation has the potential to be beneficial for understanding gene expression evolution in multiple
455 tissues.

456

457 Open Chromatin Predictions Do Not Seem to Be Associated with Genome Quality

458 Since the Zoonomia genomes vary in quality, we evaluated whether our open chromatin status
459 predictions are associated with genome quality [3, 4, 39]. We computed the correlation between mean
460 predicted brain open chromatin status across the mouse brain OCR orthologs in each Glires species and
461 scaffold and contig N50's. We found a weak Pearson correlation and even weaker Spearman correlation
462 between the scaffold and contig N50's and the mean predicted brain open chromatin status
463 (**Supplemental Figure 10a**). We repeated this process in for the liver predictions of the mouse liver OCR
464 orthologs and obtained similar results (**Supplemental Figure 10b**). To demonstrate that mean predicted
465 mouse OCR ortholog open chromatin status has a stronger relationship with divergence from mouse than
466 it does with genome quality, we created generalized linear models for mean predicted mouse OCR
467 ortholog open chromatin status with covariates for divergence from mouse and scaffold or contig N50.
468 The coefficients for divergence from mouse were all statistically significantly different from zero and larger
469 in magnitude than the coefficients for scaffold or contig N50, and the coefficients for scaffold or contig
470 N50 were never statistically significantly different from zero (**Supplemental Table 18**). These results
471 suggest that lower-quality genomes are not strongly associated with less confident OCR ortholog open
472 chromatin status predictions.

473 To further evaluate the relationship between genome quality and our predictions, we investigated
474 whether the extent to which OCR ortholog open chromatin status predictions vary within a species is
475 associated with genome quality. To do this, we computed the correlation between standard deviation of
476 predicted brain open chromatin status across the mouse brain OCR orthologs in each Glires species and
477 scaffold and contig N50's. We found a weak negative Pearson correlation and even weaker negative
478 Spearman correlation between the scaffold N50's and the standard deviation of predicted brain open

479 chromatin status; for contig N50, the Pearson correlation was weak and negative, while the Spearman
480 correlation was weak and positive (**Supplemental Figure 10c**). We repeated this process for the liver open
481 chromatin status predictions of the mouse liver OCR orthologs and obtained similar results except that
482 the Spearman correlation between the contig N50 and standard deviation of predicted open chromatin
483 status was weak and negative (**Supplemental Figure 10d**). To demonstrate that standard deviation of
484 predicted mouse OCR ortholog open chromatin status had a stronger relationship with divergence from
485 mouse than it did with genome quality, we created generalized linear models for standard deviation of
486 predicted mouse OCR ortholog open chromatin status with covariates for divergence from mouse and
487 scaffold or contig N50. The coefficients for divergence from mouse were all statistically significantly
488 different from zero and larger in magnitude than the coefficients for scaffold or contig N50 (**Supplemental**
489 **Table 19**). These results further demonstrate that genome quality does not substantially influence our
490 OCR ortholog open chromatin status predictions.

491

492

493 **DISCUSSION**

494

495 The ability to link differences in genome sequence across species to the evolution of complex
496 traits is becoming tractable due to increased availability in genomic datasets and advances in
497 computational methodologies. Although gene regulatory mechanisms have been strongly linked with the
498 evolution of a number of complex traits [6, 8, 11], enhancers, which play a crucial role in gene regulation,
499 have been difficult to study due to their tissue-specificity [26] and relatively poor sequence conservation
500 across species [27, 28]. In recent years, computational models have been constructed that can predict
501 enhancer-associated regulatory genomics features with relatively high accuracy [30, 113-115]. However,

502 it has yet to be demonstrated that these models can achieve high lineage-specific accuracy. Indeed, when
503 we evaluated the performance of CNN models of open chromatin, their performance was often
504 substantially stronger on regions with conserved open chromatin relative to those whose open chromatin
505 status differed across species. We were able to improve this performance by shifting our negative sets in
506 model training from negative sets like dinucleotide-shuffled OCRs to negative sets consisting of closed
507 chromatin regions whose orthologs in other species are OCRs. It is possible that other negative sets, such
508 as genome-wide negatives or combinations of negative sets that work well, could further improve
509 performance, but such negative sets would require substantially more training time than our novel
510 negative set, making training and tuning such models potentially prohibitive for researchers who do not
511 have access to multiple GPUs.

512 Efforts of the Vertebrate Genomes Project, the Zoonomia Project [3], the Bat1K Project [2], the
513 Genome 10K Project [1], and others are making the genomes of hundreds of vertebrate species available
514 to the scientific community. In parallel, computational tools are being developed to improve assembly,
515 alignment, and identification of orthologous genes and regulatory elements [4, 40, 116-119]. Leveraging
516 these resources, we demonstrated the ability of our computational models to predict OCR ortholog open
517 chromatin status in dozens of species beyond those that we used for training, including correctly
518 predicting differences in open chromatin status between OCR orthologs. Furthermore, we found a
519 relationship between evolutionary distance and predicted open chromatin status differences in the brain
520 and the liver. Although there are OCR orthologs, such as species-specific OCRs and OCRs with
521 convergently evolved open chromatin [16], whose open chromatin status across species is not associated
522 with phylogenetic distance, we think that such OCRs in most tissues are a minority. This evaluation does
523 not require experimentally determined OCR data from multiple species; it requires only genomes from a
524 large number of species, which are becoming increasingly available [1-3].

525 In addition to showing that our novel models achieved lineage-specific OCR accuracy, achieved
526 tissue-specific OCR accuracy, and made predictions that had phylogeny-matching correlations, we also
527 showed that they were substantially more effective at predicting conservation or lack of conservation of
528 open chromatin than mean conservation scores generated by PhastCons [82] or PhyloP [81]. In some
529 cases, the OCR orthologs whose open chromatin status conservation or lack of conservation were
530 correctly predicted by our models but not by mean conservation scores were near important genes in the
531 tissue whose expression has the same cross-species trend as the open chromatin [108, 112],
532 demonstrating the potential machine learning model predictions have to identify candidate OCRs involved
533 in regulating important genes. We think our predictions work better than mean conservation scores
534 because conservation-based methods for predicting open chromatin status conservation consider every
535 nucleotide within an OCR to be equally important [16, 81, 82], whereas CNNs automatically learn which
536 nucleotides within a sequence are most strongly associated with open chromatin. Thus, if most of an OCR
537 is not conserved but the motif of a TF that is important for establishing open chromatin is conserved, then
538 the region may have conserved open chromatin in spite of having overall low conservation. Furthermore,
539 the high rate of turnover of specific transcription factor binding sites suggests that the regulatory function
540 of a region can be conserved in regions for which nucleotide-level conservation is difficult to detect [19,
541 120]. In addition, a region might be open in multiple tissues and be highly conserved because its open
542 chromatin status in some tissues is highly conserved even though its open chromatin status in a tissue of
543 interest is not conserved; thus, this further demonstrates the necessity of showing that models achieve
544 high tissue-specific accuracy.

545 Our novel brain and liver OCR ortholog open chromatin status prediction models have advantages
546 beyond meeting most of the criteria that we recommended. First, in order to directly help the model
547 achieve high lineage-specific accuracy, we used non-OCR orthologs of OCRs in the same tissue as
548 negatives. As open chromatin data from more species and tissues are generated, we anticipate that using

549 such a negative set will lead to increasingly good performance on our criteria. A further advantage is that
550 our models require as input sequences that are only 500bp long, as compared to other models that require
551 sequences of 3kb [30] or over 100kb [31]. Using short sequences as input enables us to make predictions
552 in species whose assemblies have short scaffolds, which include many of the species in the Zoonomia
553 Project [3, 4], and species whose genomes are substantially rearranged relative to the species with the
554 open chromatin data, such as gibbons [121]. In addition, it allows us to identify short regions with
555 sequence changes that are likely to have caused open chromatin changes, which is useful for downstream
556 work to identify mechanisms underlying evolutionary changes in open chromatin. Finally, using short
557 sequences allows us to experimentally validate predictions in a reporter assay [38, 122]. While our
558 negative set requires open chromatin data from at least two and ideally at least three species, we found
559 that using a large number of random, G/C-matched negatives [30] works well for training models on 500bp
560 sequences with open chromatin data from only one species.

561 Although our models excel in many criteria for their stated purpose of predicting lineage- and
562 tissue-specific open chromatin status, there are some cases where the models may not work well. Despite
563 the numerous advantages of predicting open chromatin status with short sequences, using shorter input
564 sequences also has limitations. Some enhancers, such as super-enhancers, are much longer than 500 base
565 pairs, and such enhancers have been shown to play important roles in the brain [123]. Another such case
566 is the example of long-range gene regulation [124], where open chromatin status can be affected by DNA
567 sequences that are more than a few hundred base pairs away from open chromatin peak summits. For
568 example, a recent study showed that many variants associated with open chromatin occur at least a few
569 hundred base pairs away from OCRs [125]. Encouragingly, our knowledge of 3D chromatin structure is
570 advancing rapidly, and, in the future, such data could be incorporated into open chromatin status
571 prediction models to account for the spatial and structural aspects of lineage-specific gene regulatory
572 differences. Furthermore, open chromatin status changes over evolutionary history can be affected by

573 factors not influenced by local sequence, such as changes in TFs' protein structures that affect their ability
574 to interact with other TFs [126], so any model with only DNA sequence underlying OCRs as input will not
575 be able to predict every OCR ortholog open chromatin status difference between species. An exciting
576 direction for future work would be to incorporate TF protein sequence and, if available, gene expression
577 into machine learning models for OCR ortholog open chromatin status prediction.

578 Our model set-up also has inherent limitations. For instance, since we predict open chromatin
579 status of only OCR orthologs, we cannot identify OCRs whose orthologs are not OCRs in any species for
580 which we have open chromatin data. An exciting extension to our work would be training a model on a
581 few species with open chromatin data to predict OCRs genome-wide in species without open chromatin
582 data; evaluating such models using the evaluation metrics we developed will be necessary for
583 demonstrating their utility in studying gene expression evolution. In addition, we treat open chromatin
584 status as binary, where, in reality, in bulk open chromatin data, open chromatin is a continuous signal.
585 Modifying our models to predict continuous open chromatin signal across species and our evaluation
586 metrics to evaluate such models' utility in understanding gene expression evolution would be an exciting
587 extension to this work. A previous study trained CNNs to predict continuous open chromatin signal across
588 species [31], suggesting that accomplishing this task might be feasible, but such models' ability to
589 accurately predict changes in open chromatin between species has yet to be systematically evaluated.

590 In addition, while our CNN has great performance, using a CNN for our machine learning model
591 has limitations. CNNs require inputs of a fixed size; this prevented us from accounting for differences in
592 peak length between OCRs and would make using CNNs in future work incorporating long-range
593 interactions difficult. CNNs also require extensive hyper-parameter tuning. It is possible that, with more
594 extensive hyper-parameter tuning, we would have been able to train models with better performance on
595 our criteria for some of the negative sets whose models had poor performance or to obtain models trained
596 on only mouse sequences with comparably good performance to the multi-species models. While

597 multiple Bayesian optimization methods exist for automating much of the hyper-parameter tuning
598 process [127-129], these methods often require extensive compute time that is not available to many
599 researchers. SVMs do not have CNNs' input size limits, have only a few hyper-parameters to tune, and
600 have been shown to work well on related tasks [30, 49, 115], but their prediction time can be slow because
601 their kernels need to be computed for every DNA sequence, which could make using SVMs for predicting
602 open chromatin status of orthologs of hundreds of thousands of OCRs in each of dozens of species
603 intractable. In addition, CNNs continue to be less directly interpretable than methods with user-defined
604 features that cannot account for complex combinatorial relationships between sequence patterns
605 involved in open chromatin, even though many advances have been made to improve the interpretability
606 of CNNs [52, 53, 130]. Interpreting models for open chromatin prediction could reveal the mechanisms
607 through which enhancer orthologs have lost activity over evolutionary history, such as losses in TF motifs
608 and changes in DNA shape.

609 Training and evaluating any machine learning model using regulatory genomics data from
610 multiple species is inherently limited because raising different species in the same type of controlled
611 environment is infeasible, and this can make differentiating between species-specific OCRs and
612 confounding factor-specific OCRs difficult. For example, the activity of many enhancers has been
613 associated with aging [131, 132]. Although all of our data were from adults, the mouse [43, 48] and rat
614 data were from younger adults, whereas the human and macaque data came from a combination of
615 younger and older adults [22, 41, 42]. Part of our motivation for conservatively defining clade-specific
616 OCRs was the desire to prevent Glires-specific OCRs from being young adult-specific OCRs. In addition,
617 time of day and the amount of time since waking up has been shown to affect enhancer activity [133],
618 and controlling for these factors is challenging when obtaining post-mortem human data or data from
619 different animal colonies. Although our macaque and rat samples were collected approximately two
620 hours after animals woke up, time of day of collection relative to sleep cycle for the remaining samples

621 used was either not described or not able to be controlled [22, 41-43, 48]. Thus, some individual OCR
622 ortholog open chromatin status differences between species and tissues could be affected by the time
623 that the animal had been was awake, in addition to species and tissue differences. Furthermore, an
624 animal's sex has been shown to be associated with the activity of both brain [134, 135] and liver [136]
625 enhancers. Although all of our datasets with multiple biological replicates had both males and females,
626 the number of male and female replicates differed between datasets. We hope that our conservative
627 definitions of clade-specific, species-specific, and tissue-specific OCRs prevented these OCRs from being
628 sex-specific OCRs. In addition, our conservative definitions of clade-specific, species-specific, and tissue-
629 specific OCRs limited our power to compare different models on these OCRs.

630 The metrics that we presented for evaluating OCR ortholog open chromatin status prediction
631 methods and our approach to constructing a negative set for such methods can be applied to any tissue
632 or cell type with data from open chromatin assays in multiple species. Beyond the data employed in this
633 study, there are publicly available open chromatin data from both human and mouse from many tissues
634 [23, 43, 47, 137, 138] and from dog [139], cow, and pig [140] from multiple tissues. An exciting extension
635 would be to train new machine learning models to predict the open chromatin status of OCR orthologs in
636 additional species and tissues, making use of these datasets as well as the wealth of similar OCR datasets
637 soon to be released. As single-cell ATAC-seq (scATAC-seq) [141] becomes more widely used, these
638 methods could be applied to predicting OCR ortholog open chromatin status in specific cell types and
639 determining if specific cell types have been lost in some species.

640 The results from our method can also be used in a forward genomics approach [142] to help
641 identify the mechanisms underlying the evolution of the expression of genes of interest or of phenotypes
642 that have evolved through gene expression. This can be done by identifying OCR orthologs whose changes
643 in predicted open chromatin status correspond to changes in gene expression or phenotypes. Many multi-
644 species gene expression datasets in tissues with data from an assay that can serve as a proxy for enhancers

645 are publicly available [47, 80, 108, 143], and more will likely be generated in the near future [140].
646 Additional multi-species single-cell RNA-seq datasets are being generated from some of these tissues
647 [144, 145] and will likely soon be supplemented by scATAC-seq. Many of these tissues have been
648 associated with phenotypes that have evolved through gene expression [8, 11], so we can use the gene
649 expression data along with predictions from models like ours to gain insights how these phenotypes
650 evolved. Such insights may also reveal mechanisms underlying diseases associated with these phenotypes
651 [29].

652

653

654 **CONCLUSIONS**

655 The specific enhancer sequence differences that underlie phenotypes that have evolved through gene
656 expression are largely unknown and are difficult to determine. One can identify candidate differences by
657 using open chromatin data from one species to predict open chromatin status of OCR orthologs in another
658 and finding those OCR orthologs whose predicted activity is associated with a phenotype. We developed
659 a set of criteria for evaluating OCR ortholog open chromatin status prediction methods and used them to
660 evaluate multiple methods for this task, including a new method developed explicitly for identifying
661 sequence differences associated with changes in open chromatin between species. We found that some
662 methods, including ours, satisfy most of our criteria. While we applied our method and evaluation criteria
663 to mammalian brain and liver, they can be applied to any tissue or cell type with enhancer data available
664 from multiple species. As this is, to our knowledge, the first study to predict the open chromatin status
665 of OCR orthologs across more than a few species, we anticipate that our work will serve as a foundation
666 for identifying OCRs whose open chromatin status is involved in gene expression evolution, providing

667 insights into transcriptional regulatory mechanisms underlying phenotypes that have evolved through
668 gene expression and their associated diseases.

669

670

671 **METHODS**

672

673 Assaying Open Chromatin in Mouse Liver

674 We performed ATAC-seq experiments on two 10-week-old heterozygous Pvalb-2A-Cre mice
675 (B6.Cg-Pvalb$^{tm1.1(cre)Aibs}$/J; Jackson Stock No: 012358) [146], one male and one female. We euthanized the
676 mice by isoflurane and decapitation. We quickly dissected fresh liver tissue and extracted nuclei by 30
677 strokes of Dounce homogenization with the loose pestle (0.005 in. clearance) in 5mL of cold lysis buffer
678 [45]. The nuclei suspensions were filtered through a 70µm cell strainer, pelleted by centrifugation at
679 2,000 x g for 10 minutes, resuspended in water, and filtered a final time through a 40µm cell strainer.
680 Sample aliquots were stained with DAPI (Invitrogen #D1206), and nuclei concentrations were quantified
681 using a manual hemocytometer under a fluorescent microscope. Approximately 50,000 nuclei were
682 input into a 50µL ATAC-seq tagmentation reaction as described in [45] and [44]. The resulting libraries
683 were amplified to 1/3 qPCR saturation, and fragment length distributions as estimated by the Agilent
684 TapeStation System showed high quality ATAC-seq periodicity. The samples were paired-end sequenced
685 on the Illumina NovaSeq 6000 System through Novogene services. We obtained 165,337,124 reads
686 from the male mouse and 225,752,264 reads from the female mouse.

687

688

689 Identifying Brain and Liver OCRs

690 We used open chromatin data from four species: *Homo sapiens* [22, 42, 47], *Macaca Mulatta* [41],
691 *Mus musculus* [43, 48], and *Rattus norvegicus* [41]. For human brain OCRs, we used NeuN+ primary motor
692 cortex (4 biological replicates), putamen (4 biological replicates), and nucleus accumbens (1 biological
693 replicate) ATAC-seq data from [42] and caudate and putamen DNase hypersensitivity data (2 biological
694 replicates) from [22]. For macaque brain OCRs, we used orofacial motor cortex (2 biological replicates),
695 hand motor cortex (2 biological replicates), caudate (2 biological replicates), putamen (2 biological
696 replicates), and nucleus accumbens (1 biological replicate) ATAC-seq data from [41]. For macaque liver
697 OCRs, we used liver ATAC-seq data (1 biological replicate) from [41]. For mouse brain OCRs, we used
698 cortex and striatum ATAC-seq data from seven-week-old and twelve-week-old mice from [48] (2 biological
699 replicates each). For mouse liver OCRs, we used the mouse liver ATAC-seq data that we generated as well
700 as mouse liver ATAC-seq data from [43] (4 biological replicates). For rat brain OCRs, we used primary
701 motor cortex (3 biological replicates) and striatum data (2 biological replicates) from [41]. For rat liver
702 OCRs, we used liver ATAC-seq data (2 biological replicates) from [41]. We combined reads from technical
703 replicates for all of the datasets.

