Augmented prediction of multi-species protein–RNA interactions using evolutionary conservation of RNA-binding proteins

Jiale He[1,*], Tong Zhou[1,*], Lu-Feng Hu[2,3,*], Yuhua Jiao[1], Junhao Wang[1], Shengwen Yan[1], Siyao Jia[1], Qiuzhen Chen[1], Yangming Wang[2,3], Yucheng T. Yang[4,5,#], Lei Sun[1,6,#]

[1] Shandong Provincial Key Laboratory of Animal Cell and Developmental Biology, School of Life Sciences, Shandong University, Qingdao 266237, China

[2] State Key Laboratory of Gene Function and Modulation Research, Institute of Molecular Medicine, College of Future Technology, Peking University, Beijing 100871, China

[3] Beijing Advanced Center of RNA Biology (BEACON), Peking University, Beijing 100871, China

[4] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China

[5] MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, and MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200433, China

[6] State Key Laboratory of Microbial Technology, Shandong University, Qingdao 266237, China

*These authors contributed equally to this work

[#]Correspondence: yangyy@fudan.edu.cn (Y.T.Y.); sunlei0227@sdu.edu.cn (L.S.)

Highlights

- MuSIC integrates RBP-binding peaks with conservation-weighted label smoothing to predict RBP–RNA interactions across eleven species
- MuSIC outperforms state-of-the-art computational methods in predicting cross-species RBP–RNA interactions
- Cross-species prediction accuracy of RBP-binding peaks correlates with the conservation of RBPs
- MuSIC quantifies the effects of homologous SNVs on RBP binding with experimental validation in mouse

Abstract

RNA-binding proteins (RBPs) play critical roles in gene expression regulation. Recent studies have begun to detail the RNA recognition mechanisms of diverse RBPs. However, given the array of RBPs studied so far, it is implausible to experimentally profile RBP-binding peaks for hundreds of RBPs in multiple non-model organisms. Here, we introduce MuSIC (**Mu**lti-**S**pecies RBP–RNA **I**nteractions using **C**onservation), a deep learning-based framework for predicting cross-species RBP–RNA interactions by leveraging label smoothing and evolutionary conservation of RBPs across 11 diverse species ranging from human to yeast. MuSIC outperforms state-of-the-art computational methods, and provides predicted RBP-binding peaks across species with high accuracy. The prediction confidence is higher in the closely related species, partially due to the RBP conservation patterns. Finally, the effects of homologous genetic variants on RBP binding can be computationally quantified across species, followed by experimental validations. The target transcripts with disrupted binding events are enriched with the ubiquitination-associated pathways. To summarize, MuSIC provides a useful computational framework for predicting RBP–RNA interactions cross-species and quantifying the effects of genetic variants on RBP binding, offering novel insights into the RBP-mediated regulatory mechanisms implicated in human diseases.

Keywords

RBP–RNA interaction, deep learning, protein conservation, cross-species prediction, post-transcriptional regulation

Introduction

RNA-binding proteins (RBPs) play an important role that interact with intracellular RNAs by recognizing specific RNA motifs, and perform important biological functions including RNA transcription, splicing, transport, translation, and stability [1–3]. Disruptions in RBP–RNA interactions have been implicated in the pathogenesis of various diseases, particularly cancer and neurodegenerative diseases [4,5].

Genetic variants, including single nucleotide polymorphisms (SNPs) and single nucleotide variants (SNVs), can alter the RBP-binding affinity, thereby disrupting RNA recognition and contributing to the development of various diseases [6–8]. For example, the genetic variants in the iron-responsive element of *FTL* can impair the binding of IRP1 to *FTL* mRNA, leading to hyperferritinaemia-cataract syndrome [9]. Similarly, the genetic variants in *TCF3* can hinder the binding of hnRNPH1 to exon 18b, thus disrupting alternative splicing of *TCF3* and promoting Burkitt lymphoma development [10]. These cases highlight the critical roles of RBP–RNA interactions in human disease pathogenesis.

Model organisms, such as mouse and zebrafish, are widely used in revealing the pathogenic mechanisms of complex human diseases[11,12], where normal RBP–RNA interactions could be potentially disrupted by genetic variants [13,14]. Various high-throughput experimental methods have been applied to identify RBP-binding peaks both *in vitro* [15–17] and *in vivo* [18–21] . For example, PAR-CLIP technology has been applied to the mouse for exploring the regulatory roles of IGF2BP3 and LIN28B in hematopoietic reprogramming [22]. However, it is still unrealistic to generate the binding profiles of hundreds of RBPs in human and animal models using *in vivo* high-throughput experimental methods.

With the recent advancements in AI technologies, computational methods leveraging large-scale RBP–RNA binding data can effectively learn RBP binding characteristics [19,23,24]. Several computational methods using machine learning [25,26] and deep learning [27–31] technologies have been developed to predict RBP-binding

peaks and model their RNA sequence preferences. However, the predictions by these traditional computational methods are generally restricted in the specific species, where the CLIP-seq datasets are generated. Given the fact that only a minority of the RBPs that have been studied by CLIP-seq technologies are from non-human organisms—65 RBPs identified in yeast, 45 in mouse, and only 5-6 in species such as fly, worm, and *A. thaliana* [32], it is vital to develop novel computational methods to predict RBP–RNA interactions in the cross-species tasks[33,34].

The style transfer methods have been successfully applied in the field of computer vision, where similarity-guided label smoothing strategy could improve the cross-style performance [35–37]. In our study, the analogies of the similarity features in label smoothing can be the sequence and structural conservation of RBPs. RBPs have been shown to be relatively highly conserved across species, particularly for the core RNA-binding domains (RBDs) [38–40]. Consistently, the RBP-binding motifs from the RBPs containing highly conserved RBDs also show notable degree of conservation [41–43]. These observations inspire us to apply the label smoothing strategy to the prediction of cross-species RBP–RNA interactions.

In this study, we develop MuSIC (**Mu**lti-**S**pecies RBP–RNA **I**nteractions using **C**onservation), a deep learning framework for cross-species RBP–RNA interaction prediction by leveraging the multi-level conservation patterns of RBPs. We show that MuSIC can accurately predict RBP-binding peaks across species and outperforms existing state-of-the-art computational methods, particularly in the closely related species. We then apply MuSIC to systematically profile the RBP-binding peaks and motifs for 186 RBPs across 11 species. Finally, we use MuSIC to quantify the effects of SNVs on RBP binding, providing insights into the RBP-mediated regulatory mechanisms associated with human diseases.

115 ## Results

116 ### Evolutionarily conserved RBPs exhibit similar RNA-binding preference

117 Previous studies have revealed that RBDs, the core functional elements in RBPs,
118 exhibit a significantly slower rate of sequence evolution compared to genomic
119 sequence context surrounding these RBDs, indicating their critical functions in
120 regulating RNA recognition [38]. We collected protein sequences and predicted
121 structures of 216 RBPs from 11 species ranging from human to yeast, including human,
122 orangutan, monkey, mouse, rat, chicken, frog, zebrafish, fly, *A. thaliana* and yeast, to
123 systematically assess the conservation patterns of RBPs and their core RBDs (see
124 **Methods** for details). Briefly, the conservation patterns of RBPs and their core RBDs
125 was evaluated based on sequence and structure similarity, respectively (**Fig. 1a**,
126 **Supplementary Fig. 1** and **Supplementary Table 1**). As expected, both sequence
127 and structural similarities between human and non-human species declined with the
128 increasing evolutionary distance (**Fig. 1a**). In particular, the protein sequence exhibited
129 higher conservation for the core RBDs compared to the full-length RBPs; however, the
130 structural similarity for both core RBDs and full-length RBPs remained relatively high
131 across the 11 species ranging from human to yeast (**Fig. 1a** and **Supplementary Fig.**
132 **1**), indicating the fundamental importance of local structures of core RBDs for RNA
133 recognition.

134

135 We then sought to partition the 216 RBPs into different groups based on their
136 conservation patterns across the 11 species (**Supplementary Table 2**) and showed
137 cases in each group (**Fig. 1b**): (i) highly conserved (1169, accounting for 49.0%),
138 showing conservation in both sequence and structure in closely related species such
139 as human and mouse, exemplified with the case of AGO2 protein; (ii) considerable
140 conserved (769, accounting for 32.2%), showing conservation in full-length and RBD
141 structure, such as DDX51; (iii) intermediately conserved (426, accounting for 17.8%),
142 showing conservation in only RBD structure, such as RBM10; and (iv) weakly
143 conserved (23, accounting for 1.0%), such as SERBP1. We found that almost all the
144 RBPs (>99% of the total) exhibit structural conservation in their RBDs, consistent with
145 previous studies showing that RBDs of the same RBP are highly structurally conserved
146 across species [44,45].

Previous studies have shown that RBPs containing structurally conserved RBDs tend to recognize and bind their RNA target with similar sequence patterns [41,45,46]. We first compared the distribution of RBP-binding peaks from the homologous RBPs along mRNAs in human and mouse, exemplified with the case of SRRM4 and LIN28A (**Fig. 1c** and **Supplementary Fig. 2a**). As expected, we observed a high correlation of the RBP-binding peak distributions between human and mouse (SRRM4 r = 0.96, LIN28A r = 0.91, Spearman's rank correlation test). Notably, SRRM4 showed stronger preference of binding on the stop codon and a depletion around the start codon (**Fig. 1c**), suggesting the critical roles of SRRM4 in regulating alternative splicing of RNA within these regions [47]. Next, we explored the motif-level conservation of RBPs using the ATtRACT database [48], which contains experimentally validated binding motifs in multiple species. Comparative analysis of RBP-binding motifs between human and mouse revealed a high degree of conservation (e.g., UUCUGUG motif for RC3H1, GGGGGCUG motif for HNRNPF, GGAG motif for LIN28A) (**Fig. 1d**). Moreover, both the structures derived from X-ray diffraction (**Fig. 1e**) and AlphaFold3 [49] prediction (**Supplementary Fig. 2b-e**) suggest that the homologous RBPs from different species exhibit similar three-dimensional interaction patterns between the interface residues in RBPs and their recognized target binding motifs (PDB ID 4QIL in human and 4QI2 in mouse) (**Fig. 1e**). Taken together, these results suggest that RBPs show high sequence and structural conservation across species, and maintain high conservation in their RNA-binding patterns, including motifs, binding regions within transcripts, and key interacting residues.

Inspired by the aforementioned observations and previous studies[30,33,43], we aim to develop a novel computational model to systematically predict the RBP–RNA interactions in the target species (e.g., non-human species) by integrating the RBP-binding peaks from the source species (e.g., human) and the RBP similarity patterns between species. We leveraged the large-scale RBP-binding peaks from different species (human, mouse, and zebrafish) to construct the computational model. We initially found 23 homologous RBPs between human and mouse, and one homologous RBP between human and zebrafish with available datasets from POSTAR3

179 (Supplementary Fig. 2f), among which 38 datasets from 14 RBPs were curated for
180 the training and validation of the cross-species model (see Methods for details;
181 Supplementary Table 3).

182

183 **Overview of MuSIC algorithm**

184 We developed MuSIC, a deep learning framework that could integrate RBP-binding
185 peaks and cross-species conservation information of RBPs to accurately model and
186 predict RBP–RNA interactions in diverse non-human species. Briefly, the MuSIC
187 framework consists of four components (Fig. 2a): (i) pre-processing input data:
188 collecting peaks of RBPs from multiple species, and constructing training and
189 validation sets; (ii) feature engineering: incorporating nucleotide sequences and RNA
190 secondary structures to learn RBP-binding preferences and enhance the prediction
191 ability of the model (Fig. 2b); (iii) predicting RBP–RNA interactions**:** dynamic
192 optimization on the prediction results using label smoothing regularization and gradient
193 weight adaptation in a convolutional neural network (CNN); and (iv) analyzing RBP–
194 RNA interactions: cross-species motif identification and other downstream
195 applications.

196

197 We realized that the cross-species prediction tasks often have great challenges
198 in generalization, e.g., the ability to accurately predict RBP–RNA interactions in mouse
199 based on the same distribution as the training data in human [33,34]. There are several
200 disadvantages in the existing deep learning-based methods, such as overfitting
201 caused by the limitations of one-hot label loss functions, and heightened sensitivity to
202 label noise and boundary ambiguities inherent in one-hot supervision [36]. To this end,
203 we leveraged the label smoothing strategy that has been widely used in the field of
204 computer vision to facilitate the cross-style classification [35]. In MuSIC, the conservation
205 of RBPs across species can be used to enhance the model generalization and mitigate
206 overfitting in a similar way. Briefly, MuSIC employs a novel approach in applying the
207 label smoothing strategy, in which the one-hot labels were converted into soft labels
208 based on the pre-defined RBP conservation scores [37]. The cross-entropy loss was
209 computed for both label types, with the final loss obtained by weighting and summing
210 the two components (Fig. 2c). In addition, we used gradient weight adaptation

211 algorithm to dynamically adjust the weight of each loss component during the model
212 training (**Fig. 2d**). Next, the MuSIC model combines several convolutional layers,
213 residual blocks and the Squeeze-and-Excitation (SE) module to capture and enhance
214 RBP binding features learning, in which overfitting could be minimized by dropout and
215 early stopping (**Fig. 2a, e**). Finally, the learned features were integrated using the fully
216 connected layers to generate the prediction outputs (see **Methods** for details).