704 We processed DNase hypersensitivity data by using the Kundaje Lab open chromatin pipeline
705 [147] to map reads to hg38 [148], filter reads, call peaks, evaluate which peaks are reproducible, and
706 remove peaks overlapping the ENCODE black list [149]. We used the default settings for the pipeline.
707 Human brain DNase hypersensitivity data from the caudate nucleus and the putamen were downloaded
708 from the ENCODE portal [47]. Since the caudate nucleus and the putamen are both parts of the striatum
709 but came from different people, we treated them as biological replicates. The final set of peaks was the
710 larger set of the peaks that were reproducible according to IDR across biological replicates and the peaks
711 that were reproducible according to IDR across pooled pseudo-replicates (the “optimal set”).

712 We processed the mouse brain ATAC-seq data using the Kundaje Lab open chromatin pipeline
713 [147] and the mouse liver, human, macaque, and rat ATAC-seq data using the ENCODE ATAC-seq pipeline
714 [150]. For the mouse brain ATAC-seq data, we began with the filtered bam files from [48] used the default
715 parameters for the remainder of the pipeline. For the other ATAC-seq data, we used the default
716 parameters except for "atac.multimapping" : 0, "atac.cap_num_peak" : 300000, "atac.smooth_win" : 150,
717 "atac.enable_idr" : true, and "atac.idr_thresh" : 0.1; these parameter modifications enabled the
718 parameters for read filtering, peak calling, and running IDR to be the same as those used for the mouse
719 brain data. We mapped the human data to hg38 [148], the macaque data to rheMac8 [151], the mouse
720 data to mm10 [152], and the rat data to rn6 [153]. For the mouse liver ATAC-seq data from [43], we
721 treated the two female and two male samples as four biological replicates. The final set of peaks for
722 datasets with multiple biological replicates was the larger set of the peaks that were reproducible
723 according to the irreproducible discovery rate (IDR) [154] across biological replicates and the peaks that
724 were reproducible according to IDR across pooled pseudo-replicates (the “optimal set”); the final set of
725 peaks for datasets with only 1 biological replicates were the peaks that were reproducible according to
726 IDR across self-pseudo-replicates.

727 We then used the percentage of mapped reads, number of filtered reads, periodicity, TSS
728 enrichment, number of IDR reproducible peaks, rescue ratio, and self-consistency ratio analyses
729 generated by the pipelines [45, 155] to evaluate data quality. We found that most of the samples were
730 high-quality. However, we excluded the second macaque nucleus accumbens biological replicate because
731 it had only about sixteen million filtered reads and poor periodicity and because the two replicates had
732 rescue ratio 4.01 and self-consistency ratio 2.04. We also excluded the second macaque liver replicate
733 because it had only about two million filtered reads and poor periodicity and because the two replicates
734 had self-consistency ratio 3.53. In addition, we excluded the first rat liver biological replicate it had only
735 35,593 reproducible peaks according to IDR across self-pseudo-replicates in spite of having over sixty-

736 eight million filtered reads. As a result, for macaque nucleus accumbens and liver, we used the peaks
737 from the first biological replicate that were reproducible according to IDR across self-pseudo-replicates,
738 and, for rat liver, we used the "optimal set" from running the ENCODE ATAC-seq pipeline on only biological
739 replicates 2 and 3.

740 To obtain OCRs in each species, we intersected the IDR "optimal set" reproducible peaks from
741 each of the brain regions and datasets for brain and each of the liver datasets for liver and defined OCRs
742 to be the intersected peaks that are likely to be enhancers. Specifically, for each species, we selected one
743 set of reproducible open chromatin peaks to be the "base peaks," used bedtools intersect with the -wa
744 and -u options to intersect it with each of the other reproducible peak sets in series, and then used
745 bedtools closestBed with the -t first and -d options to identify the "base peaks" that overlapped at least
746 one peak from each other set that were over 20kb from the nearest protein-coding TSS (not promoters),
747 at most 1kb long (not super-enhancers), and non-exconic [156]. The base peaks for human brain were the
748 IDR "optimal set" from NeuN+ cells in the primary motor cortex from [42], for the macaque brain were
749 the IDR "optimal set" from the orofacial motor cortex from [41], for the macaque liver were the IDR
750 reproducible peaks across self-pseudo replicates from the first macaque liver replicate from [41], for
751 mouse brain were the IDR "optimal set" from the cortex from the seven-week-old mouse from [48], for
752 mouse liver were the IDR "optimal set" from our mouse liver ATAC-seq dataset, for the rat brain were the
753 IDR "optimal set" from the primary motor cortex from [41], and for the rat liver were the IDR "optimal
754 set" from the second and third rat liver replicates from [41]. To determine the distance from the nearest
755 protein-coding TSS, we used the GENCODE protein-coding TSSs for human (version 27) and mouse
756 (version M15) [157, 158], the union of the RefSeq rheMac8 protein-coding TSSs [159] and the human
757 GENCODE protein-coding TSSs mapped to rheMac8 using liftOver [160] for macaque, and the union of the
758 RefSeq rn6 protein-coding TSSs [159] and the mouse GENCODE protein-coding TSSs mapped to rn6 using
759 liftOver [160] for rat. To identify non-exonic peaks, we used bedtools [156] subtract with option -A to

760 identify peaks that did not overlap protein-coding exons, where human protein-coding exons were
761 obtained from GENCODE (version 27), mouse protein-coding exons were obtained from GENCODE
762 (version M15) [157, 158], macaque protein-coding exons were obtained from RefSeq for rheMac8, and
763 rat protein-coding exons were obtained from RefSeq for rn6 [159]. We defined the peak summit of an
764 OCR to be the peak summit of the corresponding base peak, and we constructed positive examples by
765 taking likely enhancer peaks summits +/- 250bp and their reverse complements. We centered peaks on
766 their summits because previous work has shown that there is a concentration of TF motifs at peak summits
767 [116, 161, 162]. By requiring OCRs to be reproducible open chromatin peaks according to IDR, intersecting
768 OCRs across multiple datasets, and filtering OCRs in a conservative way, we limited the number of false
769 positive OCRs being used to train our machine learning models.

770

771 Constructing Negative Sets

772 *Flanking Regions*

773 We constructed the flanking region negative set by using bedtools [156, 163] to identify the subset
774 of regions flanking mouse brain OCRs that are not OCRs. Specifically, we first identified the 500bp flanking
775 regions of each of our OCRs +/- 500bp; we required a 500bp separation between flanks and OCRs to
776 ensure that we would not have false positives in our negative set due to poorly defined peak boundaries.
777 We then removed all flanking regions that overlapped any open chromatin peaks called from the pooled
778 set of reads across biological replicates from cortex or striatum from either age [147, 150]; we used these
779 peaks instead of the subset of such peaks that we defined as OCRs because non-reproducible peaks have
780 the potential to be enhancers, and we wanted to limit the number of false negatives in our negative set.
781 For each remaining flanking region, we used its underlying sequence and that sequence's reverse

782 complement. Thus, although there could be up to two negatives for every positive, our negative:positive
783 training data ratio was approximately 1.65:1 (**Table 5**).

784 *OCRs from Other Tissues*

785 We constructed the OCRs from other tissues negative set by identifying OCRs in non-cortex and
786 non-striatum tissues that do not overlap our brain OCRs. We first used the ENCODE ATAC-seq pipeline
787 [150] with the same parameters that we used for the brain samples to process all of the ATAC-seq data
788 from tissues that do not overlap cortex and striatum from the mouse ENCODE post-natal samples [138]
789 and from [43]. We then used the same quality control metrics that we used for selecting datasets to
790 include as OCRs to evaluate the quality of these datasets and removed those that were low-quality. The
791 mouse ENCODE datasets that we included were from liver (We did not use this for our liver positive set
792 because it came from an embryonic sample, and the other datasets were from adults.), intestine, and
793 cerebellum. The datasets from [43] that we included were from female abdominal fat, female adrenal
794 gland, female kidney, male kidney, female liver, male liver (Male and female samples were processed
795 separately for the purposes of creating this negative set.), female lung, male lung, female pancreas, male
796 small intestine, male spleen, male stomach, female thymus, and male thymus. We obtained the union of
797 the IDR "optimal set" peaks across all of these datasets as well as our mouse liver data and used bedtools
798 subtract with the -A option [156] to remove those peaks that overlapped open any chromatin peaks called
799 from the pooled set of replicates from cortex or striatum from either age. For each filtered peak, we used
800 the sequence underlying its summit +/- 250bp and that sequence's reverse complement. Our
801 negative:positive training data ratio was approximately 19.78:1 (**Table 5**).

802 *G/C- and Repeat-Matched Regions*

803 We identified G/C- and repeat-matched regions for our OCRs using a combination of R packages
804 and bedtools [156]. We first created a repeat-masked mm10 genome by running

805 forgeMaskedBSgenomeDataPkg from the BSgenome R package [164] on mm10 [152] with masks
806 downloaded from the UCSC Genome Browser [165]. We then ran genNullSeqs from the gkmSVM R
807 package [115, 166] on the sequences of the brain OCR peak summits +/- 250bp and our masked mouse
808 genome with default parameters except for the following: length_match_tol=0.00, which ensures that all
809 of our sequences are 500bp; nMaxTrials=100, which allows for more attempts to find G/C- and repeat-
810 matched regions than the default; and xfold=10 for the larger G/C- and repeat-matched region negative
811 set and =2 for the smaller G/C- and repeat-matched region negative set. Although we allowed for more
812 trials, getNullSeqs found fewer G/C- and repeat-matched regions than we had requested. After
813 generating these regions, we used bedtools subtract [156] with the -A option to remove any regions that
814 overlapped any open chromatin peaks called from the pooled set of replicates from cortex or striatum
815 from either age. For each filtered G/C- and repeat-matched region, we used its underlying sequence and
816 that sequence's reverse complement. As a result, for the larger G/C- and repeat-matched negative set,
817 the negative:positive training data ratio was approximately 8.15:1, and, for the smaller G/C- and repeat-
818 matched negative set, the negative:positive training data ratio was approximately 1.64:1 (**Table 5**).

819 *Dinucleotide-Shuffled OCRs*

820 We obtained dinucleotide shuffled OCRs by running the MEME suite's [167] fasta-shuffle-letters
821 on the sequences of our brain OCR peak summits +/- 250bp. We used the default parameters except for
822 -kmer 2, which enabled us to preserve dinucleotide frequencies, and -copies 10, which enabled us to
823 generate a negative set that was ten times larger than our positive set. We used every shuffled sequence
824 and its reverse complement. Thus, the negative:positive training data ratio was exactly 10:1 (**Table 5**).

825 *Non-OCR Orthologs of OCRs*

826 We obtained the non-OCR orthologs of OCRs (red-brown regions in **Figure 1**) by obtaining the
827 orthologs of OCRs in each species in each other species for which we had data from the same tissue and

828 filtering the orthologs to include only those that did not overlap open chromatin in the same tissue. For
829 example, mouse chr4:127435564-127436049 does not overlap any OCRs in mouse cortex or striatum, but
830 its human ortholog, chr1:34684126-34684689, is an OCR in both cortex and striatum, so this mouse region
831 was a member of our novel negative set. To ensure that we would have enough negatives for training the
832 model, we created a less conservative set of OCRs, which we called "loose OCRs." For each species, tissue
833 combination, the loose OCRs are the base peaks that are non-exonic, at least 20kb from a TSS, and at most
834 1kb long (same criteria used in constructing the positive set) and intersect at least 1 peak from the pooled
835 reads across replicates from each of other datasets that are used for the species, tissue combination; we
836 obtained these loose OCRs using bedtools [156]. We defined the peak summit of a loose OCR to be the
837 peak summit of the corresponding base peak.

838 We identified orthologs of our loose OCRs in each other species with open chromatin data from
839 the same tissue using halLiftover [117] followed by HALPER [116]. We used these tools instead of liftOver
840 [160] because they map regions using Cactus alignments [40], which, unlike the pairwise alignment
841 liftOver chains, contain many species and account for a wide range of structural rearrangements, including
842 inversions. We ran halLiftover with default parameters on our loose OCRs and their peak summits using
843 the Zoonomia Cactus alignment [4]. We then constructed contiguous loose OCR orthologs from the
844 outputs of halLiftover by running HALPER with parameters -max_frac 2.0, -min_len 50, and -protect_dist
845 5. Finally, we used bedtools subract with the -A option [156] to remove the loose OCR orthologs in each
846 species that overlapped peaks from the pooled reads across replicates from any of the datasets from the
847 same tissue in that species.

848 For the models trained only on mouse sequences, we used only the mouse orthologs of loose
849 OCRs in non-mouse species. For the multi-species brain models, we used the mouse, human, macaque,
850 and rat orthologs of loose brain OCRs from each of the other species, and for the multi-species liver
851 models, we used the mouse, macaque, and rat orthologs of loose liver OCRs from each of the other

852 species. When constructing negatives of the non-OCR orthologs of OCRs, we used sequence underlying
853 the ortholog of the base peak summit +/- 250bp and that sequence's reverse complement. Our
854 negative:positive training set ratios were approximately 1.16:1 for the brain model trained on only mouse
855 sequences, 0.704:1 for the liver model trained on only mouse sequences, 1.49:1 for the multi-species
856 brain model, and 0.822:1 for the multi-species liver model.

857

858 Constructing Training, Validation, and Test Sets

859 For the models trained using only mouse sequences, we divided the positives and negatives
860 (except for the dinucleotide-shuffled OCRs) into training, validation, and test sets based on chromosomes
861 to ensure that there would be no overlap between the sets. For the dinucleotide-shuffled OCRs negatives,
862 we put them into the set that corresponded to the positive example from which they were constructed.
863 Our training set consisted of regions on mm10 chromosomes 3-7, 10-19, and X. Our validation set that
864 we used for developing our positive and negative set definitions (for example, validation set performance
865 was used to determine that we should use orthologs of loose OCRs instead of OCRs for our novel negative
866 set), early-stopping, and hyper-parameter tuning consisted of regions on mm10 chromosomes 8-9. Our
867 test set consisted of mm10 chromosomes 1-2. All presented evaluations on mouse genomic regions,
868 including those on types of regions not used in model training, were done on only regions from mm10
869 chromosomes 1-2.

870 For the models and model evaluations using sequences from non-mouse species, we divided
871 sequences into training, validation, and test sets based on the chromosomes to which their mouse
872 orthologs mapped. In other words, we mapped such regions to mm10 using halLiftover [117] with the
873 Zoonomia Cactus [4] followed by HALPER [116] with parameters -max_frac 2.0, -min_len 50, and -
874 protect_dist 5 and put them into the training set if their mm10 orthologs were on chromosomes 3-7, 10-

875 19, or X; put them into the validation set if their mm10 orthologs were on chromosomes 8-9; put them
876 into the test set if their mm10 orthologs were on chromosomes 1-2; and excluded them if their orthologs
877 were elsewhere in mm10 or if they had no orthologs. Although many non-mouse regions were excluded
878 from evaluation, because some OCRs have high sequence conservation, not accounting for the location
879 of mouse orthologs when constructing training, validation, and test sets could lead to test set sequences
880 that are almost identical to training set sequences [47, 168].

881

882 Training Machine Learning Models

883 We used CNNs [51] for our machine learning model because they have achieved state-of-the-art
884 performance in related tasks [36, 114, 169]; they can learn complex combinatorial relationships between
885 sequences, which we know can play an important role in enhancer activity [32]; and they do not require
886 an explicit featurization of the data, enabling them to learn yet-to-be-discovered sequence patterns that
887 are important for enhancer activity. Our inputs were one-hot-encoded DNA sequences [50], and our
888 outputs were the probability of the sequence being an OCR in the tissue for which the model was trained.
889 We first tuned hyper-parameters for the brain model trained on only mouse sequences with our novel
890 negative set by comparing the validation set performances of various models. Our final architecture was
891 five convolutional layers with 300 filters of width 7 and stride 1 in each, followed by a max-pooling layer
892 with width and stride 26, followed by a fully connected layer with 300 units, followed by a fully-connected
893 layer that went into a sigmoid. All convolutional layers had dropout 0.2 and L2 regularization 0.00001.
894 The model was trained using stochastic gradient decent with learning rate 0.001, Nesterov momentum
895 0.99, and batch size 100, and each class was assigned a weight equal to the fraction of the other class in
896 the training set. The model was trained using the training set until there were three consecutive epochs
897 with no improvement in recall at eighty percent precision (or, if there were more positives than negatives,

898 no improvement in specificity at eighty percent NPV) on the validation set. Before training, weights were
899 initialized to be those from a pre-trained neural network with the same hyper-parameters and the
900 negative set randomly down-sampled to be the size of the positive set (or a positive set randomly down-
901 sampled to be the size of the negative set if the positive set was larger). The weights for the pre-training
902 were initialized using Keras's He normal initializer [170, 171].

903 We then tuned hyper-parameters for the CNNs for the other negative sets, the CNNs for the liver
904 data, and the CNNs for the multi-species models. We began with the architecture described above and
905 adjusted the number of convolutional filters per layer and the learning rate, ultimately selecting the values
906 that provided the best performance on the validation set. For the model with flanking regions negatives,
907 we used 250 convolutional filters per layer and learning rate 0.001. For the model with the OCRs from
908 other tissues negatives, we used 250 convolutional filters per layer and learning rate 0.0005. For the
909 model with the larger number of random G/C- and repeat-matched negatives, the multi-species brain
910 model, and both liver models, we used 350 convolutional filters per layer and learning rate 0.001. For the
911 models with the smaller number of random G/C- and repeat-matched negatives and the dinucleotide-
912 shuffled negatives, we used 300 convolutional filters per layer and learning rate 0.001. All models were
913 implemented and trained using Keras [171] version 1.2.2 with the Theano backend [172] and evaluated
914 using Scikit-learn [173] and PRROC [174, 175].

915

916 Calibrating Machine Learning Models

917 Because the machine learning models trained with some of the negative sets had high sensitivity
918 and low specificity, we tried re-calibrating them with the training data from our novel negative set. More
919 specifically, we first made predictions with the model we wanted to re-calibrate on the training data from
920 the positive set and our novel negative set. We next trained a logistic regression to use the model's

921 predictions as features to predict the real open chromatin status for these training examples. We then
922 used the logistic regression to make predictions on the test set. The training and prediction was done
923 using Scikit-learn [173].

924

925 Evaluating Machine Learning Models

926 *Evaluating Models' Lineage-Specific OCR Accuracy*

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

928 To obtain OCR orthologs that are open in one species but not in another, we used as positives the
929 sequences and reverse complements of sequences underlying OCRs from one species whose orthologs in
930 the other species have closed chromatin and as negatives sequences and reverse complements of
931 sequences underlying non-OCRs whose orthologs in at least one other species are OCRs. More specifically,
932 for evaluating mouse OCRs whose open chromatin status differs in other species, we used halLiftover
933 [117] followed by HALPER [116] with the same parameters we used previously to identify mouse OCR
934 orthologs in human, macaque, and rat for brain and macaque and rat for liver that did not overlap any of
935 the peaks from the pooled replicates from any of the datasets that we used for identifying OCRs. This
936 gave us 1,570 positive test set examples for evaluating brain models and 3,738 positive test set examples
937 for evaluating liver models. For evaluating mouse non-OCRs whose open chromatin status differs in other
938 species, we used the same approach that we took to construct our novel negative set except that we
939 identified non-OCR orthologs of OCRs instead of loose OCRs. This gave us 2,062 negative test set examples
940 for evaluating brain models and 4,080 negative test set examples for evaluating liver models (**Figure 2c,**
941 **Supplemental Figure 6a**). For evaluating mouse OCRs whose rat orthologs are not OCRs, we identified
942 mouse OCRs whose rat orthologs do not overlap pooled peaks across replicates for any rat OCRs from any
943 of the datasets used to define OCRs; this gave us 674 positive test set examples for evaluating brain

944 models and 2,482 positive test set examples for evaluating liver models. For evaluating mouse closed
945 chromatin regions whose orthologs are open in rat, we used the subset of our novel negative set that
946 came from rat OCR orthologs (not rat loose OCR orthologs); this gave us 890 negative test set examples
947 for evaluating brain models and 2,050 negative test set examples for evaluating liver models
948 (**Supplemental Figure 1a, Supplemental Figure 6a**). For evaluating macaque OCRs whose open chromatin
949 status differs in mouse, we identified macaque OCRs whose mouse orthologs do not overlap peaks from
950 the pooled replicates of any of the datasets we used for identifying mouse OCRs; this gave us 734 positive
951 test set examples for evaluating brain models and 2,384 positive test set examples for evaluating liver
952 models. For identifying macaque non-OCRs whose OCR status differs in mouse, we identified mouse OCR
953 orthologs in macaque that do not overlap peaks from the pooled replicates from any of the datasets we
954 used for identifying macaque OCRs; this gave us 788 negative test set examples for evaluating brain
955 models and 2,228 negative test set examples for evaluating liver models (**Figure 2d**, **Supplemental Figure**
956 **6a**). We obtained human and rat OCRs and non-OCRs whose OCR status is different in mouse using same
957 process that we used for macaque. We obtained 416 positive and 896 negative test set examples for
958 evaluating brain models (**Supplemental Figure 1b**). For evaluating rat OCRs and non-OCRs whose mouse
959 orthologs have different OCR statuses, we obtained 990 positive and 676 negative test set examples for
960 evaluating brain models and 2,050 positive and 2,482 negative test set examples for evaluating liver
961 models (**Supplemental Figure 1c, Supplemental Figure 6a**).

962 Clade- and Species-Specific OCRs:

963 We defined a clade-specific OCR in a tissue as an OCR whose ortholog is open in that tissue in
964 every species within a clade for which we have data and closed in every other species for which we have
965 data, and we defined a species-specific OCR as an OCR whose ortholog is open a species for which we
966 have data and is closed in the most closely related species for which we have data. More specifically, we
967 identified clade-active OCRs in each clade – OCRs in a "base species" whose ortholog in the other species

968 in that clade (if there was another) overlaps an open chromatin peak from the pooled reads across
969 replicates in all of the datasets we used in that tissue from that species. We did not require the OCR
970 ortholog in the non-base species to overlap a reproducible open chromatin peak so that we could have at
971 least one hundred test set examples for each evaluation. We chose the "base species" to be the species
972 in each clade with the highest-quality genomes – mouse for Glires for brain and liver, human for
973 Euarchonta for brain, and macaque for Euarchonta for liver. We then identified the subset of clade-active
974 peaks from the base species whose orthologs in all species in the other clade do not overlap any open
975 chromatin peaks from the pooled replicates from any of the datasets we used to identify OCRs; these
976 were our clade-specific OCRs. To obtain clade-specific non-OCRs for a clade, we identified orthologs of
977 clade-active OCRs from the other clade in the base species in the clade whose orthologs in all species in
978 the clade did not overlap open chromatin peaks from pooled replicates in any of the datasets we used for
979 identifying OCRs. The sequences of clade-specific OCRs and non-OCRs used for evaluating the models
980 were those from the base species and their reverse complements. This gave us 230 Glires-specific test
981 set brain OCRs, 134 Glires-specific test set brain non-OCRs (**Supplemental Figure 1d, Supplemental Figure**
982 **2a, Figure 6a**), 134 Euarchonta-specific test set brain OCRs, 230 Euarchonta-specific test set brain non-
983 OCRs (**Figure 2e, Supplemental Figure 2b, Figure 6a**), 1,024 Glires-specific test set liver OCRs, 1,826 Glires-
984 specific test set liver non-OCRs (**Supplemental Figure 6a, Figure 6b**), 1,826 Euarchonta-specific test set
985 liver OCRs, and 1,024 Euarchonta-specific test set liver non-OCRs (**Supplemental Figure 6a, Figure 6b**).
986 When evaluating the multi-species models, we combined the clade-specific OCRs and non-OCRs from the
987 two clades.

988 When evaluating species-specific OCRs and non-OCRs, we identified mouse OCRs and non-OCRs
989 whose orthologs' open chromatin status differs in rat, rat OCRs and non-OCRs whose orthologs' open
990 chromatin status differs in mouse, human OCRs and non-OCRs whose orthologs' open chromatin status
991 differs in macaque, and macaque OCRs and non-OCRs whose orthologs' open chromatin status differs in

992 human. We obtained 66 human-specific brain OCRs, 188 human-specific brain non-OCRs, 188 macaque-
993 specific brain OCRs, and 66 macaque-specific brain non-OCRs (**Figure 6a**). We did not include macaque-
994 specific OCRs and non-OCRs when evaluating the multi-species liver model because we did not have liver
995 open chromatin from any other Euarchonta species. We combined the species-specific OCRs and non-
996 OCRs from different species when evaluating the multi-species models (**Figures 6a-b**).

997 *Evaluating Models' Tissue-Specific OCR Accuracy*

998 To evaluate the performance of models trained in one tissue on OCRs from another tissue, we
999 defined our positives to be OCRs that are shared between the two tissues (shows our models were not
1000 learning only the sequences involved in tissues-specific open chromatin), and we defined our negatives
1001 to be OCRs in the evaluation tissue that do not overlap OCRs in the training tissue (shows our models were
1002 not learning only sequences involved in general open chromatin). More specifically, we used bedtools
1003 intersectBed with options -wa and -u [156] to identify OCRs from our training tissue that overlap OCRs
1004 from the evaluation tissue. For brain models, we obtained 1,040 positives for mouse, 846 positives for
1005 macaque, and 1,770 positives for rat (**Figure 3, Supplemental Figure 3**). For liver models, we obtained
1006 2,012 positives for mouse, 946 positives for macaque, and 1,130 positives for rat (**Supplemental Figure**
1007 **6b**). We used bedtools subtractBed with option -A [156] to identify liver OCRs that do not overlap any
1008 open chromatin from pooled replicates in any of the datasets that were used to generate the brain OCRs
1009 and brain OCRs that do not overlap open chromatin from pooled replicates in any of the datasets that
1010 were used to generate liver OCRs. For brain, we obtained 3,382 negatives for mouse, 1,898 negatives for
1011 macaque, and 3,518 negatives for rat (**Figure 3, Supplemental Figure 3**). For liver, we obtained 2,212
1012 negatives for mouse, 1,428 negatives for macaque, and 2,942 negatives for rat (**Supplemental Figure 6b**).
1013 We combined the data from all three species for evaluating the multi-species models (**Figures 6a-b**).

1014 For the negative set comparison, we also compared the distributions of test set predictions for
1015 the brain OCRs that do not overlap liver OCRs, the brain OCRs that overlap liver OCRs, the liver OCRs that
1016 do not overlap brain OCRs, and the negative set. We defined these groups of OCRs as we did for other
1017 evaluations, and we used predictions for sequences and their reverse complements. We compared the
1018 distributions for the brain OCRs that overlap liver OCRs to the liver OCRs that do not overlap brain OCRs
1019 using a Wilcoxon rank-sum test and multiplied the p-values by 6 to do a Bonferroni correction.

1020 *Evaluating if Models' Predictions Had Phylogeny-Matching Correlations*

1021 To evaluate the relationship between OCR ortholog open chromatin status and phylogenetic
1022 distance, we identified test set mouse OCR orthologs in all of the fifty-six Glires species from Zoonomia,
1023 predicted the open chromatin statuses of those orthologs, and computed the correlation between those
1024 predictions and the species' phylogenetic divergences from mouse. This provides us with an approximate
1025 measure of how predicted OCR ortholog open chromatin statuses change over evolution. We identified
1026 the test set mouse OCR orthologs and OCR summit orthologs in Glires using halLiftover [117] with the
1027 Cactus alignment [40] from Zoonomia [4]; we used brain OCRs for evaluating brain OCR models and liver
1028 OCRs for evaluating liver OCR models. We next constructed contiguous orthologs from the outputs of
1029 halLiftover using HALPER [116] with parameters -max_frac 2.0, -min_len 50, and -protect_dist 5. We
1030 constructed inputs for our models from the contiguous OCR orthologs by using bedtools fastaFromBed
1031 [156] with fasta files downloaded from NCBI [3, 39] and the UCSC Genome Browser [165] to obtain the
1032 sequences underlying their summit orthologs +/- 250bp. We constructed the reverse complements of
1033 sequences, used our models to predict each sequence and its reverse complement's open chromatin
1034 status, and averaged the predictions between the forward and reverse strands. We then removed all
1035 predictions from OCRs with orthologs in less than one quarter of species. After that, for each model, we
1036 computed the mean OCR ortholog open chromatin status prediction and the standard deviation of
1037 predictions across all remaining OCR orthologs in each species. We finally computed the Pearson and

1038 Spearman correlations between these means and standard deviations of predictions and the millions of
1039 years since divergence from mouse, which we obtained from TimeTree [176]. We did this for brain OCR
1040 orthologs using brain models trained on mouse sequences from each negative set and the multi-species
1041 brain model as well as for liver OCR orthologs using the liver model trained on only mouse sequences and
1042 the multi-species liver model.

1043 *Evaluating the Relationship between OCR Ortholog Open Chromatin Status and Genome Quality*

1044 To evaluate the relationship between predicted OCR ortholog open chromatin status and genome
1045 quality, we computed the correlation between the mean and standard deviation of predicted OCR
1046 ortholog open chromatin status obtained in the previous evaluation and the Glires' genome assemblies'
1047 scaffold and contig N50's. We obtained the scaffold and contig N50's from NCBI [3, 39] and computed
1048 the log base ten of each of them. We computed the correlations for predictions from multi-species brain
1049 and liver models, using brain and liver OCR orthologs, respectively. We also determined the relative
1050 association of phylogenetic distance and genome quality with predictions by fitting generalized linear
1051 models of mean and standard deviation of predicted OCR ortholog open chromatin status as a
1052 combination of divergence from mouse and scaffold or contig N50. In addition to comparing the effect
1053 sizes for the generalized linear models, we also computed the p-values on the coefficients and multiplied
1054 them by four to do a Bonferroni correction.

1055

1056 <u>Interpreting Deep Learning Models</u>

1057 We interpreted the deep learning models by computing the importance of every nucleotide in
1058 each true positive example in the validation set and then using these importance values to construct
1059 motifs. We computed the importance of every nucleotide in every true positive example in the validation
1060 set using deepLIFT, which calculates the extent to which each input contributes to the prediction relative

1061 to a reference [52]. We used the deepLIFT version 0.5.5-theano with the Rescale rule scores from the
1062 sequence layer with the target of the final convolutional layer, where our reference was a sequence of
1063 N's. We also used an extension to deepLIFT, also with the Rescale rule, to compute the "hypothetical
1064 scores" for each nucleotide at each position for each sequence, which can be thought of as the preference
1065 of the model for observing each nucleotide at each position in the sequence [53].

1066 We combined the scores and hypothetical scores using the TF-MoDISco method to construct "TF-
1067 MoDISco Motifs" [53]. TF-MoDISco first identifies frequently occurring sequence patterns with high
1068 deepLIFT scores within the sequences of each OCR (called seqlets), next computes a similarity matrix
1069 between all seqlets, and then uses the similarity matrix to cluster the seqlets into nonredundant motifs.
1070 We used the following settings for TF-MoDIsco: seqlet FDR threshold = 0.2; gapped k-mer settings for
1071 similarity computation k-mer length = 8, number of gaps = 1, and number of mismatches = 0; and final
1072 motif width = 50. We visualized our TF-MoDISco motifs from TF-MoDISco using the aggregated
1073 hypothetical scores of the seqlets supporting each motif. We created position frequency matrices from
1074 TF-MoDISco motifs by averaging the one-hot-encoded sequences at all of the seqlet coordinates
1075 belonging to the motifs and compared them to known motifs by running TomTom [101] on them with the
1076 *Mus musculus* motifs from CIS-BP [177].

1077

1078 <u>Comparing Predictions to Mean Conservation Scores</u>

1079 We compared the predictions to mean conservation scores by identifying OCR orthologs with
1080 conserved and non-conserved open chromatin status between species, computing the mean conservation
1081 scores of those OCR orthologs, and comparing those scores to the predicted open chromatin status of
1082 those OCR orthologs. We defined an OCR ortholog with conserved open chromatin status between mouse
1083 and another species as a mouse OCR whose ortholog in the other species overlaps an OCR in the same

1084 tissue. For mouse brain test set OCRs, we identified 441 OCRs with conserved open chromatin status in
1085 macaque, 195 OCRs with conserved open chromatin status in human, and 670 OCRs with conserved open
1086 chromatin status in rat. For mouse liver test set OCRs, we identified 689 OCRs with conserved open
1087 chromatin status in macaque and 580 OCRs with conserved open chromatin status in rat (**Figure 5,**
1088 **Supplemental Figure 7, Figure 7, Supplemental Figure 9**). We defined an OCR ortholog with non-
1089 conserved open chromatin status between mouse and another species as a mouse OCR whose ortholog
1090 in another species does not overlap any OCR from the pooled replicates from any dataset used in defining
1091 an OCR in that tissue in that species. For mouse brain test set OCRs, we identified 394 OCR orthologs with
1092 non-conserved open chromatin status in macaque, 448 OCR orthologs with non-conserved open
1093 chromatin status in human, and 338 OCR orthologs with non-conserved open chromatin status in rat. For
1094 mouse liver test set OCRs, we identified 1,114 OCR orthologs with non-conserved open chromatin status
1095 in macaque and 1,241 OCR orthologs with non-conserved open chromatin status in rat (**Figure 5,**
1096 **Supplemental Figure 7, Figure 7, Supplemental Figure 9**). We think that the differences in numbers of
1097 OCR orthologs with conserved and non-conserved open chromatin status between species is due not only
1098 to differences in evolutionary relatedness but also to differences between species in numbers of datasets
1099 used to define OCRs and differences in sequencing depths of those datasets [22, 41, 42].

1100 We used our models to predict the OCR ortholog open chromatin status for the open chromatin
1101 status-conserved and open chromatin status non-conserved OCR orthologs in the non-mouse species and
1102 compared it to the conservations scores of the mouse OCRs. We computed mean conservation scores of
1103 the mouse OCRs by calculating the mean PhastCons [82] and PhyloP [81] scores at the peak summits +/-
1104 250bp. We evaluated if the distributions of the predictions and each type of conservation score differed
1105 between the open chromatin status-conserved and open chromatin status non-conserved orthologs using
1106 a Wilcoxon rank-sum test. We did a Bonferroni correction by multiplying all p-values by 20 (2 conservation

1107 score comparisons and 2 model predictions comparisons – models trained on only mouse sequences and
1108 multi-species models – for 5 species, tissue pairs).

1109 We then evaluated whether the predictions were more effective than the mean conservation
1110 scores at differentiating between open chromatin status-conserved and open chromatin status non-
1111 conserved OCR orthologs. We first averaged the predictions of the sequence underlying the non-mouse
1112 OCR ortholog's summit +/-250bp and its reverse complement so that each OCR ortholog would have a
1113 single prediction value. We next combined our open chromatin status-conserved and open chromatin
1114 status non-conserved OCR orthologs and ranked them according to each of PhastCons score, PhyloP score,
1115 and OCR ortholog open chromatin status prediction. We then did a Wilcoxon sign-rank test to compare
1116 the ranking distributions of the open chromatin status-conserved OCR orthologs between the OCR
1117 ortholog open chromatin status predictions and each type of conservation score. We also did this for the
1118 ranking distributions of the open chromatin status non-conserved OCR orthologs. We did this for
1119 predictions made by the models trained using only mouse sequences and by the multi-species models.
1120 Finally, we did a Bonferroni correction by multiplying all p-values by 40 (2 conservation score comparisons
1121 for each of open chromatin status-conserved and open chromatin status non-conserved OCR orthologs in
1122 2 tissues for 5 species, tissue pairs).

1123

1124 <u>Obtaining and Visualizing Signal Tracks</u>

1125 We obtained the signal tracks used in **Figures 7b-e** using the pooled replicates fold-change bigwigs
1126 from the data processing pipelines. For the H3K27ac ChIP-seq data, we downloaded the mouse and
1127 macaque H3K27ac ChIP-seq data from [80] and reprocessed it using the AQUAS Transcription Factor and
1128 Histone ChIP-Seq processing pipeline [178] with default parameters, mapping reads to mm10 and
1129 rheMac8, respectively. We evaluated the data quality of each biological replicate based on the percentage

1130 of mapped reads, number of filtered reads, NSC, RSC, number of IDR reproducible peaks, rescue ratio, and
1131 self-consistency ratio analyses generated by the pipelines and found that all four biological replicates from
1132 each species were high-quality. Visualizations for these figures were created using the New WashU
1133 Epigenome Browser [179].