217

218 Existing state-of-the-art computational methods typically utilize input sequence
219 lengths of 101nt [28,30,31] or 400nt [26,50]. We found that increasing the input sequence
220 length to 200nt could significantly improve the Area Under the Curve (AUC) score in
221 cross-species prediction (**Fig. 2f** and **Supplementary Fig. 3a**). The inclusion of label
222 smoothing, and gradient adaptation could further improve the AUC score (**Fig. 2f** and
223 **Supplementary Fig. 3b**). Notably, incorporating predicted RNA structures could
224 further enhance the AUC score to 0.85 in cross-species prediction (**Fig. 2f**). We also
225 evaluated the prediction performance using the experimentally measured *in vivo* RNA
226 structures instead of the predicted RNA structures (**Supplementary Fig. 3c**). The
227 predicted RNA structures were used in the model because of the limited performance
228 gains by using the *in vivo* RNA structures, as well as very few *in vivo* RNA structure
229 datasets are available in other species. Taken together, these results reveal that
230 extending the sequence length and applying label smoothing could significantly
231 improve the performance of cross-species generalization.

232

233 We then characterized the contribution of each component of the MuSIC
234 framework to the performance of feature learning. The positive and negative samples
235 of input RNA sequences cannot be clearly separated in the original feature space
236 using t-SNE (**Supplementary Fig. 3d**). After adding the SE module and residual block
237 processing, the t-SNE clustering performance on the positive and negative samples
238 could be significantly improved, thus leading to the apparent discriminative power of
239 the final output from the fully connected layers (**Supplementary Fig. 3d**).

240

241 The classification performance of MuSIC's classifier was benchmarked against
242 diverse machine learning and deep learning classifiers using the identical input
243 features. We found that MuSIC consistently outperformed all the baseline classifiers,
244 confirming the robustness and effectiveness of MuSIC in modeling RBP binding
245 across species (**Supplementary Fig. 3e, f**).

246

247 MuSIC outperforms existing computational methods in predicting cross-
248 species RBP–RNA interactions

249 We evaluated the prediction performance of MuSIC using 262 datasets from 186
250 RBPs, benchmarked under both within-species (224 datasets from 186 RBPs) and
251 cross-species (38 datasets from 14 RBPs) scenarios (**Supplementary Table 3**). All
252 the datasets were processed using a uniform preprocessing pipeline to ensure
253 fairness in the performance evaluation (see **Methods** for details). Briefly, we
254 performed a systematic comparison of prediction accuracy with four state-of-the-art
255 computational methods, including HDRNet [31], PrismNet [30], DeepBind [28], and
256 GraphProt [26]. As expected, MuSIC showed better prediction accuracy than other
257 methods for the within-species prediction (AUC: MuSIC = 0.88, HDRNet = 0.87,
258 PrismNet = 0.86, DeepBind = 0.67, GraphProt = 0.72) (**Supplementary Fig. 4a**).

259

260 We then evaluated the prediction accuracy of MuSIC and other computational
261 methods in the cross-species prediction, where we trained the model in human and
262 validated in mouse and zebrafish (19 datasets). MuSIC outperformed other methods,
263 showing an obvious improvement of in AUC (AUC: MuSIC = 0.85, HDRNet = 0.73,
264 PrismNet = 0.73, DeepBind = 0.71, GraphProt = 0.70; **Fig. 3a**), as well as in the Area
265 Under the Precision-Recall Curve (PRAUC) and accuracy (ACC) metrics
266 (**Supplementary Fig. 4b, c**). In addition, we found that MuSIC a significant
267 performance improvement of AUC for predicting RBP–RNA interactions in the closely
268 related predictions, e.g., trained in human and validated in mouse (**Supplementary
269 Fig. 4d-f**). For example, MuSIC showed AUCs of 0.88 for YTHDC2 and 0.89 for TAF15
270 in the cross-species prediction, a significant improvement than other computational
271 methods (**Fig. 3b** and **Supplementary Fig. 5a**). Taken together, these results showed
272 that MuSIC could improve the prediction accuracy of RBP–RNA interactions than other

273 computational methods, especially under the cross-species scenario. Consistently,
274 MuSIC showed better performance in separating the positive and negative samples of
275 input RNA sequences in the feature space than other computational methods (**Fig. 3c**
276 and **Supplementary Fig. 5b**).

277

278 Finally, we compared the prediction accuracy of MuSIC and other computational
279 methods by evaluating to what degree the RBP-binding peaks from CLIP-seq
280 experiments could be rediscovered (see **Methods** for details). Briefly, the model was
281 trained based on the RBP-binding peaks in human, followed by predicting peaks
282 across all the transcripts in mouse (for 22 RBPs) and zebrafish (for 1 RBP)
283 (**Supplementary Table 4**). For each RBP, we observed that MuSIC showed
284 consistently better performance than other computational methods (**Fig. 3d**),
285 particularly on the cases of AGO2, SRRM4, and Hnrnpa1 (**Fig. 3e**). Notably, HDRNet
286 was excluded in this analysis due to its reliance on the experimental RNA structure
287 features, which were not available in the relevant mouse datasets. For all the 23 RBPs
288 analyzed here, the predicted peaks by MuSIC showed high correlation with the
289 experimentally derived ones (r = 0.73, Spearman's rank correlation test;
290 **Supplementary Fig. 6a-c**). Here, we showcased the site-level prediction performance
291 of MuSIC using SRSF3 within the local genomic region (chr7: 96,200,000–96,400,000)
292 in mouse, where 13 out of 17 experimentally derived peaks could be successfully
293 recovered by MuSIC (**Fig. 3f**). In contrast, only 4 peaks could be predicted by other
294 computational methods. These findings confirm the feasibility and effectiveness of
295 MuSIC in accurately capturing RBP–RNA interactions across species.

296

297 **MuSIC generates systematic RBP-binding predictions in diverse species**

298 The aforementioned results reveal that MuSIC could generate better cross-species
299 predictions than other computational methods, particularly when the source and target
300 species are more evolutionarily closed. It will be useful to explore to what degree the
301 source and target species are evolutionarily closed, then the MuSIC can generate
302 reliable prediction results. We then evaluated the relationship between the prediction
303 accuracy and the RBP conservation between species in terms of both their sequence
304 and structure (**Supplementary Fig. 7a**; see **Methods** for details). Notably, we

305 observed a strong positive correlation (r = 0.76, Pearson's correlation test; **Fig. 4a**)
306 between the prediction accuracy (measured by AUC) and the RBP conservation based
307 on 14 RBPs shared between mouse and zebrafish, indicating that the degree of RBP
308 conservation could be used to quantificationally threshold the prediction results of
309 RBP–RNA interactions. We then predicted the RBP–RNA interactions for the 10
310 species in a broad evolutionary range using MuSIC, and filtered low-confidence
311 prediction results based on the degree of RBP conservation equivalent to AUC of 0.8
312 (see **Methods** for details).

313 As expected, more than 90% of the RBPs showed high-confidence prediction
314 results (i.e., with prediction AUC > 0.8) in the species closely related to human,
315 including orangutan, monkey, mouse and rat (**Supplementary Fig. 7b**). In contrast,
316 only a small fraction (32.9%–68.1%) of the RBPs showed high-confidence prediction
317 results in the remaining distantly related species, such as yeast, *A. thaliana* and fly
318 (**Supplementary Fig. 7b**). Taken together, these findings confirm that MuSIC performs
319 significantly better in the closely related species, by leveraging the RBP conservation
320 patterns in terms of both sequence and structure (**Supplementary Fig. 7b** and
321 **Supplementary Table 5**).

322

323 **MuSIC improves RBP-binding motif identification in diverse species**

324 Previous studies have revealed that conserved RBPs tend to have conserved RBP-
325 binding motifs during evolution, which has been validated by some specific RBPs [41,42].
326 However, this hypothesis has not been systematically tested in a broad set of RBPs.
327 To this end, we first predicted the RBP-binding peaks across all the 11 species using
328 the trained 2,046 RBP models, and identified the enriched RBP-binding motifs in the
329 human and non-human species (**Supplementary Table 5**). The predicted enriched
330 motifs in the non-human species could be successfully recapitulated by the well-known
331 RBP binding sequence preference [48]. For example, the motif CAUCU for YBX3, GGA
332 for SRSF1 and polyC for KHSRP could be rediscovered in mouse, rat and chicken,
333 respectively (**Fig. 4b** and **Supplementary Fig. 7c**). We then generated a catalog of
334 RBP-binding motifs in all the 11 species ranging from human to yeast, and visualized
335 the motif clustering based on their similarities (**Fig. 4c**; see **Methods** for details).
336 Notably, RBP-binding motif prediction failed for certain RBPs due to their weak or

337 undetectable binding signals in the distantly related species, such as yeast, fly, and
338 zebrafish.

339

340 The inferred RBP-binding motifs with high confidence showed distinct degree of
341 conservation during evolution (**Supplementary Fig. 7d** and **8**, **Supplementary Table**
342 **5**). For example, HNRNPA1 was predicted to contain GA-rich (e.g., GAA/AG) motifs
343 in the closely related species (e.g., orangutan, monkey, and mouse) (**Supplementary**
344 **Fig. 7d**). In contrast, the UGUANA-like motifs for PUM2 could be predicted in diverse
345 species ranging from human to yeast (**Supplementary Fig. 7d**), consistent with the
346 well-known motif UGUANAUA for PUM2 across multiple species [51]. Overall, the
347 predicted RBP-binding peaks by MuSIC enables the identification of RBP-binding
348 motifs in diverse non-human species, revealing the conserved and diverged binding
349 preference of specific RBPs.

350

351 **MuSIC reveals evolutionarily conserved RBP-binding peaks**

352 It has been shown that the RBP-binding peaks from an RBP are either evolutionarily
353 conserved or species-specific [43,52]. The conservation patterns of RBP-binding peaks
354 could be influenced by both the RBP itself and the conservation of the target RNA
355 sequence. To systematically evaluate the conservation of RBP-binding across species,
356 we analyzed the predicted RBP-binding peaks of 186 RBPs across 11 species in the
357 conserved transcripts (**Supplementary Fig. 9a, b**; see **Methods** for details). As
358 expected, we observed higher proportion of conserved RBP-binding peaks in the more
359 closely related species (**Fig. 5a**). Almost 64% of the total RBPs analyzed here
360 exhibited considerable global patterns of RBP-binding peaks conservation (**Fig. 5a**).
361 We highlighted two RBPs, RBM15B and ALKBH1, with their RBP-binding peaks from
362 9 species on the evolutionarily conserved genomic region (**Fig. 5b**). Notably, there
363 were fewer or none evolutionarily conserved RBP-binding peaks in the distantly
364 related species.

365

366 We further investigated the conservation of RBP-binding peaks by analyzing the
367 RBPs containing the same protein family from different species. The YTH family RBPs,

368 which contain YTH protein domains, exhibited consistently strong RBP-binding peaks
369 conservation across species (**Supplementary Fig. 9c**). In contrast, the CPSF family
370 RBPs showed lower RBP-binding peaks conservation in the distantly related species
371 (**Supplementary Fig. 9c**), consistent with previous studies reporting gradient
372 conservation of the RBP-binding peaks [43]. Further exploration of the relationship
373 between RBP-binding peaks conservation and species evolution revealed that, while
374 CPSF family RBPs maintained the steady degree of RBP-binding peaks conservation
375 of in mammals, an obviously decrease showed among the distantly related species
376 (i.e., from frog; **Fig. 5c**). Similar patterns could be also observed in the YTH family
377 RBPs (**Supplementary Fig. 9d**). Taken together, these results indicate that the RNA-
378 binding domains may play critical roles in maintaining the conserved RBP-binding
379 peaks across species.