1134

1135

1136 **DECLARATIONS**

1137 **Ethics approval and consent to participate:** All human data is publicly available, and all other animal data
1138 is either publicly available or was collected in experiments approved by Carnegie Mellon University's
1139 Institutional Animal Care and Use Committee.

1140 **Consent for publication:** Not applicable.

1141 **Availability of data and materials:** Human DNase hypersensitivity data analyzed in this study was
1142 downloaded from the ENCODE portal (https://www.encodeproject.org/) [22, 47, 137], and human ATAC-
1143 seq data analyzed in this study was downloaded from Gene Expression Omnibus accession GSE96949 [42].
1144 Macaque and rat ATAC-seq data analyzed in this study can be accessed by contacting the authors of [41].
1145 Mouse liver ATAC-seq data in this manuscript can be accessed by contacting the authors. Mouse brain
1146 data analyzed in this study can be obtained by contacting the authors of [48]. Publicly available mouse
1147 liver ATAC-seq data was downloaded from China National Gene Bank accession CNP0000198 [43]. Other
1148 mouse ATAC-seq data was downloaded from China National Gene Bank accession CNP0000198 [43] and
1149 the ENCODE portal (https://www.encodeproject.org/) [138]. Mouse and macaque H3K27ac ChIP-seq data
1150 was downloaded from ArrayExpress accession E-MTAB-2633 [80]. All machine learning models trained in

this paper and code for using them can be found in this github repository: https://github.com/pfenninglab/OCROrthologPrediction.

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., 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., A.R.P., and M.E.W. designed the study. 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. I.M.K. and A.R.P. designed the machine learning model evaluation metrics with assistance from M.K. and M.E.W. I.M.K. wrote the manuscript with assistance from A.J.L. All authors reviewed and helped revise the manuscript.

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

FIGURE CAPTIONS

1174

1175 **Figure 1: OCR Ortholog Open Chromatin Status Prediction Framework Overview**

1176 We trained a CNN for predicting brain open chromatin using sequences underlying brain OCR orthologs
1177 in a small number of species and used the CNN to predict brain OCR ortholog open chromatin status across
1178 all of the species in the Zoonomia Project. In this toy example, the green rectangles on the left are brain
1179 OCRs, and the corresponding regions in other species are their orthologs. The orthologs colored in red-
1180 brown are from species for which we have brain open chromatin data and do not overlap brain OCRs, and
1181 the orthologs colored in white are from species for which we do not have open chromatin data. We used
1182 the sequences underlying the orthologs for which we have brain open chromatin data to train a CNN for
1183 predicting open chromatin. Then, we used the CNN to predict the probability of brain open chromatin for
1184 all brain OCR orthologs; predictions are illustrated on the right. Animal silhouettes were obtained from
1185 PhyloPic [180].

1186

1187 **Figure 2: Lineage-Specific OCR Accuracy Evaluation – Performance of Models Trained on Brain OCRs and**
1188 **Different Negative Sets**

1189 **a)** Illustration of different negative sets.

1190 **b)** Performance on test sets from negative sets used in model training.

1191 **c)** Performance on mouse brain test set OCRs whose orthologs in at least one other species are not brain
1192 OCRs and mouse brain test set non-OCRs whose orthologs in at least one other species are brain OCRs.

1193 **d)** Performance on macaque brain OCRs whose orthologs in mouse are not brain OCRs and are on test set
1194 chromosomes and macaque brain non-OCRs whose orthologs in mouse are brain OCRs and are on test set
1195 chromosomes.

1196 **e)** Performance on Euarchonta-specific brain OCRs and Euarchonta-specific brain non-OCRs whose mouse
1197 orthologs are on test set chromosomes.

1198 Animal silhouettes were obtained from PhyloPic [180]. *'s indicate the species from which sequences
1199 were obtained for making predictions.

1200

1201 **Figure 3: Tissue-Specific OCR Accuracy Evaluation – Predictions from Models Trained on Brain OCRs on**
1202 **Brain, Liver OCRs**

1203 We made test set predictions with each negative set on brain OCRs that do not overlap liver OCRs, brain
1204 OCRs that overlap liver OCRs, liver OCRs that do not overlap brain OCRs, and negative sets. p-Values were
1205 computed with a Wilcoxon rank-sum test and a Bonferroni correction. Mouse silhouette was obtained
1206 from PhyloPic [180].

1207

1208 **Figure 4: Phylogeny-Matching Correlations Evaluation – Divergence from Mouse versus Predictions in**
1209 **Glires**

1210 We compared the relationship between the phylogenetic distance between each Glires species and
1211 mouse and the mean brain OCR ortholog open chromatin status prediction across each Glires species.
1212 The yellow curve is the best fit exponential function of the form $y = ae^{bx}$. The yellow dotted line is the
1213 average prediction across the negatives in the test set. Animal silhouettes were obtained from PhyloPic
1214 [180].

1215

1216 **Figure 5: Mean Conservation Score or Open Chromatin Status Prediction Rankings versus Open**
1217 **Chromatin Conservation – Macaque**

1218 We ranked the PhastCons and PhyloP mouse mean conservation scores from as well as the brain and liver
1219 open chromatin status predictions of macaque OCR orthologs of the mouse OCRs whose open chromatin
1220 status is and is not conserved. We then plotted the cumulative distribution functions (CDFs) of the reverse
1221 rankings (largest has the highest ranking). The black dashed line is what we would expect from random
1222 rankings, and the blue dashed arrows show what we would expect from rankings that perfectly
1223 correspond to open chromatin status conservation. Animal silhouettes were obtained from PhyloPic
1224 [180].

1225

1226 **Figure 6: Multi-Species Model Performance**

1227 **a)** Performance of multi-species brain model on full test set, subset of test set consisting of clade-specific
1228 brain OCRs and non-OCRs, subset of test set consisting of species-specific brain OCRs and non-OCRs, and
1229 subset of positive test set consisting of OCRs shared between brain and liver as well as subset of liver OCRs
1230 from test set genomic regions that do not overlap brain OCRs.

1231 **b)** Performance of multi-species liver model on full test set, subset of test set consisting of clade-specific
1232 liver OCRs and non-OCRs, subset of test set consisting of species-specific liver OCRs and non-OCRs, and
1233 subset of positive test set consisting of OCRs shared between liver and brain as well as subset of brain
1234 OCRs from test set genomic regions that do not overlap liver OCRs. We reported area under the negative
1235 predictive value (NPV)-specificity curve instead of the area under the precision-recall curve because these
1236 test sets have more positives than negatives.

1237 **c)** Divergence from mouse versus mean multi-species brain model predictions across mouse brain OCR
1238 orthologs in Glires. The green curve is the best fit exponential function of the form $y = ae^{bx}$. The green
1239 dotted line is the average prediction across test set negatives. Animal silhouettes were obtained from
1240 PhyloPic [180].

1241 **d)** Divergence from mouse versus mean multi-species liver model predictions across mouse liver OCR
1242 orthologs in Glires. The red curve is the best fit exponential function of the form $y = ae^{bx}$. The red dotted
1243 line is the average prediction across the negatives in the test set. Animal silhouettes were obtained from
1244 PhyloPic [180].

1245

1246 **Figure 7: Mean Conservation Score or Open Chromatin Status Prediction Rankings versus Open**
1247 **Chromatin Conservation – Macaque**

1248 **a)** CDFs of the reverse rankings of the PhastCons and PhyloP mouse mean conservation scores and the
1249 multi-species brain and liver models' open chromatin status predictions of the macaque orthologs of the
1250 mouse OCRs whose open chromatin status is and is not conserved. The black dashed line is what we
1251 would expect from random rankings, and the blue dashed arrows show what we would expect from
1252 rankings that perfectly correspond to open chromatin status conservation.

1253 **b)** 7-week-old mouse cortex and striatum and macaque orofacial motor cortex ("Cortex") and putamen
1254 ("Striatum") open chromatin signal for a mouse brain OCR that is 50,328 bp away from the *Stx16* TSS.
1255 Experimentally identified and predicted brain open chromatin statuses are conserved even though mean
1256 mouse PhastCons score is low.

1257 **c)** 7-week-old mouse cortex and striatum and macaque orofacial motor cortex ("Cortex") and putamen
1258 ("Striatum") open chromatin signal for a mouse brain OCR that is 144,474 bp away from the *Lnpk* TSS.
1259 Experimentally identified and predicted brain open chromatin statuses are not conserved even though
1260 mean mouse PhastCons score is high.

1261 **d)** Our mouse liver and macaque liver open chromatin signal for a mouse liver OCR that is 24,814 bp away
1262 from the *Rxra* TSS. Experimentally identified and predicted liver open chromatin statuses are conserved
1263 even though mean mouse PhastCons score is low.

1264 **e)** Our mouse liver and macaque liver open chromatin signal for a mouse liver OCR that is 154,404 bp
1265 away from the *Fn1* TSS. Experimentally identified and predicted liver open chromatin statuses are not
1266 conserved even though mean mouse PhastCons score is high.

1267 Animal silhouettes were obtained from PhyloPic [180]. In **b-e**, regions are mouse cortex open chromatin
1268 peak summits +/-250bp and their macaque orthologs, signals are from pooled reads across biological
1269 replicates, and liver H3K27ac ChIP-seq data comes from [80].

1270

1271 **TABLES:**

1272

1273 **Table 1: Mouse Brain OCR Predictions by Mouse Sequence Brain Model on Macaque Orthologs versus**
1274 **Conservation Scores**

Conservation Score Type	Brain Open Chromatin Conserved	Brain Open Chromatin Not Conserved
PhastCons	3.69×10^{-4}	6.88×10^{-4}
PhyloP	9.30×10^{-6}	1.85×10^{-7}

1275

1276 **Table 2: Mouse Liver OCR Predictions by Mouse Sequence Liver Model on Macaque Orthologs versus**
1277 **Conservation Scores**

Conservation Score Type	Liver Open Chromatin Conserved	Liver Activity Not Conserved
PhastCons	6.25×10^{-17}	3.63×10^{-11}
PhyloP	1.35×10^{-22}	3.42×10^{-18}

1278

1279 **Table 3: Mouse Brain OCR Predictions by Multi-Species Brain Model on Macaque Orthologs versus**
1280 **Conservation Scores**

Conservation Score Type	Brain Open Chromatin Conserved	Brain Open Chromatin Not Conserved
PhastCons	7.80×10^{-5}	4.34×10^{-5}
PhyloP	1.53×10^{-6}	8.08×10^{-9}

1281

1282 **Table 4: Mouse Liver OCR Predictions by Multi-Species Liver Model on Macaque Orthologs versus**
1283 **Conservation Scores**

Conservation Score Type	Liver Open Chromatin Conserved	Liver Open Chromatin Not Conserved
PhastCons	3.68×10^{-22}	7.15×10^{-14}
PhyloP	1.25×10^{-27}	9.82×10^{-22}

1284

1285 **Table 5: Number of Positives and Negatives Used for Training, Tuning, and Testing Each Model**

Genomes Used in Training	Tissue	Negative Set	Positives (Training, Validation, Test)	Negatives (Training, Validation, Test)	Negatives:Positives (Training, Validation, Test)
mm10	Brain	Flanking Regions	21594, 2416, 4576	35640, 4018, 7440	1.65:1, 1.66:1, 1.63:1
mm10	Brain	OCRs in Other Tissues	21594, 2416, 4576	427174, 70504, 82172	19.78:1, 29.18:1, 17.96:1
mm10	Brain	Large G/C- and Repeat-Matched	21594, 2416, 4576	175912, 23880, 32008	8.15:1, 9.88:1, 6.99:1
mm10	Brain	Small G/C- and Repeat-Matched	21594, 2416, 4576	35358, 4776, 6654	1.64:1, 1.98:1, 1.45:1
mm10	Brain	Dinucleotide-Shuffled Enhs.	21594, 2416, 4576	215940, 24160, 45760	10:1, 10:1, 10:1
mm10	Brain	Non-OCR Orths. of OCRs	21594, 2416, 4576	25086, 3456, 4694	1.16:1, 1.43:1, 1.03:1
mm10	Liver	Non-OCR Orths. of OCRs	32498, 4032, 7752	22890, 2994, 4434	1:1.42, 1:1.35, 1:1.75
mm10, hg38, rheMac8, rn6	Brain	Non-OCR Orths. Of OCRs	74688, 9036, 15266	111206, 14650, 19688	1.49:1, 1.62:1, 1.29:1
mm10, rheMac8, rn6	Liver	Non-OCR Orths. of OCRs	81886, 10428, 17688	67278, 8680, 14544	1:1.22, 1:1.20, 1:1.22

1286

1287 **REFERENCES**

1288 1. Koepfli KP, Paten B, O'Brien SJ, Genome 10K Community of Scientists: **The Genome 10K Project:**
1289 **a way forward.** *Annu Rev Anim Biosci* 2015, **3:**57-111.
1290 2. Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, Bat1K Consortium: **Bat**
1291 **Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All**
1292 **Living Bat Species.** *Annu Rev Anim Biosci* 2018, **6:**23-46.
1293 3. Zoonomia Consortium: **A comparative genomics multitool for scientific discovery and**
1294 **conservation.** *Nature*, *in press*.
1295 4. Armstrong J, Hickey G, Diekhans M, Deran A, Fang Q, Xie D, Feng S, Stiller J, Genereux D,
1296 Johnson J, et al: **Progressive alignment with Cactus: a multiple-genome aligner for the**
1297 **thousand-genome era.** *bioRχiv*, 2019; doi:https://doi.org/10.1101/730531.
1298 5. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, London D, Song L, Lee
1299 BK, Iyer VR, et al: **Extensive evolutionary changes in regulatory element activity during human**
1300 **origins are associated with altered gene expression and positive selection.** *PLoS Genet* 2012,
1301 **8:**e1002789.
1302 6. King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees.** *Science* 1975,
1303 **188:**107-116.
1304 7. Wray GA: **The evolutionary significance of cis-regulatory mutations.** *Nat Rev Genet* 2007,
1305 **8:**206-216.
1306 8. Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell
1307 PV, Ganapathy G, et al: **Convergent transcriptional specializations in the brains of humans and**
1308 **song-learning birds.** *Science* 2014, **346:**1256846.
1309 9. Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, Blanco-Aguiar JA,
1310 Plyusnina IZ, Trut L, Villafuerte R, et al: **A comparison of brain gene expression levels in**
1311 **domesticated and wild animals.** *PLoS Genet* 2012, **8:**e1002962.
1312 10. Trut L, Oskina I, Kharlamova A: **Animal evolution during domestication: the domesticated fox**
1313 **as a model.** *Bioessays* 2009, **31:**349-360.
1314 11. Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov AV, Yim SH, Buffenstein R, Lee SR, Chang KT,
1315 Rhee H, et al: **Gene expression defines natural changes in mammalian lifespan.** *Aging Cell*
1316 2015, **14:**352-365.
1317 12. Ma S, Gladyshev VN: **Molecular signatures of longevity: Insights from cross-species**
1318 **comparative studies.** *Semin Cell Dev Biol* 2017, **70:**190-203.
1319 13. Fraser HB, Khaitovich P, Plotkin JB, Pääbo S, Eisen MB: **Aging and gene expression in the**
1320 **primate brain.** *PLoS Biol* 2005, **3:**e274.
1321 14. McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke
1322 DB, Schaar BT, et al: **Human-specific loss of regulatory DNA and the evolution of human-**
1323 **specific traits.** *Nature* 2011, **471:**216-219.
1324 15. Berger MJ, Wenger AM, Guturu H, Bejerano G: **Independent erosion of conserved transcription**
1325 **factor binding sites points to shared hindlimb, vision and external testes loss in different**
1326 **mammals.** *Nucleic Acids Res* 2018, **46:**9299-9308.
1327 16. Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL:
1328 **Subterranean mammals show convergent regression in ocular genes and enhancers, along**
1329 **with adaptation to tunneling.** *Elife* 2017, **6:**e25884.
1330 17. Shen YY, Liang L, Li GS, Murphy RW, Zhang YP: **Parallel evolution of auditory genes for**
1331 **echolocation in bats and toothed whales.** *PLoS Genet* 2012, **8:**e1002788.

1332 18. Young RL, Ferkin MH, Ockendon-Powell NF, Orr VN, Phelps SM, Pogány Á, Richards-Zawacki CL,
1333 Summers K, Székely T, Trainor BC, et al: **Conserved transcriptomic profiles underpin monogamy**
1334 **across vertebrates.** *Proc Natl Acad Sci U S A* 2019, **116:**1331-1336.
1335 19. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S,
1336 Martinez-Jimenez CP, Mackay S, et al: **Five-vertebrate ChIP-seq reveals the evolutionary**
1337 **dynamics of transcription factor binding.** *Science* 2010, **328:**1036-1040.
1338 20. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves n, Kutter C, Brown GD, Marshall A,
1339 Flicek P, Odom DT: **Waves of retrotransposon expansion remodel genome organization and**
1340 **CTCF binding in multiple mammalian lineages.** *Cell* 2012, **148:**335-348.
1341 21. Schwalie PC, Ward MC, Cain CE, Faure AJ, Gilad Y, Odom DT, Flicek P: **Co-binding by YY1**
1342 **identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate**
1343 **genomes.** *Genome Biol* 2013, **14:**1-15.
1344 22. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis
1345 AB, Wang H, Vernot B, et al: **The accessible chromatin landscape of the human genome.** *Nature*
1346 2012, **489:**75-82.
1347 23. Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E, et
1348 al: **Global reference mapping of human transcription factor footprints.** *Nature* 2020, **583:**729-
1349 736.
1350 24. Vermunt MW, Tan SC, Castelijns B, Geeven G, Reinink P, de Bruijn E, Kondova I, Persengiev S,
1351 Bontrop R, Cuppen E, et al: **Epigenomic annotation of gene regulatory alterations during**
1352 **evolution of the primate brain.** *Nat Neurosci* 2016, **19:**494-503.
1353 25. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK,
1354 Ching CW, et al: **Histone modifications at human enhancers reflect global cell-type-specific**
1355 **gene expression.** *Nature* 2009, **459:**108-112.
1356 26. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z,
1357 Wang J, Ziller MJ, et al: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015,
1358 **518:**317-330.
1359 27. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford
1360 GE, Dekker J, et al: **Defining functional DNA elements in the human genome.** *Proc Natl Acad Sci*
1361 *U S A* 2014, **111:**6131-6138.
1362 28. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM,
1363 Bilmes JA, Birney E, et al: **Integrative annotation of chromatin elements from ENCODE data.**
1364 *Nucleic Acids Res* 2013, **41:**827-841.
1365 29. Minnoye L, Taskiran II, Mauduit D, Fazio M, Van Aerschot L, Hulselmans G, Christiaens V,
1366 Makhzami S, Seltenhammer M, Karras P, et al: **Cross-species analysis of enhancer logic using**
1367 **deep learning.** *Genome Res* 2020.
1368 30. Chen L, Fish AE, Capra JA: **Prediction of gene regulatory enhancers across species reveals**
1369 **evolutionarily conserved sequence properties.** *PLoS Comput Biol* 2018, **14:**e1006484.
1370 31. Kelley DR: **Cross-species regulatory sequence activity prediction.** *PLoS Comput Biol* 2020,
1371 **16:**e1008050.
1372 32. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R: **Absence of a simple code:**
1373 **How transcription factors read the genome.** *Trends in Biochemical Sciences* 2014, **39:**381-399.
1374 33. Peng PC, Khoueiry P, Girardot C, Reddington JP, Garfield DA, Furlong EEM, Sinha S: **The Role of**
1375 **Chromatin Accessibility in cis-Regulatory Evolution.** *Genome Biol Evol* 2019, **11:**1813-1828.
1376 34. Rodríguez-Martínez JA, Reinke AW, Bhimsaria D, Keating AE, Ansari AZ: **Combinatorial bZIP**
1377 **dimers display complex DNA-binding specificity landscapes.** *eLife* 2017, **6:**e19272.