380

381 **MuSIC prioritizes genetic variation with consistent effects influencing RBP-**
382 **binding across species**

383 Genetic variations, such as SNVs, can potentially alter RBP-binding affinity by
384 disrupting RNA target recognition, and thus cause functional abnormalities and severe
385 diseases [8,53,54]. The ability of predicting RBP-binding peaks across species by MuSIC
386 enables us to quantify and analyze the effects of SNVs on RBP binding in the
387 homologous regions between human and non-human species, which will be valuable
388 for leveraging model organisms such as mouse to interpret the RBP-mediated
389 mechanisms of human disease-associated SNVs [10,55].

390

391 We first quantified and compared the binding affinity on the local regions with (i.e.,
392 alternative allele) and without (i.e., reference allele) SNVs using the eCLIP data [56] from
393 103 RBPs (**Supplementary Fig. 10a**; see **Methods** for details). We found that SNVs
394 detected in CLIP data were relevant to modulating RBP-binding affinity
395 (**Supplementary Table 6**). Inspired by these findings, we further evaluated the
396 consistency between MuSIC predictions and experimental data for SNV effects. We
397 found that MuSIC could accurately predict the RBP-binding affinity on the local region
398 containing the reference and alternative alleles, which was exemplified with the cases

399 of SRSF1 and DDX3X (**Fig. 6a, b** and **Supplementary Fig. 10b**; see **Methods** for
400 details).

401

402 Next, we ask whether MuSIC could be applied in quantifying the effects of SNVs
403 in the homologous regions between species (**Fig. 6c**; see **Methods** for details). By
404 applying the analysis framework, we systematically screened the 32,207 synonymous
405 SNVs in human genome, corresponding to 85,599 SNVs in the homologous regions
406 in mouse genome. We predicted 59,811 SNVs in mouse genome that could potentially
407 disrupt RBP binding (i.e., ΔSNV < -0.5) (**Fig. 6c**). Here, we showed an example: the
408 SNV NC_000006.11: g.4059022A>G in the human transcript NM_003913.5 was
409 predicted as disrupting PUM2 binding (ΔSNV = -0.898) (**Fig. 6d**), consistent with the
410 predicted effect of the SNV located in the homologous mouse transcript
411 ENSMUST00000220965.2 (ΔSNV = -0.805) (**Fig. 6e**). Furthermore, we performed
412 RNA pull-down *in vitro* and POND-qPCR [57] *in vivo* assays to validate the predictions
413 by MuSIC. The POND (**P**rotein nanocage-emp**O**wered **N**on-**D**estructive) method is a
414 recently developed novel technique for *in vivo* detection of RBP–RNA interactions,
415 which enables selective packaging of RBP–RNA complexes into extracellular
416 nanovesicles [57] (**Supplementary Fig. 11a**). We focused on three SNVs predicted by
417 MuSIC showing disrupted binding affinity by PUM2 in both human and mouse
418 (**Supplementary Table 7**). The RNA pull-down *in vitro* results confirmed that the
419 alternative allele could indeed significantly impact PUM2 binding on the target RNA
420 sequences in human and mouse (**Fig. 6f** and **Supplementary Fig. 11b, c**). In addition,
421 the POND-qPCR *in vivo* results also supported these observations in human and
422 mouse (**Fig. 6g**), confirming that MuSIC can generate accurate and reliable prediction
423 on the effects of homologous SNVs on RBP binding across species.

424

425 **Pervasive influenced RBP-binding implicated in disease mechanisms**

426 Among all the 186 RBPs analyzed for their binding affinity changes on the SNVs
427 in mouse genome, 79 (42.5%) RBPs were found to be enriched with the biological
428 processes associated with ubiquitination regulation and protein degradation (**Fig. 6h**
429 and **Supplementary Fig. 12a, b**). Previous studies have revealed that ubiquitination-
430 related processes are implicated in the pathogenesis of complex diseases, such as

431 cancer [58–61] and neurodegenerative diseases [62–65], which can be also validated by the
432 predictions by MuSIC (**Supplementary Fig. 12c**). These findings suggest the critical
433 roles of SNVs in modulating RBP binding and influencing the ubiquitination-associated
434 processes, which may be implicated in complex disease mechanisms through
435 disrupting protein homeostasis.

436

437 In our analysis, TARDBP was prioritized as highly related to the ubiquitination-
438 related processes. TARDBP has been reported to be critical in regulating protein
439 homeostasis and protein folding, and abnormal TARDBP binding on the target RNA
440 transcripts can lead to several diseases in human and mouse [66–70]. We confirmed that
441 the SNV in mouse could disrupt TARDBP binding to *Usp25*, a gene involved in
442 ubiquitin metabolism (**Supplementary Fig. 13a**). Moreover, the TARDBP-bound
443 genes perturbed by SNVs were significantly enriched in the functions related to neural
444 development, ubiquitination, deubiquitination, and protein metabolism
445 (**Supplementary Fig. 13b, c**).

446

447 ## Discussion

448 Existing computational methods for predicting RBP–RNA interactions [26,28,30,31] were
449 generally developed for the within-species tasks, having difficulties in predicting cross-
450 species. In this study, we developed MuSIC, a novel computational method that can
451 accurately predict RBP-binding peaks in the species lacking available CLIP datasets
452 by leveraging the multi-level conservation patterns of the RBP. Compared to other
453 similar computational methods, MuSIC showed superior performance in predicting
454 RBP-binding peaks on the cross-species tasks. We then systematically predicted
455 RBP-binding peaks and their enriched sequence motifs for 186 RBPs in human and
456 10 non-human species using MuSIC. We confirmed that the conservation patterns of
457 the predicted RBP-binding peaks were consistent with previous studies [42,43,52]. Finally,
458 we applied MuSIC to investigate the effects of SNVs on RBP binding, particularly
459 focusing on the imbalance in RBP binding induced by SNVs and exploring their
460 potential connections to diseases such as cancer and neurodegenerative disorders.

461

The following improvements in the model architecture are critical for the augmented prediction performance of MuSIC in the cross-species tasks: optimization of input feature length and label smoothing. Previous studies have showed that increasing input sequence length can dramatically improve the model performance [24,71]. In our study, we revealed that increasing the input sequence length from 101nt to 200nt could significantly improve the prediction performance across species (AUC from 0.55 to 0.75). In contrast, increasing the input sequence length from 200nt to 400nt would lead to decreased model performance. Increasing the input sequence length can not only provide more useful sequence information, but also may include noisy and redundant sequence information for modeling RBP binding. Inspired by the concepts from computer vision, we implemented label smoothing in the model to further improve prediction performance [35–37]. We showed that label smoothing and the gradient weight adaptation mechanism could significantly improve cross-species predictions. This strategy reflects the inherently dynamic nature of biomolecular recognition, which operates on a continuum rather than a binary decision framework [72].

The MuSIC model exhibited lower prediction confidence in the distantly related species, such as fly, *A. thaliana* and yeast, partially due to the decreased evolutionary conservation of RBPs in these species. With the advancement in large language models, RBP-guided RNA generation strategies could increase the prediction confidence of RBP-binding peaks in distantly related species [73]. For example, BERT-RBP [74] was developed by adapting the pretrained BERT architecture and leveraging long-range RNA sequence features in predicting RNA-RBP interactions with improved performance. Similarly, ZHMolGraph [75] was proposed to predict novel RBP–RNA interactions by combining graph neural networks with unsupervised large language models. In the future, we hope that incorporating large language model-based diffusion methods into the MuSIC model can improve the prediction performance by complementing the low conservation of RBPs in the distantly related species.

Previous studies revealed that genetic variation in RNA can potentially reduce RBP-binding affinity [8,53,76]. However, the effects of SNVs on RBP binding has not fully

explored, particularly in the non-model organisms. In our study, we quantified the effects of SNVs on RBP binding in the homologous regions between human and mouse, and furthermore experimentally validated these prediction results *in vitro* and *in vivo*. Our results revealed the evolutionary conservation of genetic variation influencing RBP binding across species, and found that the SNVs that could potentially alter RBP–RNA interactions are associated with human diseases [9,10]. We then performed a comprehensive analysis on the homologous human SNVs in mouse to quantify their binding affinity changes for each of the 186 RBPs. Interestingly, the disrupted binding events generally are recognized by TARDBP, and the target transcripts are enriched with the ubiquitination-associated pathways [31]. The dysregulation of ubiquitin–proteasome system (UPS) can lead to protein misfolding and accumulation, which has been linked to human diseases including cancer and neurodegenerative disorders [65,77].

In summary, we propose MuSIC as a model to predict RBP-binding peaks across species. MuSIC outperformed state-of-the-art prediction models, and generated a catalog of predicted RBP-binding peaks of 186 RBPs with high accuracy in 11 species. The detailed analysis quantified the effects of genetic variants on RBP binding, particularly highlighted in the homologous regions between human and mouse. MuSIC provides insights into the RNA-RBP interactions in non-model organisms, and generates novel hypotheses on the interplays between RBP-mediated regulation and disease mechanisms.

Acknowledgements

We thank Chenqian Wang, Shaozhen Yin, Ruobin Zhao, Liangyu Li, Yongkang Tang and Suiru Lu for fruitful discussions and invaluable feedback. We thank Jindong Sun for running the software comparison benchmark. This work was supported by the National Natural Science Foundation of China (No.82341086, No.32300521, and No.32422013 to L.S., No.32025007 to Y.W.); the Open Grant from the Pingyuan Laboratory (No.2023PY-OP-0104 to L.S.); the State Key Laboratory of Microbial Technology Open Projects Fund (No.M2023-20 to L.S.); the Intramural Joint Program Fund of the State Key Laboratory of Microbial Technology (NO.SKLMTIJP-2024-02 to L.S., and T.Z.); the Shandong Province Postdoctoral Innovation Project (NO. SDCX-ZG-202400146 to T.Z.), the Qingdao Postdoctoral Science Foundation (NO. QDBSH20240102200 to T.Z.), the Double-First Class Initiative of Shandong University School of Life Sciences; the Young Innovation Team of Shandong Higher Education Institutions, the Taishan Scholars Youth Expert Program of Shandong Province, and the Program of Shandong University Qilu Young Scholars.

Author's Contribution

L.S. conceived the project. L.S., Y.T.Y. and Y.W. supervised the project. J.H., and T.Z. established the initial technical framework. J.H. performed the computational analyses, data curation, and methods development. J.H., Q.C. and S.Y. trained the deep learning models. L.H. and Y.W. conducted *in vivo* and *in vitro* experimental design and validation. Y.J., T.Z. and S.J. collected and re-analyzed the RBP similarity data. J.H. and J.W. collected and re-analyzed RBP-binding peaks data based on CLIP-seq. The initial draft of the manuscript was written by J.H. and T.Z., with Y.T.Y., L.H. and Q.C. providing critical feedback and suggestions for improvement. J.H. formatted the figures, and the final version of the manuscript was prepared by L.S. and Y.T.Y.

Availability of data and materials

An implementation of the MuSIC script and benchmarking data is available at https://github.com/GALE1228/MuSIC.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

559 **Methods**

560 **Compiling RBP-binding peak datasets**

561 We obtained the RBP-binding peaks of 216 RBPs from POSTAR3 [32] for the within-
562 and cross-species model training and testing. We applied a uniform preprocessing
563 pipeline to the peak datasets to minimize putative bias in the datasets. Briefly, we first
564 removed the duplicated and short (length < 5nt) peaks in each dataset. Then, we
565 selected the top 5000 peaks with the highest binding strength in each dataset for the
566 downstream analysis. Next, the peaks were extended by 200nt on both sides. We
567 used GFF Annotation Parser (https://github.com/lipan6461188/GAP) to extract the
568 RNA sequences of these extended peaks, which were considered as the positive
569 peaks (i.e., RBP-binding peaks) in the model training and validation. We randomly
570 selected 5000 RNA sequences of 200nt in length from the remaining genomics regions
571 as the negative ones.

572

573 For the within-species prediction, 224 datasets of 186 RBPs from human, mouse,
574 and zebrafish were used. For the cross-species prediction, 17 datasets of 13 RBPs
575 shared between human and mouse, and 2 datasets of one RBP shared between
576 human and zebrafish were used. Notably, the datasets of the same RBP generated by
577 different CLIP-seq technologies were excluded in the cross-species analysis. For the
578 within-species prediction, we used randomly 80% of the positive and negative peaks
579 for the model training, whereas 20% were used for the model validation. For the cross-
580 species prediction, we simply maintained a training-to-validation ratio of 4:1 by
581 randomly selecting the positive and negative peaks from the source and target species.

582

583 **Reconstructing phylogenetic tree among the 11 species**

584 We aim to develop a novel computational method to systematically predict RBP–RNA
585 interactions in 11 species, including *Homo sapiens* (human), *Pongo*
586 *abelii* (orangutan), *Macaca fascicularis* (monkey), *Mus musculus*(mouse), *Rattus*
587 *norvegicus* (rat), *Gallus gallus* (chicken), *Xenopus laevis* (frog), *Danio*
588 *rerio* (zebrafish), *Drosophila melanogaster* (fly), *Arabidopsis thaliana* (*A. thaliana*),
589 and *Saccharomyces cerevisiae* (yeast).