1378 35. LeCun Y, Jackel LD, Boser B, Denker JS, Graf HP, Guyon I, Henderson D, Howard RE, Hubbard W:
1379 **Handwritten digit recognition: applications of neural network chips and automatic learning.**
1380 *IEEE Communications Magazine* 1989, **27:**41-46.
1381 36. Quang D, Xie X: **DanQ: a hybrid convolutional and recurrent deep neural network for**
1382 **quantifying the function of DNA sequences.** *Nucleic Acids Res* 2016, **44:**e107.
1383 37. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J: **A novel connectionist**
1384 **system for unconstrained handwriting recognition.** *IEEE Trans Pattern Anal Mach Intell* 2009,
1385 **31:**855-868.
1386 38. Klein JC, Keith A, Agarwal V, Durham T, Shendure J: **Functional characterization of enhancer**
1387 **evolution in the primate lineage.** *Genome Biol* 2018, **19:**99.
1388 39. **Assembly [Internet] from National Library of Medicine (UC), National Center for Biotechnology**
1389 **Information.** *Bethesda, MD*. 2012. https://www.ncbi.nlm.nih.gov/assembly/. Accessed June
1390 2019.
1391 40. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D: **Cactus: Algorithms for genome**
1392 **multiple sequence alignment.** *Genome Res* 2011, **21:**1512-1528.
1393 41. Wirthlin M, Kaplow IM, Lawler AJ, He J, Phan BN, Brown AR, Stauffer WR, Pfenning, AR: **The**
1394 **Regulatory Evolution of the Primate Fine-Motor System.** *bioRxiv* 2020;
1395 doi:https://doi.org/10.1101/2020.10.27.356733.
1396 42. Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru MD, Reach SM, Motl J, Ehrlich ME, Hurd YL,
1397 Roussos P: **An atlas of chromatin accessibility in the adult human brain.** *Genome Res* 2018,
1398 **28:**1243-1252.
1399 43. 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**
1400 **atlas of chromatin accessibility in mouse tissues.** *Sci Data* 2019, **6:**65.
1401 44. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ: **ATAC-seq: A Method for Assaying Chromatin**
1402 **Accessibility Genome-Wide.** *Curr Protoc Mol Biol* 2015, **109:**21.29.21-29.
1403 45. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ: **Transposition of native chromatin**
1404 **for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and**
1405 **nucleosome position.** *Nature methods* 2013, **10:**1213-1218.
1406 46. John S, Sabo PJ, Canfield TK, Lee K, Vong S, Weaver M, Wang H, Vierstra J, Reynolds AP,
1407 Thurman RE, Stamatoyannopoulos JA: **Genome-scale mapping of DNase I hypersensitivity.** *Curr*
1408 *Protoc Mol Biol* 2013. doi:10.1002/0471142727.mb2127s103.
1409 47. The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human**
1410 **genome.** *Nature* 2012, **489:**57-74.
1411 48. Srinivasan C, Phan BN, Lawler AJ, Ramamurthy E, Kleyman M, Brown AR, Kaplow IM, Wirthlin
1412 ME, Pfenning AR: **Addiction-associated genetic variants implicate brain cell type- and region-**
1413 **specific cis-regulatory elements in addiction neurobiology.** *bioRχiv*, 2020;
1414 doi:https://doi.org/10.1101/2020.09.29.318329.
1415 49. Arvey A, Agius P, Noble WS, Leslie C: **Sequence and chromatin determinants of cell-type-**
1416 **specific transcription factor binding.** *Genome Res* 2012, **22:**1723-1734.
1417 50. Alipanahi B, Delong A, Weirauch MT, Frey BJ: **Predicting the sequence specificities of DNA- and**
1418 **RNA-binding proteins by deep learning.** *Nat Biotechnol* 2015, **33:**831-838.
1419 51. LeCun Y, Bengio Y, Hinton G: **Deep learning.** *Nature* 2015, **521:**436-444.
1420 52. Shrikumar A, Greenside P, Kundaje A: **Learning Important Features Through Propagating**
1421 **Activation Differences.** *PMLR* 2017, **70:**3145-3153.
1422 53. Shrikumar A, Tian K, Shcherbina A, Avsec Ž, Banerjee A, Sharmin M, Nair S, Kundaje A: **TF-**
1423 **MoDISco v0.4.2.2-alpha: Technical Note.** *arχiv* 2018; arXiv:1811.00416.

1424 54. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K,
1425 Mishiro T, et al: **Cohesin mediates transcriptional insulation by CCCTC-binding factor.** *Nature*
1426 2008, **451:**796-801.
1427 55. Isbel L, Prokopuk L, Wu H, Daxinger L, Oey H, Spurling A, Lawther AJ, Hale MW, Whitelaw E: **Wiz**
1428 **binds active promoters and CTCF-binding sites and is required for normal behaviour in the**
1429 **mouse.** *Elife* 2016, **5:**e15082.
1430 56. Herrera DG, Robertson HA: **Activation of c-fos in the brain.** *Prog Neurobiol* 1996, **50:**83-107.
1431 57. Berretta S, Parthasarathy HB, Graybiel AM: **Local release of GABAergic inhibition in the motor**
1432 **cortex induces immediate-early gene expression in indirect pathway neurons of the striatum.** *J*
1433 *Neurosci* 1997, **17:**4752-4763.
1434 58. Joo JY, Schaukowitch K, Farbiak L, Kilaru G, Kim TK: **Stimulus-specific combinatorial**
1435 **functionality of neuronal c-fos enhancers.** *Nat Neurosci* 2016, **19:**75-83.
1436 59. Yamada K, Gerber DJ, Iwayama Y, Ohnishi T, Ohba H, Toyota T, Aruga J, Minabe Y, Tonegawa S,
1437 Yoshikawa T: **Genetic analysis of the calcineurin pathway identifies members of the EGR gene**
1438 **family, specifically EGR3, as potential susceptibility candidates in schizophrenia.** *Proc Natl*
1439 *Acad Sci U S A* 2007, **104:**2815-2820.
1440 60. Swanberg SE, Nagarajan RP, Peddada S, Yasui DH, LaSalle JM: **Reciprocal co-regulation of EGR2**
1441 **and MECP2 is disrupted in Rett syndrome and autism.** *Hum Mol Genet* 2009, **18:**525-534.
1442 61. Kawase S, Kuwako K, Imai T, Renault-Mihara F, Yaguchi K, Itohara S, Okano H: **Regulatory factor**
1443 **X transcription factors control Musashi1 transcription in mouse neural stem/progenitor cells.**
1444 *Stem Cells Dev* 2014, **23:**2250-2261.
1445 62. Zhang D, Zeldin DC, Blackshear PJ: **Regulatory factor X4 variant 3: a transcription factor**
1446 **involved in brain development and disease.** *J Neurosci Res* 2007, **85:**3515-3522.
1447 63. Xu P, Morrison JP, Foley JF, Stumpo DJ, Ward T, Zeldin DC, Blackshear PJ: **Conditional ablation of**
1448 **the RFX4 isoform 1 transcription factor: Allele dosage effects on brain phenotype.** *PLoS One*
1449 2018, **13:**e0190561.
1450 64. Chen YC, Kuo HY, Bornschein U, Takahashi H, Chen SY, Lu KM, Yang HY, Chen GM, Lin JR, Lee YH,
1451 et al: **Foxp2 controls synaptic wiring of corticostriatal circuits and vocal communication by**
1452 **opposing Mef2c.** *Nat Neurosci* 2016, **19:**1513-1522.
1453 65. Harrington AJ, Raissi A, Rajkovich K, Berto S, Kumar J, Molinaro G, Raduazzo J, Guo Y, Loerwald
1454 K, Konopka G, et al: **MEF2C regulates cortical inhibitory and excitatory synapses and behaviors**
1455 **relevant to neurodevelopmental disorders.** *Elife* 2016, **5:**e20059.
1456 66. Mitchell AC, Javidfar B, Pothula V, Ibi D, Shen EY, Peter CJ, Bicks LK, Fehr T, Jiang Y, Brennand KJ,
1457 et al: **MEF2C transcription factor is associated with the genetic and epigenetic risk architecture**
1458 **of schizophrenia and improves cognition in mice.** *Mol Psychiatry* 2018, **23:**123-132.
1459 67. **The Human Protein Atlas.** http://www.proteinatlas.org. Accessed July 2020.
1460 68. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf
1461 C, Sjostedt E, Asplund A, et al: **Tissue-based map of the human proteome.** *Science* 2015,
1462 **347:**1260419-1260419.
1463 69. Schwindt H, Akasaka T, Zühlke-Jenisch R, Hans V, Schaller C, Klapper W, Dyer MJ, Siebert R,
1464 Deckert M: **Chromosomal translocations fusing the BCL6 gene to different partner loci are**
1465 **recurrent in primary central nervous system lymphoma and may be associated with aberrant**
1466 **somatic hypermutation or defective class switch recombination.** *J Neuropathol Exp Neurol*
1467 2006, **65:**776-782.
1468 70. Tiberi L, Bonnefont J, van den Ameele J, Le Bon SD, Herpoel A, Bilheu A, Baron BW,
1469 Vanderhaeghen P: **A BCL6/BCOR/SIRT1 complex triggers neurogenesis and suppresses**
1470 **medulloblastoma by repressing Sonic Hedgehog signaling.** *Cancer Cell* 2014, **26:**797-812.

1471 71. Yan L, Miyake S, Okamura H: **Distribution and circadian expression of dbp in SCN and extra-SCN**
1472 **areas in the mouse brain.** *J Neurosci Res* 2000, **59:**291-295.
1473 72. Gachon F, Fonjallaz P, Damiola F, Gos P, Kodama T, Zakany J, Duboule D, Petit B, Tafti M,
1474 Schibler U: **The loss of circadian PAR bZip transcription factors results in epilepsy.** *Genes Dev*
1475 2004, **18:**1397-1412.
1476 73. González-Velasco O, Papy-García D, Le Douaron G, Sánchez-Santos JM, De Las Rivas J:
1477 **Transcriptomic landscape, gene signatures and regulatory profile of aging in the human brain.**
1478 *Biochim Biophys Acta Gene Regul Mech* 2020, **1863:**194491.
1479 74. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S: **Comparative**
1480 **Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture.** *Cell Rep* 2015,
1481 **10:**1297-1309.
1482 75. Song D, Chu Z, Min L, Zhen T, Li P, Han L, Bu S, yang J, Gonzale FJ, Liu A: **Gemfibrozil not**
1483 **fenofibrate decreases systemic glucose level via PPARα.** *Pharmazie* 2016, **71:**205-212.
1484 76. Yang XN, Liu XM, Fang JH, Zhu X, Yang XW, Xiao XR, Huang JF, Gonzalez FJ, Li F: **PPARα Mediates**
1485 **the Hepatoprotective Effects of Nutmeg.** *J Proteome Res* 2018, **17:**1887-1897.
1486 77. Kersten S, Stienstra R: **The role and regulation of the peroxisome proliferator activated**
1487 **receptor alpha in human liver.** *Biochimie* 2017, **136:**75-84.
1488 78. Schrem H, Klempnauer J, Borlak J: **Liver-enriched transcription factors in liver function and**
1489 **development. Part II: the C/EBPs and D site-binding protein in cell cycle control,**
1490 **carcinogenesis, circadian gene regulation, liver regeneration, apoptosis, and liver-specific gene**
1491 **regulation.** *Pharmacol Rev* 2004, **56:**291-330.
1492 79. Hatzis P, Kyrmizi I, Talianidis I: **Mitogen-activated protein kinase-mediated disruption of**
1493 **enhancer-promoter communication inhibits hepatocyte nuclear factor 4alpha expression.** *Mol*
1494 *Cell Biol* 2006, **26:**7017-7029.
1495 80. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT,
1496 Jasinska AJ, et al: **Enhancer evolution across 20 mammalian species.** *Cell* 2015, **160:**554-566.
1497 81. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on**
1498 **mammalian phylogenies.** *Genome research* 2010, **20:**110-121.
1499 82. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J,
1500 Hillier LDW, Richards S, et al: **Evolutionarily conserved elements in vertebrate, insect, worm,**
1501 **and yeast genomes.** *Genome Research* 2005, **15:**1034-1050.
1502 83. Babeu JP, Boudreau F: **Hepatocyte nuclear factor 4-alpha involvement in liver and intestinal**
1503 **inflammatory networks.** *World J Gastroenterol* 2014, **20:**22-30.
1504 84. Hoffman BG, Robertson G, Zavaglia B, Beach M, Cullum R, Lee S, Soukhatcheva G, Li L, Wederell
1505 ED, Thiessen N, et al: **Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the**
1506 **functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver.** *Genome Research*
1507 2010, **20:**1037-1051.
1508 85. Alpern D, Langer D, Ballester B, Le Gras S, Romier C, Mengus G, Davidson I: **TAF4, a subunit of**
1509 **transcription factor II D, directs promoter occupancy of nuclear receptor HNF4A during post-**
1510 **natal hepatocyte differentiation.** *Elife* 2014, **3:**e03613.
1511 86. Sakaguchi M, Cai W, Wang CH, Cederquist CT, Damasio M, Homan EP, Batista T, Ramirez AK,
1512 Gupta MK, Steger M, et al: **FoxK1 and FoxK2 in insulin regulation of cellular and mitochondrial**
1513 **metabolism.** *Nat Commun* 2019, **10:**1582.
1514 87. Wang L, Liu Q, Kitamoto T, Hou J, Qin J, Accili D: **Identification of Insulin-Responsive**
1515 **Transcription Factors That Regulate Glucose Production by Hepatocytes.** *Diabetes* 2019,
1516 **68:**1156-1167.
1517 88. Liu X, Xu J, Rosenthal S, Zhang LJ, McCubbin R, Meshgin N, Shang L, Koyama Y, Ma HY, Sharma S,
1518 et al: **Identification of Lineage-Specific Transcription Factors That Prevent Activation of**

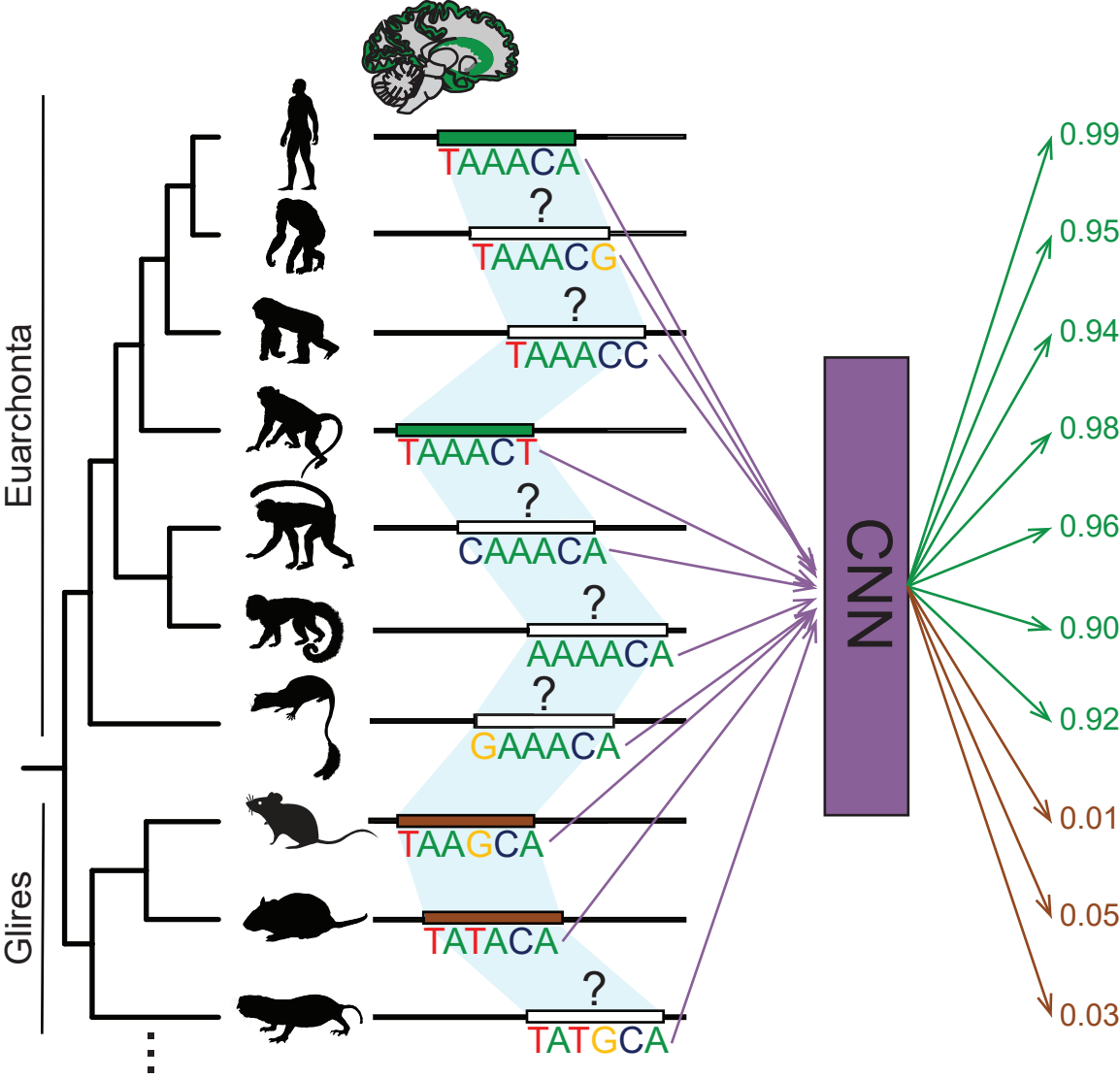
1519 **Hepatic Stellate Cells and Promote Fibrosis Resolution.** *Gastroenterology* 2020, **158:**1728-
1520 1744.e14.
1521 89. Park JS, Qiao L, Gilfor D, Yang MY, Hylemon PB, Benz C, Darlington G, Firestone G, Fisher PB,
1522 Dent P: **A role for both Ets and C/EBP transcription factors and mRNA stabilization in the**
1523 **MAPK-dependent increase in p21 (Cip-1/WAF1/mda6) protein levels in primary hepatocytes.**
1524 *Mol Biol Cell* 2000, **11:**2915-2932.
1525 90. Sugawara H, Iwata H, Souri M, Ichinose A: **Regulation of human protein Z gene expression by**
1526 **liver-enriched transcription factor HNF-4alpha and ubiquitous factor Sp1.** *J Thromb Haemost*
1527 2007, **5:**2250-2258.
1528 91. Kilbourne EJ, Widom R, Harnish DC, Malik S, Karathanasis SK: **Involvement of early growth**
1529 **response factor Egr-1 in apolipoprotein AI gene transcription.** *J Biol Chem* 1995, **270:**7004-
1530 7010.
1531 92. Plumb-Rudewiez N, Clotman F, Strick-Marchand H, Pierreux CE, Weiss MC, Rousseau GG,
1532 Lemaigre FP: **Transcription factor HNF-6/OC-1 inhibits the stimulation of the HNF-**
1533 **3alpha/Foxa1 gene by TGF-beta in mouse liver.** *Hepatology* 2004, **40:**1266-1274.
1534 93. Margagliotti S, Clotman F, Pierreux CE, Beaudry JB, Jacquemin P, Rousseau GG, Lemaigre FP: **The**
1535 **Onecut transcription factors HNF-6/OC-1 and OC-2 regulate early liver expansion by**
1536 **controlling hepatoblast migration.** *Dev Biol* 2007, **311:**579-589.
1537 94. LaPensee CR, Lin G, Dent AL, Schwartz J: **Deficiency of the transcriptional repressor B cell**
1538 **lymphoma 6 (Bcl6) is accompanied by dysregulated lipid metabolism.** *PLoS One* 2014,
1539 **9:**e97090.
1540 95. Sommars MA, Ramachandran K, Senagolage MD, Futtner CR, Germain DM, Allred AL, Omura Y,
1541 Bederman IR, Barish GD: **Dynamic repression by BCL6 controls the genome-wide liver response**
1542 **to fasting and steatosis.** *Elife* 2019, **8:**e43922.
1543 96. Tang W, Jiang YF, Ponnusamy M, Diallo M: **Role of Nrf2 in chronic liver disease.** *World J*
1544 *Gastroenterol* 2014, **20:**13079-13087.
1545 97. Xu D, Xu M, Jeong S, Qian Y, Wu H, Xia Q, Kong X: **The Role of Nrf2 in Liver Disease: Novel**
1546 **Molecular Mechanisms and Therapeutic Approaches.** *Front Pharmacol* 2018, **9:**1428.
1547 98. Chua CEL, Tang BL: **Syntaxin 16 is enriched in neuronal dendrites and may have a role in**
1548 **neurite outgrowth.** *Molecular Membrane Biology* 2009, **25**.
1549 99. Ray M, Zhang W: **Analysis of Alzheimer's disease severity across brain regions by topological**
1550 **analysis of gene co-expression networks.** *BMC Systems Biology* 2010, **4:**1-11.
1551 100. Chen L, Liu Z, Zhou B, Wei C, Zhou Y, Rosenfeld MG, Fu X-D, Chisholm AD, Jin Y: **CELF RNA**
1552 **binding proteins promote axon regeneration in C. elegans and mammals through alternative**
1553 **splicing of Syntaxins.** *eLife* 2016.
1554 101. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W: **Quantifying similarity between motifs.**
1555 *Genome Biology* 2007, **8:**R24.
1556 102. Mookherjee D, Majumder P, Mukherjee R, Chatterjee D, Kaul Z, Das SD, Sougrat R, Chakrabarti
1557 S, Chakrabarti O: **Cytosolic aggregates in presence of non-translocated proteins perturb**
1558 **endoplasmic reticulum structure and dynamics.** *Traffic* 2019, **20:**943-960.
1559 103. Spitz F, Gonzalez F, Duboule D: **A Global Control Region Defines a Chromosomal Regulatory**
1560 **Landscape Containing the HoxD Cluster.** *Cell* 2003, **113:**405-417.
1561 104. Breuss MW, An N, Song Q, Nguyen T, Stanley V, James KN, Musaev D, Chai G, Wirth SA,
1562 Anzenberg P, et al: **Mutations in LNPK, Encoding the Endoplasmic Reticulum Junction Stabilizer**
1563 **Lunapark, Cause a Recessive Neurodevelopmental Syndrome: The American Journal of Human**
1564 **Genetics.** *American Journal of Human Genetics* 2018, **103:**296-304.
1565 105. Zhang R, Wang Y, Li R, Chen G: **Transcriptional Factors Mediating Retinoic Acid Signals in the**
1566 **Control of Energy Metabolism.** *Int J Mol Sci* 2015, **16:**14210-14244.

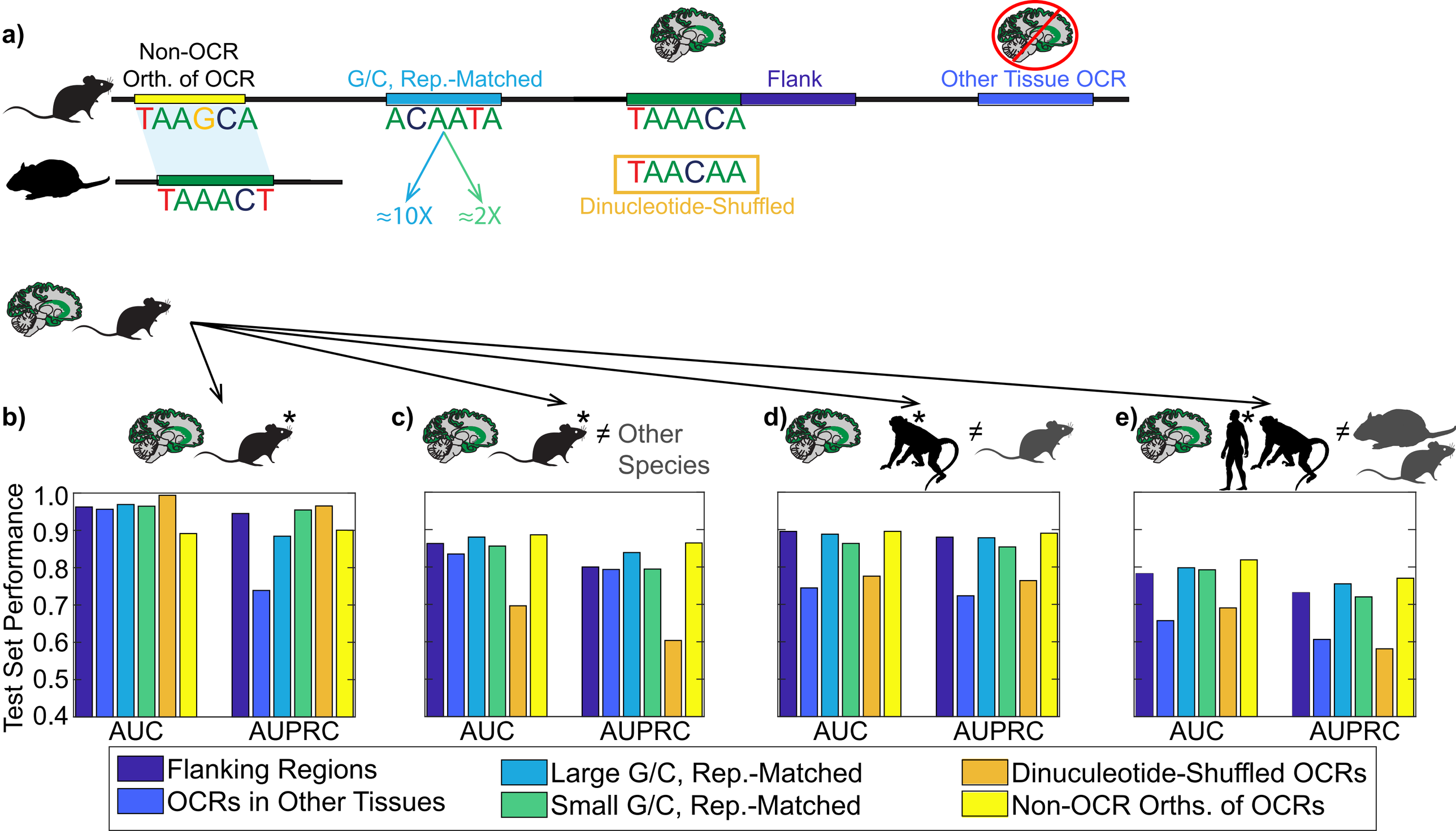
1567 106. Stossi F, Dandekar RD, Johnson H, Lavere P, Foulds CE, Mancini MG, Mancini MA: **Tributyltin**
1568 **chloride (TBT) induces RXRA down-regulation and lipid accumulation in human liver cells.** *PLoS*
1569 *One* 2019, **14:**e0224405.
1570 107. Xue Y, Guo C, Hu F, Zhu W, Mao S: **PPARA/RXRA signalling regulates the fate of hepatic non-**
1571 **esterified fatty acids in a sheep model of maternal undernutrition.** *Biochim Biophys Acta Mol*
1572 *Cell Biol Lipids* 2020, **1865:**158548.
1573 108. Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P: **Complexity and conservation of regulatory**
1574 **landscapes underlie evolutionary resilience of mammalian gene expression.** *Nat Ecol Evol*
1575 2018, **2:**152-163.
1576 109. Govaere O, Cockell S, Van Haele M, Wouters J, Van Delm W, Van den Eynde K, Bianchi A, van
1577 Eijsden R, Van Steenbergen W, Monbaliu D, et al: **High-throughput sequencing identifies**
1578 **aetiology-dependent differences in ductular reaction in human chronic liver disease.** *J Pathol*
1579 2019, **248:**66-76.
1580 110. Chen G, Wang R, Chen H, Wu L, Ge RS, Wang Y: **Gossypol ameliorates liver fibrosis in diabetic**
1581 **rats induced by high-fat diet and streptozocin.** *Life Sci* 2016, **149:**58-64.
1582 111. Hernandez-Gea V, Friedman SL: **Pathogenesis of liver fibrosis.** *Annu Rev Pathol* 2011, **6:**425-456.
1583 112. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-
1584 Petri A, Kircher M, et al: **The evolution of gene expression levels in mammalian organs.** *Nature*
1585 2011, **478:**343-348.
1586 113. Banovich NE, Li YI, Raj A, Ward MC, Greenside P, Calderon D, Tung PY, Burnett JE, Myrthil M,
1587 Thomas SM, et al: **Impact of regulatory variation across human iPSCs and differentiated cells.**
1588 *Genome Research* 2018, **28:**122-131.
1589 114. Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning-based**
1590 **sequence model.** *Nat Methods* 2015, **12:**931-934.
1591 115. Ghandi M, Lee D, Mohammad-Noori M, Beer MA: **Enhanced regulatory sequence prediction**
1592 **using gapped k-mer features.** *PLoS Comput Biol* 2014, **10:**e1003711.
1593 116. Zhang X, Kaplow IM, Wirthlin M, Park TY, Pfenning AR: **HALPER facilitates the identification of**
1594 **regulatory element orthologs across species.** *Bioinformatics* 2020.
1595 117. Hickey G, Paten B, Earl D, Zerbino D, Haussler D: **HAL: a hierarchical format for storing and**
1596 **analyzing multiple genome alignments.** *Bioinformatics* 2013, **29:**1341-1342.
1597 118. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W,
1598 Fungtammasan A, Gedman GL, et al: **Towards complete and error-free genome assemblies of**
1599 **all vertebrate species.** *bioRχiv* 2020; doi:https://doi.org/10.1101/2020.05.22.110833.
1600 119. Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, Winkler S, Jermiin LS,
1601 Skirmuntt EC, Katzourakis A, et al: **Six reference-quality genomes reveal evolution of bat**
1602 **adaptations.** *Nature* 2020, **583:**578-584.
1603 120. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM,
1604 Gifford DK, Fraenkel E: **Tissue-specific transcriptional regulation has diverged significantly**
1605 **between human and mouse.** *Nat Genet* 2007, **39:**730-732.
1606 121. Carbone L, Harris RA, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ,
1607 Herrero J, Roos C, Aken B, et al: **Gibbon genome and the fast karyotype evolution of small**
1608 **apes.** *Nature* 2014, **513:**195-201.
1609 122. Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG, Corbo JC: **Massively parallel cis-**
1610 **regulatory analysis in the mammalian central nervous system.** *Genome Res* 2016, **26:**238-255.
1611 123. Achour M, Le Gras S, Keime C, Parmentier F, Lejeune FX, Boutillier AL, Neri C, Davidson I,
1612 Merienne K: **Neuronal identity genes regulated by super-enhancers are preferentially down-**
1613 **regulated in the striatum of Huntington's disease mice.** *Hum Mol Genet* 2015, **24:**3481-3496.

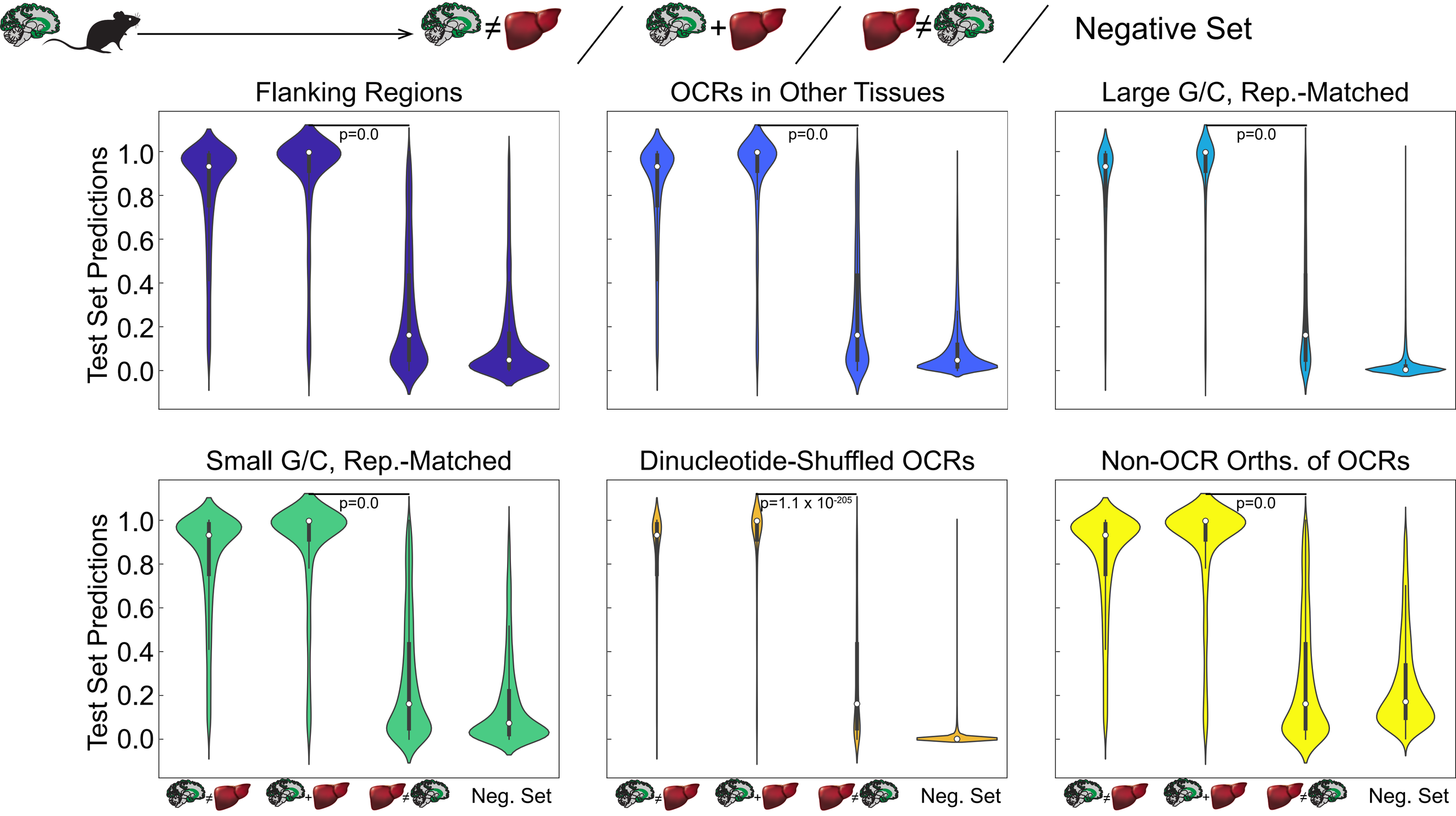
124. Long HK, Prescott SL, Wysocka J: **Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution.** *Cell* 2016, **167:**1170-1187.
125. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al: **DNase I sensitivity QTLs are a major determinant of human expression variation.** *Nature* 2012, **482:**390-394.
126. Khoueiry P, Girardot C, Ciglar L, Peng PC, Gustafson EH, Sinha S, Furlong EE: **Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity.** *Elife* 2017, **6:**e28440. **A Tutorial on Bayesian Optimization**
127. Frazier PI: **A Tutorial on Bayesian Optimization.** *arχiv* 2018; arXiv:1807.02811.
128. Bergstra J, Yamins D, Cox D: **Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures.** *PMLR* 2013, **28:**115--123.
129. Snoek J, Larochelle H, Adams RP: **Practical Bayesian Optimization of Machine Learning Algorithms.** *Advances in Neural Information Processing Systems* 2019, 2951-2959.
130. Lundberg SM, Lee S-I: **A Unified Approach to Interpreting Model Predictions.** *Advances in Neural Information Processing Systems* 2019, 4765-4774.
131. Feser J, Tyler J: **Chromatin structure as a mediator of aging.** *FEBS Lett* 2011, **585:**2041-2048.
132. Bryois J, Garrett ME, Song L, Safi A, Giusti-Rodriguez P, Johnson GD, Shieh AW, Buil A, Fullard JF, Roussos P, et al: **Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia.** *Nat Commun* 2018, **9:**3121.
133. Hor CN, Yeung J, Jan M, Emmenegger Y, Hubbard J, Xenarios I, Naef F, Franken P: **Sleep-wake-driven and circadian contributions to daily rhythms in gene expression and chromatin accessibility in the murine cortex.** *Proc Natl Acad Sci U S A* 2019, **116:**25773-25783.
134. Qureshi IA, Mehler MF: **Genetic and epigenetic underpinnings of sex differences in the brain and in neurological and psychiatric disease susceptibility.** *Prog Brain Res* 2010, **186:**77-95.
135. Forger NG: **Epigenetic mechanisms in sexual differentiation of the brain and behaviour.** *Philos Trans R Soc Lond B Biol Sci* 2016, **371:**20150114.
136. Sugathan A, Waxman DJ: **Genome-wide analysis of chromatin states reveals distinct mechanisms of sex-dependent gene regulation in male and female mouse liver.** *Mol Cell Biol* 2013, **33:**3594-3610.
137. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymurdov UK, Narayanan AK, et al: **The Encyclopedia of DNA elements (ENCODE): data portal update.** *Nucleic Acids Res* 2018, **46:**D794-D801.
138. 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:**418.
139. Megquier K, Genereux DP, Hekman J, Swofford R, Turner-Maier J, Johnson J, Alonso J, Li X, Morrill K, Anguish LJ, et al: **BarkBase: Epigenomic Annotation of Canine Genomes.** *Genes (Basel)* 2019, **10:**433.
140. Giuffra E, Tuggle CK, FAANG Consortium: **Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap.** *Annu Rev Anim Biosci* 2019, **7:**65-88.
141. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ: **Single-cell chromatin accessibility reveals principles of regulatory variation.** *Nature* 2015, **523:**486-490.
142. Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G: **A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species.** *Cell Rep* 2012, **2:**817-823.
143. Sudmant PH, Alexis MS, Burge CB: **Meta-analysis of RNA-seq expression data across species, tissues and studies.** *Genome Biol* 2015, **16:**287.