590

591 We retrieved the 18S rRNA sequences of the 11 species from RNAcentral [78]
592 (Release 24) for evaluating the evolutionary distance among these species. We then
593 performed multiple sequence alignment of the 18S rRNA sequences using the
594 Sequence Alignment tool in MEGA [79] v11.0.13, followed by constructing a phylogenetic
595 tree among the 11 species using the neighbor-joining method. We then evaluated the
596 reliability of the reconstructed phylogenetic tree using the Bootstrap test with 1000
597 replicates.

598

599 **Compiling sequence and structural datasets of RBPs**

600 We collected the protein IDs of 216 RBPs [32] that were used in this study in the 11
601 species from UniProt [51] (**Supplementary Table 1**). For each RBP, the corresponding
602 sequences and structural data were retrieved from UniProt [51]. In the cases where
603 predicted structural data for certain RBPs were unavailable in UniProt, we used
604 AlphaFold2 [80] to predict and supplement the missing ones. The sequence similarity of
605 the full-length RBPs between human and other species was computed using ClustalX
606 [81] v2.1.1. We used US-align [82] to evaluate the structural similarity of the full-length
607 RBPs by calculating two metrics: TM-score (overall structure similarity) and pLDDT
608 (confidence of predicted structure).

609

610 We obtained the annotations of 130 RBDs in the 11 species from InterPro [83]
611 (Release 103.0), including their sequence positions and respective fragments in their
612 full-length RBPs. The local structures of RBDs were extracted from the structures of
613 full-length RBPs, ensuring the local structures of RBDs were aligned with the
614 structures of full-length RBPs. Finally, ClustalX [81] v2.1.1 and US-align [82] were used to
615 calculate the sequence and structural similarity of the RBDs between human and other
616 species.

617

618 **Comparing RBP–RNA interactions in three-dimensional structures between**
619 **species**

We collected three-dimensional structures of homologous RBP RC3H1 (PDB ID 4QIL in human, PDB ID 4QI2 in mouse) and their bound RNA fragments from the PDB [84]. In addition, we used AlphaFold3 [49] webserver to predict the interactions between homologous RBP IGF2BP1 and bound RNA fragments in human and mouse based on the IGF2BP1 protein sequence in human and mouse, as well as well-known RNA sequence with strong preference by IGF2BP1 (5'-UGACUUACUUCAUUAUUGAA-3'). The cartoon and surface diagrams displaying the bound RNA, RBDs, and the residues involved in the interactions were generated using PyMOL v3.1.0 (http://www.pymol.org/pymol).

Pre-processing MuSIC input data

The RNA sequence of each peak in the datasets was first transformed into a one-hot encoded matrix of size (4, 200), where the four rows correspond to the nucleotides A, C, G, and U. The secondary structure of each RNA sequence was predicted using RNAfold [85] v2.1.0 with default settings. The resulting structure was encoded into a 2-dimensional matrix of size (2, 200), with 'P' and 'U' representing the paired and unpaired bases, respectively. The sequence and structure matrices were then concatenated along the channel axis to generate a unified (6, 200) representation encoding both sequence and structural features, which will be used as the input data for the MuSIC model.

Overview of MuSIC architecture

The input data for MuSIC is defined as *S*, which contains *N* peaks. Each sample, $s \in S$, is an RNA sequence of length *L* = 200nt. As described above, *s* corresponds to a sequence-and-structure vector of D = 6 dimensions. Therefore, *S* is encoded into a tensor $X \in R^{N \times \mathrm{D} \times \mathrm{L}}$ and the binding probability $Y \in R^{N}$ is computed by as follows:

$$Y = \mathrm{MuSIC}(X)$$

MuSIC is defined as follows:

$$\mathrm{MuSIC}(X) = \sigma(\mathrm{FC}(f_R(f_S(f_C(\mathrm{X}))))))$$

where σ is the sigmoid activation function, FC is the fully connected layer, f_R is the residual blocks, f_S is the squeeze-excitation block (SE), and f_C is the convolutional block, respectively.

Convolutional Block

First, we use a convolutional layer to extract features. The convolutional block $f_C(X)$ is defined as follows:

$$f_C(X) = \text{ReLU}\left(\text{BN}\left(\text{Conv}_{K,P}(X)\right)\right)$$

where $\text{Conv}_{K,P}(X)$ is a 2D convolutional layer with kernel K and padding P. The convolutional layer is designed to extract local sequence and structural features from the input data.

Squeeze-and-Excitation (SE) Block

We apply a Squeeze-and-Excitation (SE) block to recalibrate channel-wise feature responses. The SE block is defined as follows:

$$f_S(X) = X \otimes \text{SE}(X)$$

$$\text{SE}(X) = \sigma\left(f_{\text{ex}}\left(f_{sq}(X)\right)\right)$$

where σ is the activation function, f_{sq} is a global average pooling function to squeezes the global sequence and structural features into a channel statistic and f_{ex} applies a nonlinear transformation to the squeezed feature map. The SE block helps the model focus on the most relevant features for RBP binding.

Residual Blocks

We then use residual blocks to capture both sequence and structural features over long ranges. These residual blocks are designed to learn the hierarchical features by passing information through several convolutional layers. Two types of residual blocks are used here:

ResidualBlock2D, which operates on 2D feature maps, is defined as follows:

$$f_R(X) = \text{ResidualBlock2D}\left(\text{Conv}\left(\text{SE}(X)\right)\right)$$

ResidualBlock1D, which operates on 1D feature maps, is defined as follows:

$$f_R(X) = \text{ResidualBlock1D}\left(\text{AvgPool}\left(\text{ResidualBlock2D}(X)\right)\right)$$

Fully Connected Layer

Finally, we apply a fully connected layer to generate the binding probability prediction for each RNA input sequence, which is defined as follows:

$$Y = \sigma\left(f_{\text{FC}}(x)\right)$$

Label smoothing strategy and gradient weight adaptation

We applied the label smoothing strategy to smooth the one-hot label distribution. Given a one-hot label vector $q(x)$, the smoothed label distribution $q'(x)$ is defined as follows:

$$q'(x) = \alpha q(x) + (1 - \alpha)u(x)$$

where α is the RBP conservation score as a smoothing coefficient, and $u(x)$ is the uniform distribution.

The one-hot label distribution $q(x)$ and a predicted binding probability distribution $p(x)$ are used for computing the classification loss as follows:

$$L_{ture} = \mathcal{H}\left(q(x), p(x)\right) = -\sum_{i=1}^{C} q\,(x_i) \log\left(p(x_i)\right)$$

The smoothed label distribution $q'(x)$ and a predicted binding probability distribution $p(x)$ are used for computing the smoothing loss as follows:

$$L_{smooth} = \mathcal{H}\left(q'(x), p(x)\right) = -\sum_{i=1}^{C} q\,'(x_i) \log\left(p(x_i)\right)$$

$$L_{smooth} = \mathcal{H}\left(q'(x), p(x)\right) = -\sum_{i=1}^{C} \alpha q(x_i) \log\left(p(x_i)\right) - \sum_{i=1}^{C} (1 - \alpha)u(x_i) \log\left(p(x_i)\right)$$

The first term, L_{true}, represents the typical classification loss, while the second term, L_{smooth}, is the smoothing loss. Finally, the total loss is defined as the weighted combination of classification loss and smoothing loss as follows:

$$L_{total} = \alpha L_{ture} + \beta L_{smooth}$$

We proposed gradient weight adaptation to address the conflict between the gradient directions of L_{ture} and L_{smooth}. First, the gradients of both losses are computed with respect to the feature encoder parameters and classifier parameters as follows:

$$\nabla L_{ture(\text{encoder})} = \frac{\partial L_{ture}}{\partial \theta_{\text{encoder}}}, \quad \nabla L_{smooth(\text{encoder})} = \frac{\partial L_{smooth}}{\partial \theta_{\text{encoder}}}$$

$$\nabla L_{ture(\text{classifier})} = \frac{\partial L_{ture}}{\partial \theta_{\text{classifier}}}, \quad \nabla L_{smooth(\text{classifier})} = \frac{\partial L_{smooth}}{\partial \theta_{\text{classifier}}}$$

The weight of each loss term can be dynamically adjusted based on the ratio of the current loss to the previous one as follows:

$$\Delta_{ture} = \frac{L_{ture}}{L_{ture}^{t-1}}, \quad \Delta_{smooth} = \frac{L_{smooth}}{L_{smooth}^{t-1}}$$

$$w_{smooth} = \frac{1}{\Delta_{smooth}}, \quad w_{ture} = \frac{1}{\Delta_{\text{true}}}$$

$$w_{smooth} = \frac{w_{smooth}}{w_{smooth} + w_{ture}}, \quad w_{ture} = \frac{w_{\text{cls}}}{w_{smooth} + w_{ture}}$$

Subsequently, the gradient can be expressed as follows:

$$\nabla L_{\text{total(encoder)}} = w_{smooth} \nabla L_{smooth(\text{encoder})} + w_{ture} \nabla L_{ture(\text{encoder})}$$

$$\nabla L_{\text{total(classifier)}} = w_{smooth} \nabla L_{smooth(\text{classifier})} + w_{ture} \nabla L_{ture(\text{classifier})}$$

By dynamically adjusting w_{ture} and w_{smooth}, the model can effectively prioritize the loss terms that show significant changes in the optimization process

MuSIC model training

We applied the label smoothing strategy to smooth the one-hot label distribution and calculated the total loss as described above. The gradient weight adaptation dynamically modulated relative weights between L_{ture} and L_{smooth} according to the ratio of the current loss to the previous one, ensuring efficient gradient propagation.

The parameters in MuSIC model were optimized by minimizing the total loss—this loss integrates the classification loss and the smoothed label loss. We used Adam to optimize the model parameters, and Gradual Warmup Scheduler to adjust the learning rates. Throughout model training, we periodically evaluated the prediction performance of the model on the validation set to find the optimized parameters for the model.

MuSIC hyperparameter settings

The hyperparameters of MuSIC were empirically adjusted based on the cross-species datasets. The hyperparameter settings for MuSIC are defined as follows:

Convolutional layer: the convolutional layer uses a kernel size of (3, 3) with 16 output channels, followed by Batch Normalization (BN) and ReLU activation for non-linearity.

Pooling layer: average pooling is applied with a kernel size of (6, 1), which reduces the feature map along the first axis of the input tensor.

Full connection layer: the fully connected layer at the end of the network has 1 output unit, which corresponds to the predicted binding probability for each sample.

Dropout probability: dropout is applied with probabilities of 0.1, 0.5, and 0.3 at different layers in the network to prevent overfitting.

L2 norm penalty: the L2 norm penalty (weight decay) is set to 1×10^{-6}, which applies regularization to the model's weights during optimization.

Batch size: the batch size is set to 64 for training and evaluation.

Learning rate: the learning rate is set to 0.001 for the Adam optimizer, which is used to optimize the network's parameters.

Positive weight in loss function: the positive weight in the loss function is set to 2, which accounts for class imbalance in the dataset, giving more importance to positive samples during training.

Training epochs: the number of training epochs is set to 200, which determines how many times the model will iterate over the entire training dataset.

Early stop: early stopping is applied with a patience of 20 epochs, meaning training will halt if the model's performance does not improve on the validation set for 20 consecutive epochs.

Initial weights: the initial weights of the model are set using the Kaiming normal initialization for convolutional layers and Xavier initialization for fully connected layers.

Ablation experiments of MuSIC model

We performed the ablation experiments to explore to what degree the features could enhance the model performance: (i) input sequence length: we tried different input sequence lengths of 101nt, 200nt and 400nt; (ii) label smoothing strategy: we added and removed the label smoothing strategy in the models; (iii) sequence and structural features: we constructed the models using different types of input data, including only sequences, only predicted structures, combined sequences and predicted structures, and combined sequences and experimentally-measured structures.

Performance comparison with other computational methods

We compared the prediction performance of MuSIC with other recently published computational methods, including HDRNet (https://github.com/zhuhr213/HDRNet) [31],

PrismNet (https://github.com/kuixu/PrismNet) [30], DeepBind (https://github.com/jisraeli/DeepBind) [28], and GraphProt (https://github.com/dmaticzka/GraphProt) [26]. Notably, to make a fair comparison, each of these computational methods was trained and validated on the same datasets as used in MuSIC, using their default parameter settings.

We extracted the features from the fully connected layers of MuSIC and the other four computational methods, including HDRNet [31], PrismNet [30], DeepBind [28], and GraphProt [26]. These features were projected into a two-dimensional space using t-SNE to visualize and compare their classification performance.

We used MuSIC and the other four computational methods to train cross-species models and predict RBP-binding peaks for 22 RBPs from mouse and one RBP from zebrafish. We then estimated the prediction performance of the computational methods based on the overlap between the predicted peaks with binding probabilities above 0.9 and the experimentally-derived peaks. We also calculated the correlation between the predicted peaks and experimentally-derived peaks as we did in the meta-transcript analysis mentioned above.

Thresholding high-confidence RBP across species predicted by MuSIC

We considered the six RBP features for RBP conservation analysis, including the full-length sequence similarity, the full-length structural TM-score, the full-length structural pLDDT, the RBD sequence similarity, the RBD structural TM-score, and the RBD structural pLDDT. We then applied Principal Component Analysis (PCA) to calculate the first principal component (PC1) as RBP conservation.

We then analyzed the correlation between the RBP conservation scores and the AUC scores of 14 RBPs derived from the cross-species analysis. We observed a linear correlation and constructed a regression model between the two variables, which can be used for estimating the cross-species prediction accuracy (i.e., AUC score) of the

807 peaks (Supplementary Table 4). We applied a threshold of 0.8 for the AUC scores to
808 retain the high-confidence predicted peaks.

809

810 **Identifying and clustering RBP-binding motifs**

811 A total of 2046 models for predicting RBP–RNA interactions in 11 species were trained
812 using 186 high-quality peak datasets from human [32]. RNA sequences from each
813 species were segmented into the fragments of 200nt in length, and 20,000 fragments
814 were randomly sampled for further analysis. In each species, each of the MuSIC
815 models of the 186 RBPs generated predicted peaks on these fragments. We then
816 applied MEME [86] v5.5.7 to identify the enriched RBP-binding motifs using the
817 fragments with binding probabilities above 0.9.

818

819 We performed motif clustering in two steps. First, for each motif, we calculated
820 base frequency by averaging the nucleotide frequencies (A, C, G, U) from its position
821 weight matrix (PWM) as follows:

824 $$\text{Base Frequencies}_i = (A_i, C_i, G_i, U_i)$$

822 where A_i, C_i, G_i, U_i are the average frequencies of adenine (A), cytosine (C), guanine
823 (G), and uracil (U) for the motif i, respectively.

825

826 Second, we calculated the similarity matrix by integrating the Jensen–Shannon
827 (JS) divergence of PWM distributions and the Euclidean distance between base
828 frequency vectors:

831 $$S_{ij} = \alpha \cdot JS(i,j) + (1-\alpha) \cdot d_{\text{euclid}}(i,j)$$

832 $$JS(i,j) = \frac{1}{8}\sum_{k=1}^{8} \text{Jensen-Shannon}\left(\text{PWM}_i[k], \text{PWM}_j[k]\right)$$

833 $$d_{\text{euclid}}(i,j) = \sqrt{\left(A_i - A_j\right)^2 + \left(C_i - C_j\right)^2 + \left(G_i - G_j\right)^2 + \left(U_i - U_j\right)^2}$$

829 where $\text{PWM}_i[k]$ is the k-th row of the PWM matrix for motif i, $\text{PWM}_j[k]$ is the k-th row
830 of the PWM matrix for motif i, $d_{\text{euclid}}(i,j)$ is the Euclidean distance between the base

834 frequency vectors of motif i and j, α is a hyperparameter balancing the weight of the
835 JS divergence and Euclidean distance.

836

837 Finally, we grouped the motifs using hierarchical clustering to show their overall
838 similarity. We used the itol.toolkit package [87] v1.1.10 to visualize and map the motif
839 logo to the clustering results.

840

841 **Analyzing conservation patterns of RBP-binding peaks predicted by MuSIC**

842 Using human transcripts as the reference, we applied the Mashmap [88] v3.1.3 to
843 compare pairwise species (i.e., human and non-human species) and select the highest
844 conserved transcripts with segment length > 500nt. For each species, MuSIC was
845 used to predict RBP-binding peaks with 200nt and high attention regions [30] with 40nt
846 for 186 RBPs in these conserved transcripts.

847

848 The conserved transcripts from 11 species were subjected to multi-sequence
849 alignment by using Muscle5 [89] v5.1 with default parameters, and high attention region
850 coordinates were aligned accordingly. RBP-binding peaks overlap counts across
851 species were statistically analyzed within an extended window of 200nt. We then used
852 IGV [90] v2.16.2 to visualize the RBP-binding peaks in the conserved transcripts from
853 11 species.

854

855 **Evaluating consistency between MuSIC predictions and experimental data for**
856 **SNV effects**

857 We downloaded the eCLIP peaks of 103 RBPs in HepG2 cell line from ENCODE
858 (https://www.encodeproject.org/publication-data/ENCSR456FVU/). The SNVs in the
859 RBP-binding peaks of each RBP were then identified from the eCLIP mapping files
860 (i.e., BAM files) using the BCFtools mpileup [91] v1.20, where the replicates were
861 merged. Peaks regions were filtered (p.adj ≤ 0.05) and merged across three replicates
862 to retain high-confidence RBP-binding peaks. Using BEDTools [92] v2.27.1 , SNV
863 positions were overlapped with peak regions to classify peaks into SNV-overlapping

and non-overlapping groups. A paired t-test was then applied to compare signal differences between the two groups.

We analyzed the RNA-seq alignment files from the smartSHAPE datasets (GSE145805) to estimate the background allele ratio. For both eCLIP and RNA-seq datasets, we calculated their reference allele (REF) and alternative allele (ALT) ratios using BCFtools mpileup [91] v1.20, and then obtained $diff_{ratio}$, the difference between the REF and ALT ratios from the RNA-seq and eCLIP datasets as follows:

$$diff_{ratio} = \frac{eCLIP_{ALTratio}}{RNAseq_{ALTratio}} - \frac{eCLIP_{REFratio}}{RNAseq_{REFratio}}$$

In parallel, we also calculated $diff_{score}$, the predicted difference between the REF and ALT ratios as follows:

$$diff_{score} = Pred_{ALT} - Pred_{REF}$$

where $Pred_{ALT}$ and $Pred_{REF}$ are predicted binding possibilities for the sequences containing the ALT and REF, respectively.

We pre-processed the SNV effects by retaining $|diff_{ratio}| \geqslant 0.2$ and $|diff_{score}| \geqslant 0.2$, and then extracting the top 80% ranked by $diff_{score}$. We used the experimental labels ($label_{exp}$) and the predicted labels ($label_{pre}$) to evaluate the consistency between experimental data and MuSIC predictions for SNV effects. The $label_{exp}$ and $label_{pre}$ are defined as follows:

$$label_{exp} = 1 \text{ when } diff_{ratio} \geqslant 0;\ label_{exp} = 0 \text{ when } diff_{ratio} < 0$$

$$label_{pre} = 1 \text{ when } diff_{score} \geqslant 0;\ label_{pre} = 0 \text{ when } diff_{score} < 0$$

The consistency between $label_{exp}$ and $label_{pre}$ was evaluated using Fisher's exact test, with multiple testing corrected by the Benjamini–Hochberg method, and results with FDR ≤ 0.05 were considered statistical significance.

891 **Quantifying the effects of homologous SNVs between human and mouse on**
892 **RBP-binding using MuSIC**

893 We downloaded the human and mouse RNA sequences from Ensembl [93] (Release
894 113). We extracted the human RNA sequences overlapping with the nonsense SNV
895 positions from dbSNP [6] using the GAP (https://github.com/lipan6461188/GAP), and
896 then generated the two sequences containing the REF and ALT for each SNV. We
897 then identified the homologous regions of the human SNV-containing sequences in
898 the mouse genome using BLAST [94] v2.16.0. In total, 32,207 nonsense SNVs in human,
899 and 85,599 homologous nonsense SNVs in mouse were included for further analyses.

900

901 We then used MuSIC to predict the binding possibilities on the RNA sequences
902 containing the REF and ALT in human and mouse for each of the 186 RBPs. For each
903 sequence, the effect of the SNV on RBP binding was defined as follows:

904 $$\Delta SNV = P_{ALT} - P_{REF}$$

905 where P_{REF} is the binding probability for the REF, P_{ALT} is the binding probability for the
906 ALT.

907

908 **Cell transfection**

909 150,000 HEK293T cells per well were seeded in a well of poly-D-lysine-coated 24-well
910 plates. After 16-20 h, cells were co-transfected with 200 ng EPN24-PUM2 plus 75ng
911 GFP-REF or GFP-SNV target plasmid using JetPRIME reagent (Polyplus), the
912 medium was refreshed at 4-6 h post-transfection, and the supernatant was collected
913 for subsequent analysis 24 h after medium change.

914

915 **RNA pull-down assay**

916 The *in vitro* RNA pull-down assay was performed as described [95]. Briefly, 200 pmol
917 biotinylated RNA oligonucleotides were refolded by heating at 90 °C for 1 min, and
918 then incubated at 30 °C for 5 min. About 1 × 10^7 myc-PUM2-expressing HEK293T cells
919 were lysed in 500 µl lysis buffer (150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.5 mM

920 DTT, 50 mM Tris-HCl, pH 7.5, 0.5% sodium deoxycholate) with 5 μl PMSF (100mM,
921 Beyotime), 10 μl phosphatase inhibitor cocktail (50 ×, Beyotime), 10 μl protease
922 inhibitor cocktail (50 ×, Beyotime), 2.5 μl SUPERase In inhibitor (20U/μl Invitrogen)
923 and 5 μl RNase Inhibitor (40U/μl, Beyotime). Refolded RNA (8 μl) and 500 μl cell lysate
924 were incubated at 4 °C for 3-4 h. Then 100 μl pre-washed VAHTS CA-28 Streptavidin
925 Beads (Vazyme) were added into the buffer and incubated at 4 °C for 1h. The beads
926 were washed first with high salt buffer (50 mM Tris-HCl, pH 7.5, 1 M NaCl, 1% TRITON
927 X-100) at room temperature for 5 min and then 2 more times with low salt buffer
928 (50 mM Tris-HCl, pH 7.5, 150mM NaCl, 1% TRITON X-100), at room temperature and
929 for 5 min. Proteins were eluted in 50 μl 1× SDS-PAGE loading buffer (Beyotime) at
930 98 °C for 20 min.

931

932 Protein electrophoresis was performed using BeyoGel™ Elite Precast PAGE Gel
933 (8-16%, Beyotime), followed by protein transfer onto polyvinylidene fluoride (PVDF)
934 membrane. The Myc Tag Mouse Monoclonal Antibody (HRP Conjugated) (Beyotime,
935 AF2867) were diluted (1:1000) and incubated with the membrane at 4°C overnight.
936 After three washes with 1×TBST buffer, the membrane was treated with
937 electrochemiluminescence (ECL) reagent (Merck Millipore, WBKLS0050) and imaged
938 with eBLOT14 Touch Imager (e-BLOT).

939

940 **POND-qPCR**

941 Cell supernatant was clarified by centrifugation at 3,000 g for 5 minutes, then filtrated
942 through a cellulose acetate filter with 0.45 μm pore size (Corning). 200 μl of
943 supernatant was mixed with 800 μl TRIzol reagent (Magen) and 1 μl diluted External
944 RNA Controls Consortium (ERCC, 1:1000) synthetic spike-in RNAs (ThermoFisher).
945 RNA was extracted and dissolved in 20 μl nuclease-free water.

946

947 Next, 3 μl of supernatant RNAs or 250 ng cellular total RNA were reverse
948 transcribed using HiScript® III RT Super Mix (Vazyme) at 37°C for 15 minutes. The
949 cDNA products were diluted five-fold with 20 μl nuclease-free water, and 1 μl aliquot
950 was used for real-time qPCR with AceQ qPCR SYBR Green Master Mix (Vazyme) in

951 96-well plates on the StepOne Plus Real-Time PCR System (Applied Biosystems)
952 following standard protocols.

953

954 **Large-scale screening of RBPs affected by SNVs in UPS**

955 For each RBP, SNVs with strong effects were identified: ΔSNV < -0.2 and P_{WT} > 0.6.
956 We annotated the transcripts affected by SNVs using the GAP
957 (https://github.com/lipan6461188/GAP).

958

959 We used the STRING [96] database (https://string-db.org/) to annotate the enriched
960 biological functions of SNV-affected transcripts. Among the top five enriched biological
961 functions, we selected those related to ubiquitination or deubiquitination. Then, we
962 categorized the RBPs into four groups according to the functional enrichment with
963 ubiquitination or deubiquitination for their target transcripts affected by SNVs, i.e., the
964 highly ubiquitin-related (both ubiquitination- and deubiquitination-related), the weakly
965 ubiquitin-related (either ubiquitination- or deubiquitination-related), the not ubiquitin-
966 related (not ubiquitination- or deubiquitination-related), and the not function-enriched.