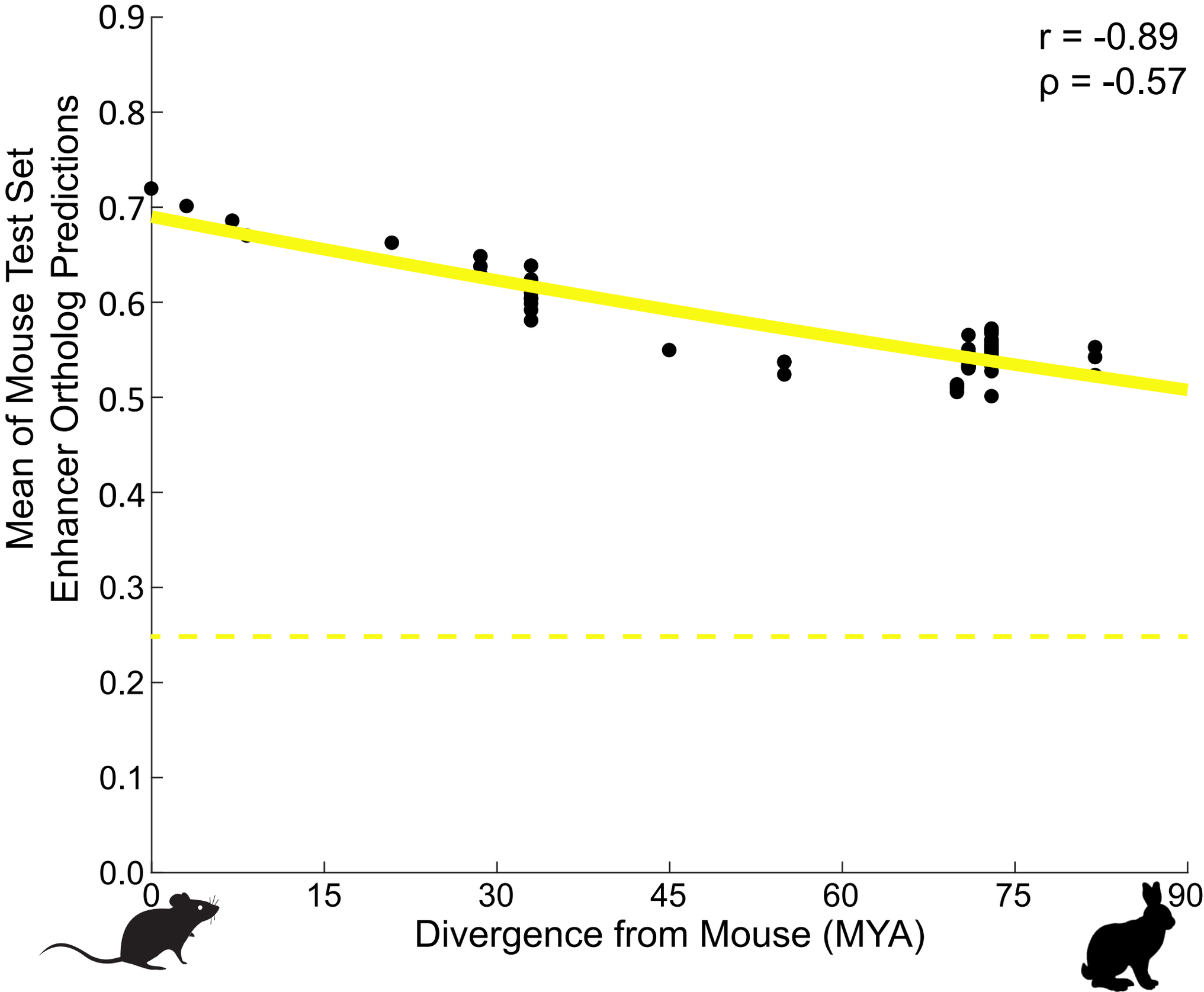
144. Tosches MA, Yamawaki TM, Naumann RK, Jacobi AA, Tushev G, Laurent G: **Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles.** *Science* 2018, **360:**881-888.
145. Zhu Y, Sousa AMM, Gao T, Skarica M, Li M, Santpere G, Esteller-Cucala P, Juan D, Ferrández-Peral L, Gulden FO, et al: **Spatiotemporal transcriptomic divergence across human and macaque brain development.** *Science* 2018, **362:**eaat8077.
146. Madisen L, Zwingman TA, Sunkin SM, Oh SW, Zariwala HA, Gu H, Ng LL, Palmiter RD, Hawrylycz MJ, Jones AR, et al: **A robust and high-throughput Cre reporting and characterization system for the whole mouse brain.** *Nat Neurosci* 2010, **13:**133-140.
147. Lee JW, Foo CS, Kim D, Boley N, Kundaje A: **ATAC-Seq / DNase-Seq Pipeline.** *Availabile at* https://github.com/kundajelab/atac_dnase_pipelines.
148. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
149. Amemiya HM, Kundaje A, Boyle AP: **The ENCODE Blacklist: Identification of Problematic Regions of the Genome.** *Sci Rep* 2019, **9:**9354.
150. **ENCODE ATAC-seq Pipeline.** https://github.com/ENCODE-DCC/atac-seq-pipeline. Accessed September 2017.
151. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, et al: **Evolutionary and biomedical insights from the rhesus macaque genome.** *Science* 2007, **316:**222-234.
152. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
153. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428:**493-521.
154. Li Q, Brown JB, Huang H, Bickel PJ: **Measuring reproducibility of high-throughput experiments.** *Annals of Applied Statistics* 2011, **5:**1752-1779.
155. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.** *Genome Research* 2012, **22:**1813-1831.
156. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.
157. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al: **GENCODE: The reference human genome annotation for the ENCODE project.** *Genome Research* 2012, **22:**1760-1774.
158. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al: **GENCODE reference annotation for the human and mouse genomes.** *Nucleic Acids Res* 2019, **47:**D766-D773.
159. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Res* 2016, **44:**D733-D745.
160. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes.** *Proc Natl Acad Sci U S A* 2003, **100:**11484-11489.
161. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based Analysis of ChIP-Seq (MACS).** *Genome Biology* 2008, **9:**R137.

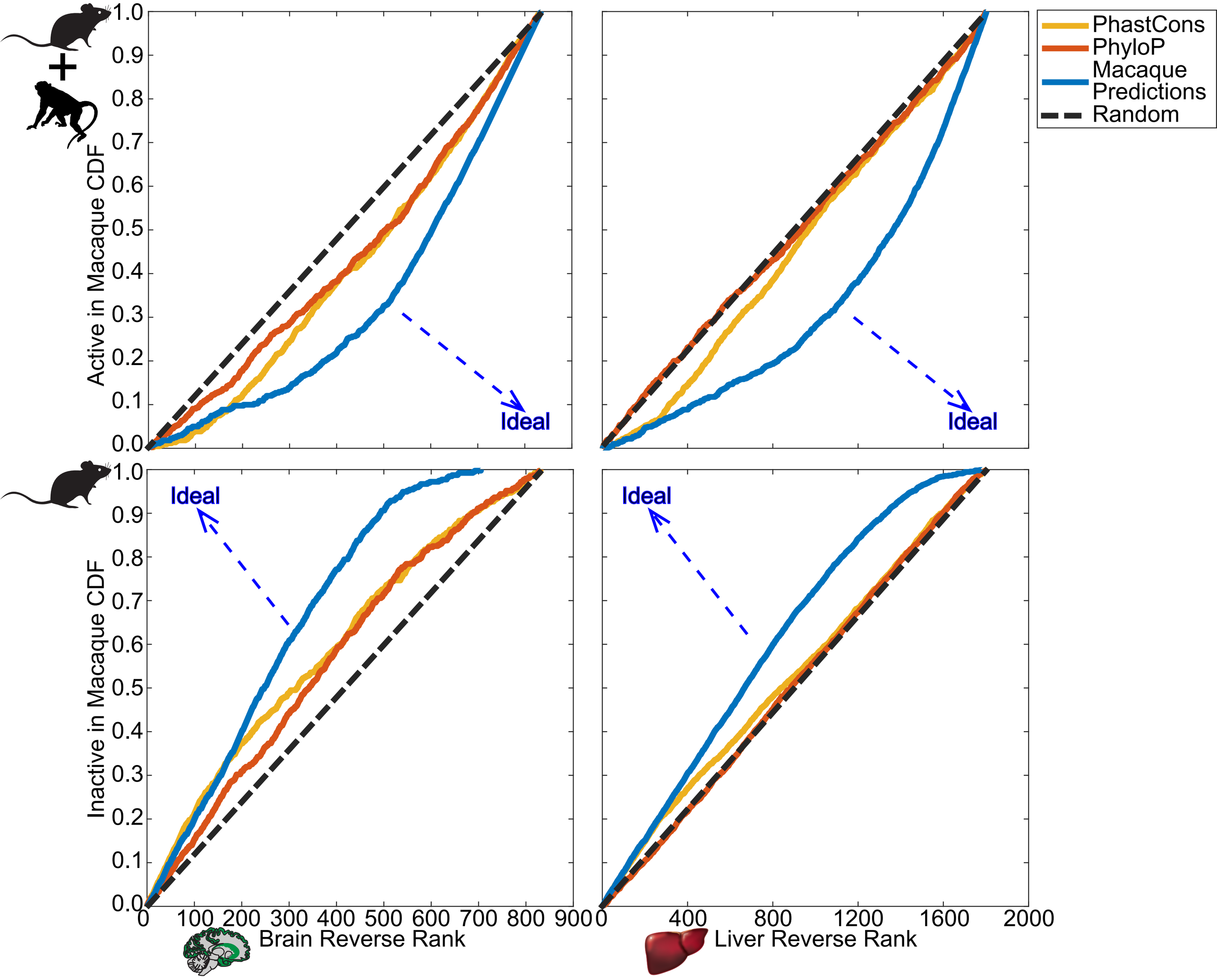
1710 162. Bailey TL, Machanick P: **Inferring direct DNA binding from ChIP-seq.** *Nucleic Acids Res* 2012,
1711 **40:**e128.
1712 163. Dale RK, Pedersen BS, Quinlan AR: **Pybedtools: A flexible Python library for manipulating**
1713 **genomic datasets and annotations.** *Bioinformatics* 2011, **27:**3423-3424.
1714 164. Pagès H: ***BSgenome: Software infrastructure for efficient representation of full genomes and***
1715 ***their SNPs*.** R package version 1.54.0. Accessed December 2019.
1716 165. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler aD: **The Human**
1717 **Genome Browser at UCSC.** *Genome Research* 2002, **12:**996-1006.
1718 166. Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA: **gkmSVM: an R**
1719 **package for gapped-kmer SVM.** *Bioinformatics* 2016, **32:**2205-2207.
1720 167. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME**
1721 **Suite: Tools for motif discovery and searching.** *Nucleic Acids Research* 2009, **37:**W202-W208.
1722 168. Patrushev LI, Kovalenko TF: **Functions of noncoding sequences in mammalian genomes.**
1723 *Biochemistry (Mosc)* 2014, **79:**1442-1469.
1724 169. Kelley DR, Snoek J, Rinn JL: **Basset: Learning the regulatory code of the accessible genome with**
1725 **deep convolutional neural networks.** *Genome Research* 2016, **26:**990-999.
1726 170. He K, Zhang X, Ren S, Sun J: **Delving deep into rectifiers: Surpassing human-level performance**
1727 **on imagenet classification.** *Proceedings of the IEEE International Conference on Computer Vision*
1728 2016, **11:**1026-1034.
1729 171. Chollet F: **Keras.** https://keras.io. Accessed October 2017.
1730 172. Team TTD, Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F,
1731 Bayer J, Belikov A, et al: **Theano: A Python framework for fast computation of mathematical**
1732 **expressions.** 2016.
1733 173. Pedregosa F, Varoquaux G: **Scikit-learn: Machine learning in Python.** *Journal of Machine*
1734 *Learning Research* 2011, **12:**2825-2830.
1735 174. Grau J, Grosse I, Keilwagen J: **PRROC: Computing and visualizing Precision-recall and receiver**
1736 **operating characteristic curves in R.** *Bioinformatics* 2015, **31:**2595-2597.
1737 175. **rpy2 - R in Python.** https://rpy2.github.io/. Accessed October 2017.
1738 176. Kumar S, Stecher G, Suleski M, Hedges SB: **TimeTree: A Resource for Timelines, Timetrees, and**
1739 **Divergence Times.** *Mol Biol Evol* 2017, **34:**1812-1819.
1740 177. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS,
1741 Lambert SA, Mann I, Cook K, et al: **Determination and Inference of Eukaryotic Transcription**
1742 **Factor Sequence Specificity.** *Cell* 2014, **158:**1431-1443.
1743 178. Lee JW, Boley N, Kundaje A: **AQUAS Transcription Factor and Histone ChIP-Seq processing**
1744 **pipeline.** https://github.com/kundajelab/chipseq_pipeline. Accessed September 2017.
1745 179. Li D, Hsu S, Purushotham D, Sears RL, Wang T: **WashU Epigenome Browser update 2019.**
1746 *Nucleic Acids Res* 2019, **47:**W158-W165.
1747 180. **PhyloPic.** http://phylopic.org/. Accessed July 2020.

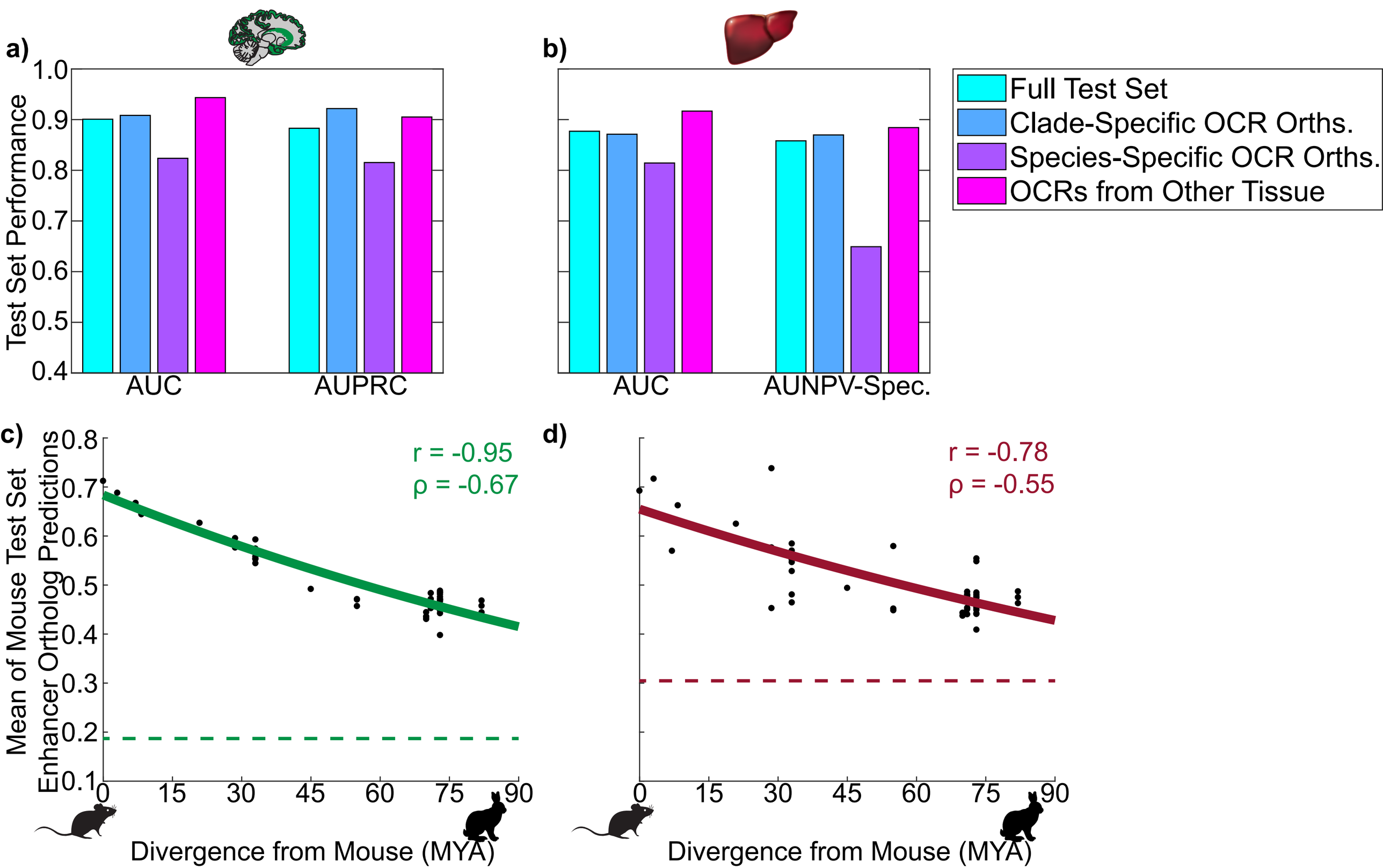

Euarchonta
Glires
TAAACA
?
TAAACG
?
TAAACC
TAAACT
?
CAAACA
?
AAAACA
?
GAAACA
TAAGCA
TATACA
?
TATGCA
CNN
0.99
0.95
0.94
0.98
0.96
0.90
0.92
0.01
0.05
0.03