967

968 We then performed the KEGG enrichment analysis on the SNV-affected
969 transcripts annotated with highly ubiquitin-related. The enriched disease pathways
970 revealed potential regulatory associations among SNVs, RBPs, and disease-relevant
971 UPS genes.

972

973 **Quantification and statistical analysis**

974 All statistical analyses were performed with R. Where applicable, the sample size
975 (N) was specified either in the plot or in the corresponding figure legends. Analysis
976 details could be found in the method section, such as statistical tests, definitions,
977 etc. Boxplots were used to visualize the distribution of the data, with the box
978 representing the interquartile range (IQR). The median was indicated by a horizontal
979 line within the box, while the upper and lower hinges represented the 75th and 25th
980 percentiles, respectively. The whiskers extended to the most extreme data points

981 within 1.5 times the IQR from the box. Any data points beyond this
982 range were considered outliers and plotted individually. Statistical significance was
983 determined using appropriate tests, with *P* reported in the figure legends or main text.

984

985 **References**

986 1. Castello, A. *et al.* Insights into RNA Biology from an Atlas of Mammalian mRNA-
987 Binding Proteins. *Cell* **149**, 1393–1406 (2012).
988 2. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins.
989 *Nat Rev Genet* **15**, 829–845 (2014).
990 3. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA:
991 Molecules and Mechanisms. *Molecular Cell* **78**, 9–29 (2020).
992 4. Li, Y. & Sun, S. RNA dysregulation in neurodegenerative diseases. *EMBO J* **44**,
993 613–638 (2025).
994 5. Gebauer, F., Schwarzl, T., Valcárcel, J. & Hentze, M. W. RNA-binding proteins in
995 human genetic disease. *Nat Rev Genet* **22**, 185–198 (2021).
996 6. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids*
997 *Res* **29**, 308–311 (2001).
998 7. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and
999 supporting evidence. *Nucleic Acids Res* **46**, D1062–D1067 (2018).
1000 8. Yang, E.-W. *et al.* Allele-specific binding of RNA-binding proteins reveals functional
1001 genetic variants in the RNA. *Nat Commun* **10**, 1338 (2019).
1002 9. Van De Sompele, S. *et al.* Functional characterization of a novel non-coding
1003 mutation “Ghent +49A > G” in the iron-responsive element of L-ferritin causing
1004 hereditary hyperferritinaemia-cataract syndrome. *Sci Rep* **7**, 18025 (2017).
1005 10. Yamazaki, T., Liu, L., Conlon, E. G. & Manley, J. L. Burkitt lymphoma-related *TCF3*
1006 mutations alter TCF3 alternative splicing by disrupting hnRNPH1 binding. *RNA*
1007 *Biology* **17**, 1383–1390 (2020).
1008 11. Patton, E. E., Zon, L. I. & Langenau, D. M. Zebrafish disease models in drug
1009 discovery: from preclinical modelling to clinical trials. *Nat Rev Drug Discov* **20**, 611–
1010 628 (2021).
1011 12. Choi, T.-Y., Choi, T.-I., Lee, Y.-R., Choe, S.-K. & Kim, C.-H. Zebrafish as an animal
1012 model for biomedical research. *Exp Mol Med* **53**, 310–317 (2021).
1013 13. Liu, J. & Cao, X. RBP–RNA interactions in the control of autoimmunity and
1014 autoinflammation. *Cell Res* **33**, 97–115 (2023).
1015 14. Bryant, C. D. & Yazdani, N. RNA -binding proteins, neural development and the
1016 addictions. *Genes Brain Behav* **15**, 169–186 (2016).

15. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
16. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* **27**, 667–670 (2009).
17. Lambert, N. *et al.* RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Mol Cell* **54**, 887–900 (2014).
18. Gilbert, C. & Svejstrup, J. Q. RNA Immunoprecipitation for Determining RNA-Protein Associations In Vivo. *Curr Protoc Mol Biol* **75**, (2006).
19. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508–514 (2016).
20. Hafner, M. *et al.* Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).
21. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).
22. Wang, S. *et al.* Enhancement of LIN28B-induced hematopoietic reprogramming by IGF2BP3. *Genes Dev.* **33**, 1048–1068 (2019).
23. Ule, J. *et al.* CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science* **302**, 1212–1215 (2003).
24. Horlacher, M. *et al.* A systematic benchmark of machine learning methods for protein–RNA interaction prediction. *Briefings in Bioinform* **24**, bbad307 (2023).
25. Orenstein, Y., Wang, Y. & Berger, B. RCK: accurate and efficient inference of sequence- and structure-based protein–RNA binding models from RNAcompete data. *Bioinformatics* **32**, i351–i359 (2016).
26. Maticzka, D., Lange, S. J., Costa, F. & Backofen, R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol* **15**, R17 (2014).
27. Ben-Bassat, I., Chor, B. & Orenstein, Y. A deep neural network approach for learning intrinsic protein-RNA binding preferences. *Bioinformatics* **34**, i638–i646 (2018).
28. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**, 831–838 (2015).

29. Pan, X., Rijnbeek, P., Yan, J. & Shen, H.-B. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* **19**, 511 (2018).
30. Sun, L. *et al.* Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. *Cell Res* **31**, 495–516 (2021).
31. Zhu, H. *et al.* Dynamic characterization and interpretation for protein-RNA interactions across diverse cellular conditions using HDRNet. *Nat Commun* **14**, 6824 (2023).
32. Zhao, W. *et al.* POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res* **50**, D287–D294 (2022).
33. Zhang, Q., Wang, S., Li, Z., Pan, Y. & Huang, D.-S. Cross-Species Prediction of Transcription Factor Binding by Adversarial Training of a Novel Nucleotide-Level Deep Neural Network. *Adv Sci (Weinh)* **11**, e2405685 (2024).
34. Cochran, K. *et al.* Domain-adaptive neural networks improve cross-species prediction of transcription factor binding. *Genome Res.* **32**, 512–523 (2022).
35. Wang, J. *et al.* Generalizing to Unseen Domains: A Survey on Domain Generalization. *arXiv* https://doi.org/10.48550/ARXIV.2103.03097 (2021).
36. Lukasik, M., Bhojanapalli, S., Menon, A. K. & Kumar, S. Does label smoothing mitigate label noise? *arXiv* https://doi.org/10.48550/ARXIV.2003.02819 (2020).
37. Wang, Y. *et al.* Label smoothing regularization-based no hyperparameter domain generalization. *Knowledge-Based Systems* **309**, 112877 (2025).
38. Gerstberger, S., Hafner, M., Ascano, M. & Tuschl, T. Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. *Adv Exp Med Biol* **825**, 1–55 (2014).
39. Beckmann, B. M. *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat Commun* **6**, 10127 (2015).
40. Matia-González, A. M., Laing, E. E. & Gerber, A. P. Conserved mRNA-binding proteomes in eukaryotic organisms. *Nat Struct Mol Biol* **22**, 1027–1033 (2015).
41. Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**, 854-867.e9 (2018).
42. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).

43. Ramakrishnan, A. & Janga, S. C. Human protein-RNA interaction network is highly stable across mammals. *BMC Genomics* **20**, 1004 (2019).
44. Grishin, N. V. KH domain: one motif, two folds. *Nucleic Acids Research* **29**, 638–643 (2001).
45. Korn, S. M., Ulshöfer, C. J., Schneider, T. & Schlundt, A. Structures and target RNA preferences of the RNA-binding protein family of IGF2BPs: An overview. *Structure* **29**, 787–803 (2021).
46. Liu, Z.-P., Liu, S., Chen, R., Huang, X. & Wu, L.-Y. Structure alignment-based classification of RNA-binding pockets reveals regional RNA recognition motifs on protein surfaces. *BMC Bioinformatics* **18**, 27 (2017).
47. Irimia, M. *et al.* A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. *Cell* **159**, 1511–1523 (2014).
48. Giudice, G., Sánchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016**, baw035 (2016).
49. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
50. Budach, S. & Marsico, A. pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics* **34**, 3035–3037 (2018).
51. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**, D609–D617 (2025).
52. Harris, S. E. *et al.* Understanding species-specific and conserved RNA-protein interactions in vivo and in vitro. *Nat Commun* **15**, (2024).
53. Bahrami-Samani, E. & Xing, Y. Discovery of Allele-Specific Protein-RNA Interactions in Human Transcriptomes. *Am J Hum Genet* **104**, 492–502 (2019).
54. Wu, K. *et al.* Exploring noncoding variants in genetic diseases: from detection to functional insights. *J Genet Genomics* **51**, 111–132 (2024).
55. Beaumont, C. *et al.* Mutation in the iron responsive element of the L ferritin mRNA in a family with dominant hyperferritinaemia and cataract. *Nat Genet* **11**, 444–446 (1995).
56. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).

1115 57. Hu, L.-F. *et al.* Longitudinal and large-scale monitoring of transcriptome and RBP-
1116 RNA interactome in living cells by engineered protein nanocages. *bioRxiv*
1117 https://doi.org/10.1101/2025.03.13.642977 (2025).
1118 58. Lau, A. W., Fukushima, H. & Wei, W. The Fbw7 and betaTRCP E3 ubiquitin ligases
1119 and their roles in tumorigenesis. *Front Biosci (Landmark Ed)* **17**, 2197–2212 (2012).
1120 59. Herbst, A. *et al.* Methylation of helicase-like transcription factor in serum of patients
1121 with colorectal cancer is an independent predictor of disease recurrence. *Eur J*
1122 *Gastroenterol Hepatol* **21**, 565–569 (2009).
1123 60. Gu, C. *et al.* TRIM50 inhibits glycolysis and the malignant progression of gastric
1124 cancer by ubiquitinating PGK1. *Int J Biol Sci* **20**, 3656–3674 (2024).
1125 61. Zhao, C. *et al.* Repurposing an antidandruff agent to treating cancer: zinc pyrithione
1126 inhibits tumor growth via targeting proteasome-associated deubiquitinases.
1127 *Oncotarget* **8**, 13942–13956 (2017).
1128 62. Xie, J. & Guo, Q. AATF protects neural cells against oxidative damage induced by
1129 amyloid beta-peptide. *Neurobiol Dis* **16**, 150–157 (2004).
1130 63. Lee, S. *et al.* PTK2/FAK regulates UPS impairment via SQSTM1/p62
1131 phosphorylation in TARDBP/TDP-43 proteinopathies. *Autophagy* **16**, 1396–1412
1132 (2020).
1133 64. Louros, S. R. & Osterweil, E. K. Perturbed proteostasis in autism spectrum
1134 disorders. *J Neurochem* **139**, 1081–1092 (2016).
1135 65. Chua, J. P., De Calbiac, H., Kabashi, E. & Barmada, S. J. Autophagy and ALS:
1136 mechanistic insights and therapeutic implications. *Autophagy* **18**, 254–282 (2022).
1137 66. Tollervey, J. R. *et al.* Characterizing the RNA targets and position-dependent
1138 splicing regulation by TDP-43. *Nat Neurosci* **14**, 452–458 (2011).
1139 67. Seddighi, S. *et al.* Mis-spliced transcripts generate de novo proteins in TDP-43-
1140 related ALS/FTD. *Sci Transl Med* **16**, eadg7162 (2024).
1141 68. Kellett, E. A., Bademosi, A. T. & Walker, A. K. Molecular mechanisms and
1142 consequences of TDP-43 phosphorylation in neurodegeneration. *Mol*
1143 *Neurodegener* **20**, 53 (2025).
1144 69. Tran, N.-N. & Lee, B.-H. Functional implication of ubiquitinating and
1145 deubiquitinating mechanisms in TDP-43 proteinopathies. *Front Cell Dev Biol* **10**,
1146 931968 (2022).
1147 70. Kabashi, E. *et al.* TARDBP mutations in individuals with sporadic and familial
1148 amyotrophic lateral sclerosis. *Nat Genet* **40**, 572–574 (2008).

71. Xie, J. & Fu, T. DeepProtein: Deep Learning Library and Benchmark for Protein Sequence Learning. *Bioinformatics* (2025) doi:10.1093/bioinformatics/btaf165.
72. Uversky, V. N. Protein intrinsic disorder and structure-function continuum. *Prog Mol Biol Transl Sci* **166**, 1–17 (2019).
73. Nori, D. & Jin, W. RNAFlow: RNA Structure & Sequence Design via Inverse Folding-Based Flow Matching. *arXiv* https://doi.org/10.48550/ARXIV.2405.18768 (2024).
74. Yamada, K. & Hamada, M. Prediction of RNA-protein interactions using a nucleotide language model. *Bioinform Adv* **2**, vbac023 (2022).
75. Liu, H., Jian, Y., Zeng, C. & Zhao, Y. RNA-protein interaction prediction using network-guided deep learning. *Commun Biol* **8**, 247 (2025).
76. Lukavsky, P. J. *et al.* Molecular basis of UG-rich RNA recognition by the human splicing factor TDP-43. *Nat Struct Mol Biol* **20**, 1443–1449 (2013).
77. Park, J., Cho, J. & Song, E. J. Ubiquitin-proteasome system (UPS) as a target for anticancer treatment. *Arch Pharm Res* **43**, 1144–1161 (2020).
78. RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res* **49**, D212–D220 (2021).
79. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol Biol Evol* **38**, 3022–3027 (2021).
80. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
81. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
82. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods* **19**, 1109–1115 (2022).
83. Blum, M. *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res* **53**, D444–D456 (2025).
84. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
85. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).
86. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res* **43**, W39–W49 (2015).

87. Zhou, T. *et al.* itol.toolkit accelerates working with iTOL (Interactive Tree Of Life) by an automated generation of annotation files. *Bioinformatics* btad339 (2023).
88. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748–i756 (2018).
89. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun* **13**, 6968 (2022).
90. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).
91. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
92. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
93. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Res* **52**, D891–D899 (2024).
94. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
95. Sun, L. *et al.* RNA structure maps across mammalian cellular compartments. *Nat Struct Mol Biol* **26**, 322–330 (2019).
96. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638–D646 (2023).

1204 **Figure Legends**

1205 **Figure 1: Evolutionary conservation of RBPs and their target RNAs**

1206 (a) Bar plots showing the sequence and structural similarity of RBPs across 11 species,
1207 in terms of full-length sequence (SeqS in blue), full-length structure (StrS in red),
1208 RBD sequence (DseqS in blue) and RBD structure (DStrS in red). The
1209 phylogenetic tree is constructed based on the 18S rRNA sequences in the 11
1210 species, and shows the relative relationships between human and non-human
1211 species.

1212 (b) Examples of structural alignment of homologous RBPs, including AGO2 (the highly
1213 conserved), DDX51 (the considerable conserved), RBM10 (the intermediately
1214 conserved), and SERBP1 (the weakly conserved). The RBP structure in human
1215 (red) is compared to that in non-human species (blue).

1216 (c) SRRM4 binding distribution along meta-transcript in human (red line) and mouse
1217 (cyan line).

1218 (d) Binding motif alignment of homologous RBPs in human (left) and mouse (right).

1219 (e) Three-dimensional structure of RC3H1 in human (PDB ID 4QIL, left) and mouse
1220 (PDB ID 4QI2, right), showing bound RNA residues (orange), RBDs residues
1221 (blue), and other remaining RNA and protein residues (grey).

1222 See also **Supplementary Fig. 1** and **2**.

1223

1224 **Figure 2: The workflow of MuSIC**

1225 (a) Schematic showing the workflow of MuSIC. Step 1 (data pre-processing):
1226 collecting peaks of RBPs from multiple species, defining the positive and negative
1227 peaks, and constructing the training and validation set. Step 2 (feature engineering):
1228 encoding for the RBP-binding peaks and predicted RNA structures. Step 3 (Deep
1229 learning model): predicting RBP–RNA interactions using a CNN framework
1230 integrated with label smoothing strategy. Step 4 (Applications): comparing
1231 prediction performance with state-of-the-art computational methods, identifying
1232 and clustering RBP-binding motifs, analyzing conservation patterns of RBP-binding
1233 peaks, and quantifying the effects of SNVs on RBP binding.

(b) Flowchart showing data pre-processing and feature encoding. Top: positive (red) and negative (blue) peaks are encoded into one-hot vectors. Bottom: computationally predicted RNA structures are encoded into one-hot vectors.

(c) Diagram showing label smoothing strategy. Top left: using the image similarities between Cartoon and Sketch in computer vision for label smoothing strategy. Top right: using the RBP similarities between human (blue) and mouse (red) for label smoothing. Bottom: transforming the one-hot labels into smoothed labels based on RBP conservation and computing cross-entropy loss for both one-hot labels and smoothed labels.

(d) Diagram showing two components of the total loss L_{total}, L_{ture} (human in black) and L_{smooth} (mouse in red).

(e) Overview of the deep learning model architecture. One-hot encoded input (b) is processed sequentially through the convolutional layers, SE modules, residual blocks, max pooling, and fully connected layers to predict the binding probability.

(f) Boxplot showing the prediction performance for different types of input features, including only sequence (101nt / 200nt) (blue), 200nt sequence with label smoothing (green), sequence with label smoothing and predicted structure (red).

See also **Supplementary Fig. 3**.

Figure 3: Comparison of prediction performance between MuSIC and other computational methods

(a) Violin plot showing the prediction accuracy of MuSIC (red) and other computational methods (blue) for 19 cross-species datasets. *** $P < 0.005$, paired paired t-test.

(b) ROC curves showing the prediction accuracy of MuSIC (red) and other computational methods (blue) for two RBPs, YTHDC2 (left) and TAF15 (right).

(c) t-SNE clustering showing the output feature maps by MuSIC and other computational methods for separating the positive (cyan) and negative (grey) peaks.

1263 (d) Bar plot showing the overlapping ratio between the experimentally-derived and
1264 predicted peaks by MuSIC (red) and other computational methods (blue). *** $P <$
1265 0.005, paired t-test; ****, $P < 0.0001$, paired t-test.

1266 (e) Heatmap showing the overlapping ratio between the experimentally-derived and
1267 predicted peaks by MuSIC and other computational methods.

1268 (f) An example of predicted peaks distribution along the local region in mouse genome
1269 (chr7:96,200,000–96,400,000). Red indicates true positives and black indicates
1270 false negatives, as evaluated against the experimentally-derived peaks.

1271 See also **Supplementary Fig. 4, 5** and **6**.

1272

1273 **Figure 4: Reliable predictions of RBP-binding peaks and motif clustering by**
1274 **MuSIC**

1275 (a) Scatter plot showing the correlation between the cross-species prediction accuracy
1276 and RBP conservation for the 14 RBPs (mouse in red, zebrafish in blue).

1277 (b) Examples showing the consistency of the predicted binding motifs and
1278 experimentally-derived motifs (from ATtRACT database) of several RBPs.

1279 (c) Hierarchical clustering of 186 predicted binding motifs in human, mouse, zebrafish,
1280 and yeast. Stars indicate that there are no motifs enriched.

1281 See also **Supplementary Fig. 7** and **8**.

1282

1283 **Figure 5: Cross-species conservation of RBP-binding peaks**

1284 (a) Heatmap showing the percentage of homologous transcripts containing the
1285 conserved peaks bound by RBPs across the 11 species, where all the 186 RBPs
1286 can be categorized into three groups. Boxplot showing the conservation degree
1287 (measured by percentage of species) of the homologous transcripts containing the
1288 conserved peaks.

1289 (b) Examples showing the predicted peaks bound by RBM15B and ALKBH1 along the
1290 transcript ENST00000688663.1.

1291 (c) Forest plot showing the conservation of the peaks bound by the CPSF family. We
1292 include the predicted peaks ranging from human to each of the non-human species
1293 shown here according to the phylogenetic tree.

1294 See also **Supplementary Fig. 9**.

1295

1296 **Figure 6: RBP–RNA interactions affected by SNVs and their associations with**
1297 **human diseases**

1298 (a) The experimentally-derived and predicted effects of SNVs on SRSF1 and DDX3X
1299 binding. Obs.REF, the binding strength from CLIP-seq data for the reference allele;
1300 Obs.ALT, the binding strength from CLIP-seq data for the alternative allele;
1301 Pre.REF, the binding strength from MuSIC prediction for the reference allele; and
1302 Pre.ALT, the binding strength from MuSIC prediction for the alternative allele.
1303 (b) Consistency between the experimentally-derived and predicted effects of SNVs on
1304 RBP binding for the 186 RBPs. Green bar indicates high consistency (-logFDR ⩾
1305 1.3), whereas grey bar indicates no significance.
1306 (c) Schematic showing the workflow for quantifying the effects of homologous SNVs
1307 on RBP binding.
1308 (d) An example showing the effect of a homologous SNV (NC_000006.11:
1309 g.4059022A>G) on PUM2 binding in human. The saliency maps show the
1310 predicted effect of REF (A) (top) and ALT (G) (bottom) on PUM2 binding. In each
1311 saliency map, the two heatmap tracks display the sequence response (top) and
1312 structure response (bottom) of the model at the base level. The homologous SNV
1313 position is highlighted by a dashed box.
1314 (e) An example showing the effect of a homologous SNV (ENSMUST00000220965.2
1315 1446 A>G) on PUM2 binding in mouse. The saliency maps show the predicted
1316 effect of REF (A) (top) and ALT (G) (bottom) on PUM2 binding.
1317 (f) Examples showing *in vitro* RNA pull-down assays (n = 3) examining the effects of
1318 three homologous SNVs on PUM2 binding in human (top) and mouse (bottom).
1319 (g) Examples showing *in vivo* POND-qPCR (n = 3) quantification of REF/ALT target
1320 RNAs binding to human and mouse PUM2 in HEK293T. Relative RNA enrichment
1321 was calculated as: (RNA ^{supernatant}/ERCC spike-in)/ (RNA ^{cellular}/GAPDH). Then the

1322 ALT group was normalized to the REF group. Green bar represents RNA
1323 enrichment for REF sequences, and blue bar for ALT sequences.
1324 (h) Enriched biological functions of the SNV-affected transcripts. The 186 RBPs can
1325 be categorized into four groups according to the enriched functions of their target
1326 transcripts. In total, 21 RBPs are identified as highly ubiquitin-related, some of
1327 which have been revealed as associated with neurodegenerative diseases (red)
1328 and cancer (blue).
1329 See also **Supplementary Fig. 10, 11, 12** and **13**.

1330

1331 **Supplementary Figure Legends**

1332 **Supplementary Figure 1: Evolutionary conservation of RBPs across 11 species**

1333 (a-b) Heatmaps showing the full-length sequence similarity (a) and structural similarity
1334 (b) of RBPs across species.

1335 (c-d) Heatmaps showing the RBD sequence similarity (c) and structural similarity (d)
1336 of RBPs across species

1337

1338 **Supplementary Figure 2: Evolutionary conservation of RBP-targeted RNAs**

1339 (a) LIN28A binding distribution along meta-transcript in human (red line) and mouse
1340 (cyan line).

1341 (b-c) Three-dimensional structures of IGF2BP1 in human (b) and mouse (c), showing
1342 the bound RNA residues and three KH domains residues.

1343 (d-e) Heatmaps showing residue interaction score for IGF2BP1 in human (d) and
1344 mouse (e), with darker green represents stronger interaction scores.

1345 (f) Venn diagrams showing the overlap of RBPs between different species (human,
1346 mouse, zebrafish, fly, worm, *A. thaliana*, and yeast) collected from POSTAR3
1347 database.

1348

1349 **Supplementary Figure 3: Optimization of input feature and model architecture**

1350 (a) Boxplot showing the prediction performance for different input sequence length
1351 (101nt, 200nt, and 400nt) on MuSIC performance. *** $P < 0.005$, paired t-test.

1352 (b) Boxplot showing the prediction performance of label smoothing, including only
1353 sequence (blue), and sequence with label smoothing (green). ** $P < 0.01$, paired
1354 t-test.

1355 (c) Boxplot showing the prediction performance of different structural features,
1356 including only predicted structure (cyan), only sequence (blue), sequence with
1357 predicted structure (red), and sequence with experimentally-measured structure
1358 (red).

1359 (d) t-SNE clustering showing the output feature maps by different blocks (input, SE
1360 block, residual block, FCL) for separating the positive (cyan) and negative (grey)
1361 peaks.

1362 (e) Boxplot showing the prediction accuracy of MuSIC (red) and other machine
1363 learning models (grey). **** $P < 0.001$, paired t-test.

1364 (f) Boxplot showing the prediction accuracy of MuSIC (red) and other deep learning
1365 models (grey). * $P < 0.05$, ** $P < 0.01$, **** $P < 0.001$, paired t-test.

1366

1367 **Supplementary Figure 4: The performance comparison between MuSIC and**
1368 **other computational methods**

1369 (a) Violin plot showing the prediction accuracy of MuSIC (red) and other computational
1370 methods (blue) across 224 within-species datasets. * $P < 0.05$, *** $P < 0.005$,
1371 paired t-test.
1372 (b-c) Violin plot showing the prediction accuracy (PRAUC and ACC) of MuSIC (red)
1373 and other computational methods (blue) across 19 cross-species datasets. *** P
1374 < 0.005, paired t-test.