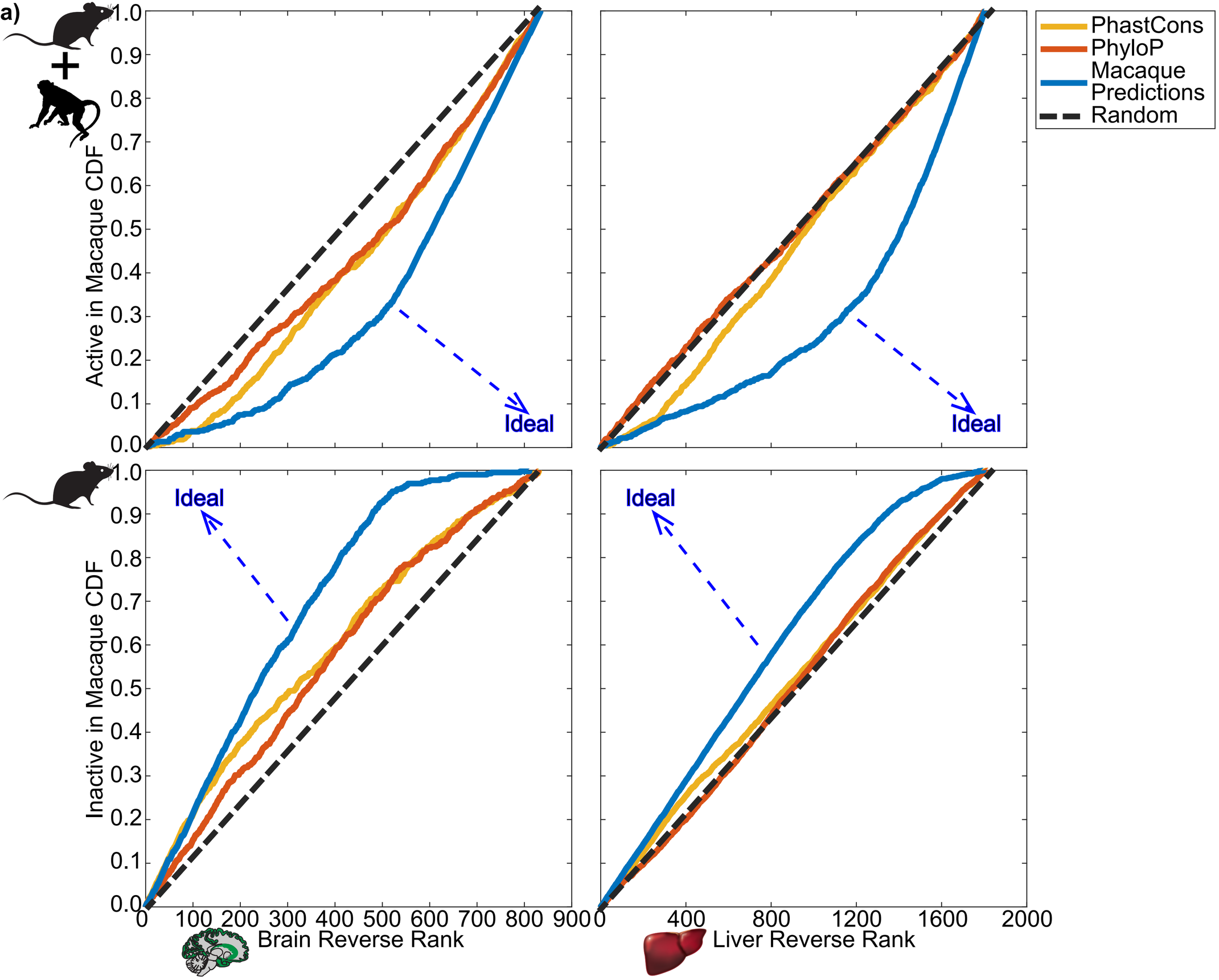

a)
Non-OCR
Orth. of OCR
G/C, Rep.-Matched
Flank
Other Tissue OCR
TAAGCA
ACAATA
TAAACA
TAAACT
≈10X
≈2X
TAACAA
Dinucleotide-Shuffled
b)
c)
≠ Other Species
d)
≠
e)
≠
Test Set Performance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
AUC
AUPRC
Flanking Regions
OCRs in Other Tissues
Large G/C, Rep.-Matched
Small G/C, Rep.-Matched
Dinuculeotide-Shuffled OCRs
Non-OCR Orths. of OCRs

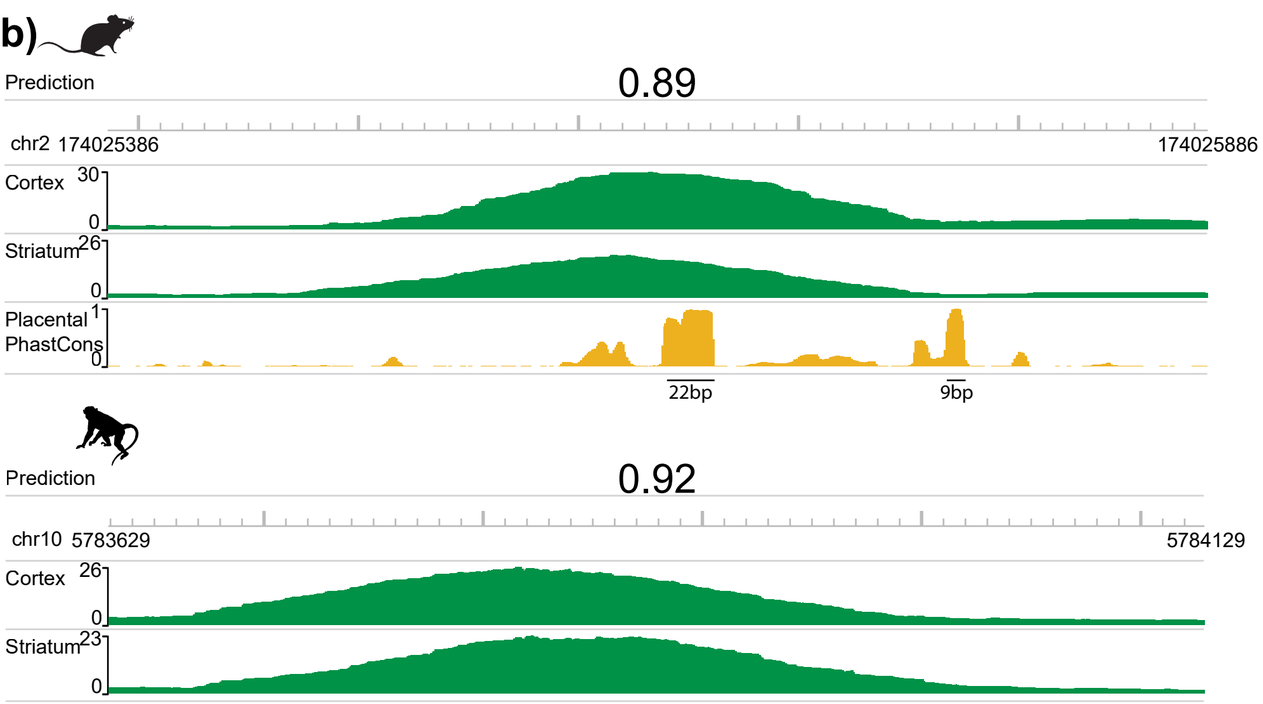

Negative Set
Flanking Regions
OCRs in Other Tissues
Large G/C, Rep.-Matched
Small G/C, Rep.-Matched
Dinucleotide-Shuffled OCRs
Non-OCR Orths. of OCRs
Test Set Predictions
1.0
0.8
0.6
0.4
0.2
0.0
p=0.0
p=0.0
p=0.0
p=0.0
p=1.1 x 10-205
p=0.0
Neg. Set
Neg. Set
Neg. Set

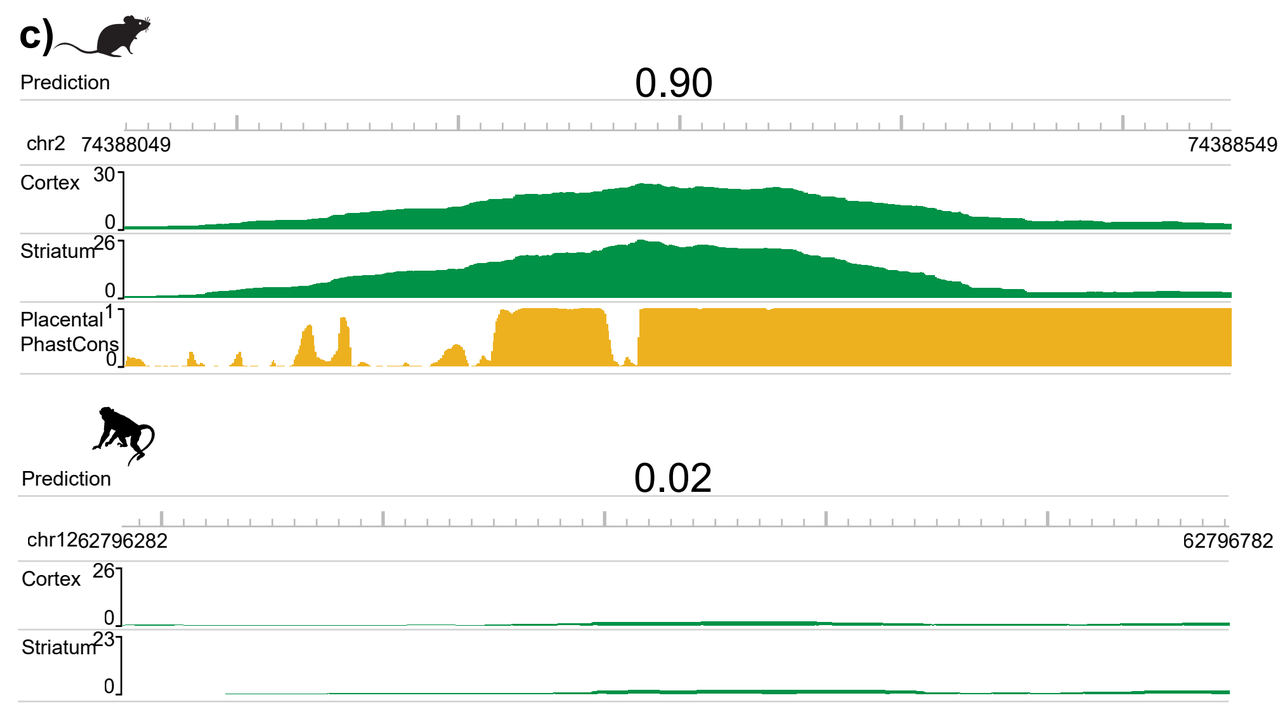

r = -0.89
ρ = -0.57
Mean of Mouse Test Set
Enhancer Ortholog Predictions
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
15
30
45
60
75
90
Divergence from Mouse (MYA)

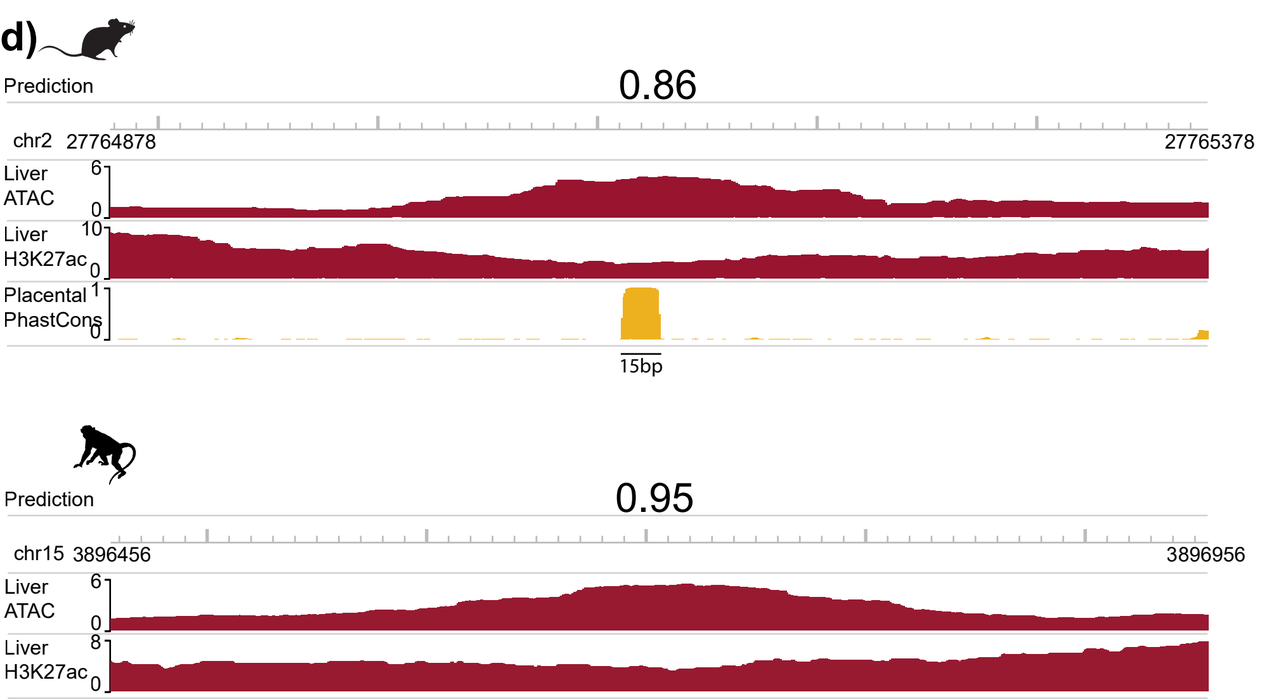

PhastCons
PhyloP
Macaque Predictions
Random
Active in Macaque CDF
Inactive in Macaque CDF
Ideal
Ideal
Ideal
Ideal
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
100
200
300
400
500
600
700
800
900
Brain Reverse Rank
0
400
800
1200
1600
2000
Liver Reverse Rank

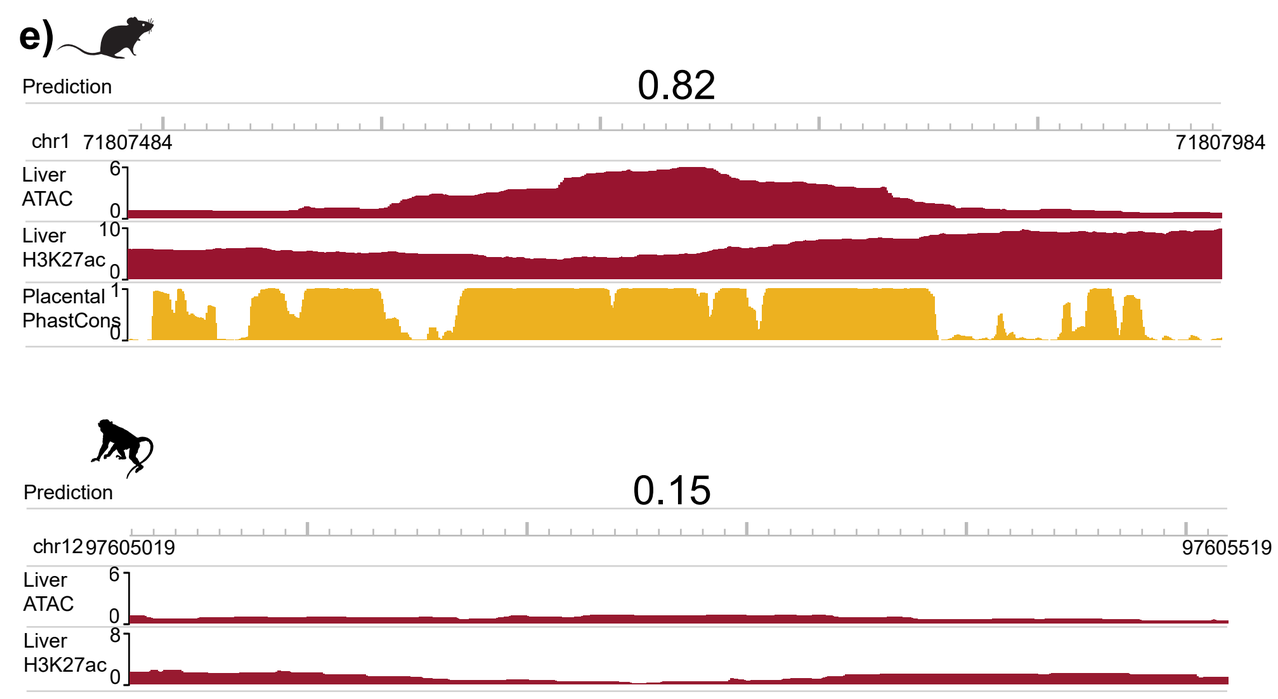

a)
b)
c)
d)
Test Set Performance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
AUC
AUPRC
AUC
AUNPV-Spec.
Full Test Set
Clade-Specific OCR Orths.
Species-Specific OCR Orths.
OCRs from Other Tissue
Mean of Mouse Test Set Enhancer Ortholog Predictions
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
r = -0.95
ρ = -0.67
r = -0.78
ρ = -0.55
0
15
30
45
60
75
90
Divergence from Mouse (MYA)
Divergence from Mouse (MYA)


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


b)
Prediction
0.89
chr2 174025386
174025886
Cortex
Striatum
Placental PhastCons
22bp
9bp
Prediction
0.92
chr10 5783629
5784129
Cortex
Striatum


c)
Prediction
0.90
chr2 74388049
74388549
Cortex
Striatum
Placental PhastCons
Prediction
0.02
chr12 62796282
62796782
Cortex
Striatum


d)
Prediction
0.86
chr2 27764878
27765378
Liver ATAC
Liver H3K27ac
Placental PhastCons
15bp
Prediction
0.95
chr15 3896456
3896956
Liver ATAC
Liver H3K27ac


e)
Prediction
0.82
chr1 71807484
71807984
Liver ATAC
Liver H3K27ac
Placental PhastCons
Prediction
0.15
chr12 97605019
97605519
Liver ATAC
Liver H3K27ac