1375 (d) Violin plot showing the prediction accuracy of MuSIC (red) and other computational
1376 methods (blue) across 38 cross-species datasets. ** $P < 0.01$, *** $P < 0.005$, paired
1377 t-test.

1378 (e) Violin plot showing the prediction accuracy of MuSIC (red) and other computational
1379 methods (blue) across 17 cross-species datasets. *** $P < 0.005$, paired t-test.

1380 (f) Violin plot showing the prediction accuracy of MuSIC (red) and other computational
1381 methods (gradient blue) across 8 cross-species datasets. * $P < 0.05$, *** $P < 0.005$,
1382 paired t-test.

1383

1384 **Supplementary Figure 5: Comparison of feature learning discriminative power**
1385 **between MuSIC and other computational methods**

1386 (a) ROC curves showing the prediction accuracy of MuSIC (red) and other
1387 computational methods (blue) for 6 RBPs, 2RBPs for human (left), 2 RBPs for
1388 mouse (middle), and 2 RBPs for zebrafish (right).

1389 (b) t-SNE clustering showing the output feature maps by MuSIC and other
1390 computational methods for separating the positive (cyan) and negative (grey)
1391 peaks.

1392

1393 **Supplementary Figure 6: Consistency of RBP-binding distributions between**
1394 **MuSIC predictions and experimental data**

1395 (a-b) ELAVL1 binding distributions along meta-transcript in MuSIC predictions (red line)
1396 and experimentally-derived data (black line). Scatter plot showing the correlation
1397 of RBP-binding distributions from MuSIC predictions and experimental data for
1398 ELAVL1.

1399 (c) Scatter plot showing the correlation of RBP-binding distributions from MuSIC
1400 predictions and experimental data for 23 RBPs.

1401

1402 **Supplementary Figure 7: Evolutionary conservation of high-confidence RBP-**
1403 **binding motifs**

1404 (a) Flowchart showing the calculation of RBP conservation.

1405 (b) Scatter plot showing the correlation between the MuSIC-predicted accuracy and
1406 RBP conservation for RBPs from the 10 non-human species.

1407 (c) Examples showing the consistency of the predicted binding motifs and
1408 experimentally-derived motifs (from ATtRACT database) of several RBPs.

1409 (d) Examples showing high-confidence RBP-binding motifs across the 11 species,
1410 including human, orangutan, monkey, mouse, rat, chicken, frog, zebrafish, fly, *A.*
1411 *thaliana*, and yeast.

1412

1413 **Supplementary Figure 8: Clustering of RBP-binding motifs**

1414 Hierarchical clustering of 186 predicted binding motifs in human, orangutan, monkey,
1415 rat, chicken, frog, fly, and yeast. Stars indicate that there are no motifs enriched.

1416

1417 **Supplementary Figure 9: Cross-species conservation of predicted RBP-binding**
1418 **peaks**

1419 (a) Schematic showing the workflow for cross-species RBP-binding peaks
1420 conservation analysis.

1421 (b) Histogram showing the conservation of RBP-binding peaks across 11 species for
1422 186 RBPs.

1423 (c) Boxplot showing the conservation degree of RBP-binding peaks across 11 species
1424 for CPSF family (top) and YTH (bottom) family.

1425 (d) Forest plot showing the conservation of the peaks bound by the YTH family. We
1426 include the predicted peaks ranging from human to each of the non-human species
1427 shown here according to the phylogenetic tree.

1428

1429 **Supplementary Figure 10: Consistency between the predicted and**
1430 **experimentally-derived effects of SNVs on RBP–RNA interactions**

1431 (a) Schematic showing the workflow for separating peak signals into the REF and ALT
1432 groups. *** $P < 0.005$, unpaired t-test.

1433 (b) Schematic showing the workflow for evaluating consistency of SNV effects
1434 between MuSIC predictions and experimental data.

1435

1436 **Supplementary Figure 11: Experimental validations for the predicted effects of**
1437 **SNVs on RBP–RNA interactions in human and mouse**

1438 (a) Flowchart of using the POND method to detect the interaction between PUM2 and
1439 REF/ALT target RNA. PUM2 is fused to the C-terminus of the EPN24 nanocage
1440 monomer, while REF and ALT RNA targets are cloned into the 3'UTR of GFP on a
1441 separate plasmid. In HEK293T cells, EPN24-PUM2 is co-transfected with either
1442 GFP-REF or GFP-ALT RNA target plasmids. After transfection, EPN24-PUM2
1443 packaged target RNAs into extracellular supernatant. Both supernatant and cells
1444 are collected for RNA extraction, followed by analysis of differential enrichment
1445 between REF and ALT target RNAs with RT-qPCR.

1446 (b) RNA pull-down assays and immunoblots for human PUM2 (left) and mouse PUM2
1447 (right). Biotin-labeled REF and ALT RNA probes were used to perform RNA pull-
1448 down assays from lysates of 293T cells overexpressing myc-tagged PUM2. The
1449 pulled-down complexes were then analyzed by immunoblotting to detect the
1450 presence of PUM2 protein. The upper blot represents the results of repeat 2, and
1451 the lower blot represents the results of repeat 3.

1452 (c) Bar plot quantifying *in vitro* RNA pull-down (n = 3) enrichment of human and mouse
1453 PUM2 protein binding to REF and ALT target RNAs. For each condition, pulled-
1454 down PUM2 levels were normalized to PUM2 in the input lysate. Then the ALT
1455 group was normalized to the REF group. Green bars represent REF sequences,
1456 and blue bars represent ALT sequences. All *P* values are calculated using two-
1457 tailed unpaired Student's t-test.

1458

1459 **Supplementary Figure 12: Predicted effects of SNVs on RBP–RNA interactions**
1460 **and their associations with human diseases**

1461 (a) Schematic showing the workflow of biological function enrichment analysis for
1462 RBPs affected by SNVs.

1463 (b) Bar chart showing the terms of biological functions associated with the weakly
1464 ubiquitin-related RBPs.

1465 (c) Sankey diagram showing the potential regulatory associations among SNVs, RBPs,
1466 and UPS-related disease.

1467

1468 **Supplementary Figure 13: Predicted effects of SNVs on TARDBP binding and**
1469 **the enriched biological functions of the transcripts containing the SNVs**

1470 (a) An example showing the effect of a homologous SNV (ENSMUST00000239066
1471 1473 U>C) on TARDBP binding in mouse. The saliency maps show the predicted
1472 effect of REF (U) (top) and ALT (C) (bottom) on TARDBP binding.

1473 (b) Gene Ontology enrichment result showing the biological process pathways for the
1474 TARDBP-bound genes perturbed by SNVs.

1475 (c) Reactome pathway enrichment result showing the biological functions for
1476 TARDBP-bound genes perturbed by SNVs.

1477

1478 **Supplementary Table Legends**

1479 **Supplementary Table 1: Evolutionary conservation of RBPs across 11 species**

1480

1481 **Supplementary Table 2: Conservation-based grouping of 216 RBPs**

1482

1483 **Supplementary Table 3: MuSIC training and validation datasets**

1484

1485 **Supplementary Table 4: Prediction performance of MuSIC**

1486

1487 **Supplementary Table 5: Predicted RBP-binding motifs from MuSIC**

1488

1489 **Supplementary Table 6: Consistency of SNV effects**

1490

1491 **Supplementary Table 7: Experimental validation of SNV effects**

1492

1493 **Supplementary Table 8: Functional enrichment of SNV-affected genes**

1494

Figure 1

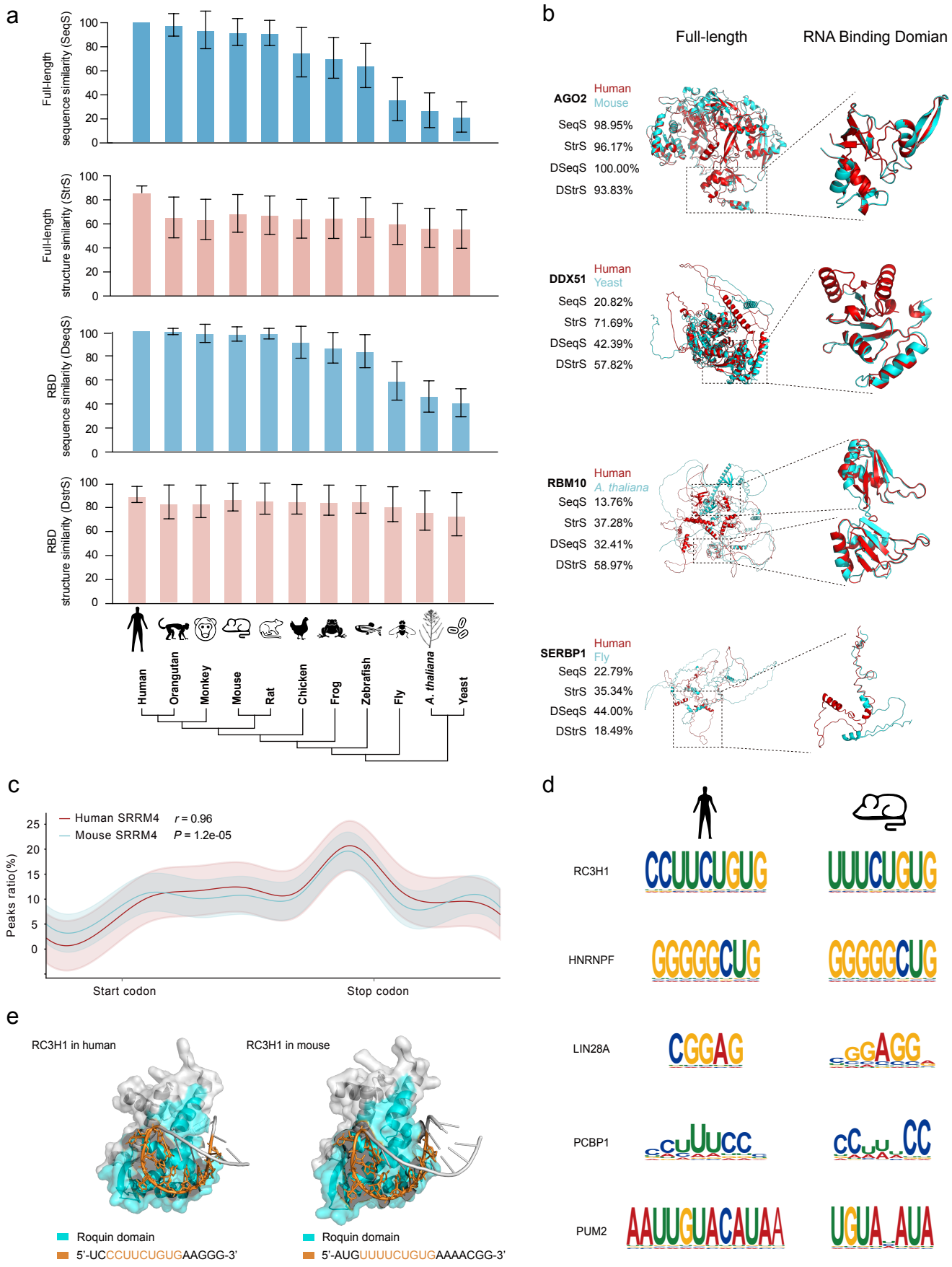

a
Full-length sequence similarity (SeqS)
Full-length structure similarity (StrS)
RBD sequence similarity (DseqS)
RBD structure similarity (DstrS)
Human
Orangutan
Monkey
Mouse
Rat
Chicken
Frog
Zebrafish
Fly
A. thaliana
Yeast
b
Full-length
RNA Binding Domian
AGO2
Human
Mouse
SeqS 98.95%
StrS 96.17%
DSeqS 100.00%
DStrS 93.83%
DDX51
Human
Yeast
SeqS 20.82%
StrS 71.69%
DSeqS 42.39%
DStrS 57.82%
RBM10
Human
A. thaliana
SeqS 13.76%
StrS 37.28%
DSeqS 32.41%
DStrS 58.97%
SERBP1
Human
Fly
SeqS 22.79%
StrS 35.34%
DSeqS 44.00%
DStrS 18.49%
c
Human SRRM4 r = 0.96
Mouse SRRM4 P = 1.2e-05
Peaks ratio(%)
Start codon
Stop codon
d
RC3H1
HNRNPF
LIN28A
PCBP1
PUM2
e
RC3H1 in human
RC3H1 in mouse
Roquin domain
5'-UCCCUUCUGUGAAGGG-3'
Roquin domain
5'-AUGUUUUCUGUGAAAACGG-3'

Figure 2

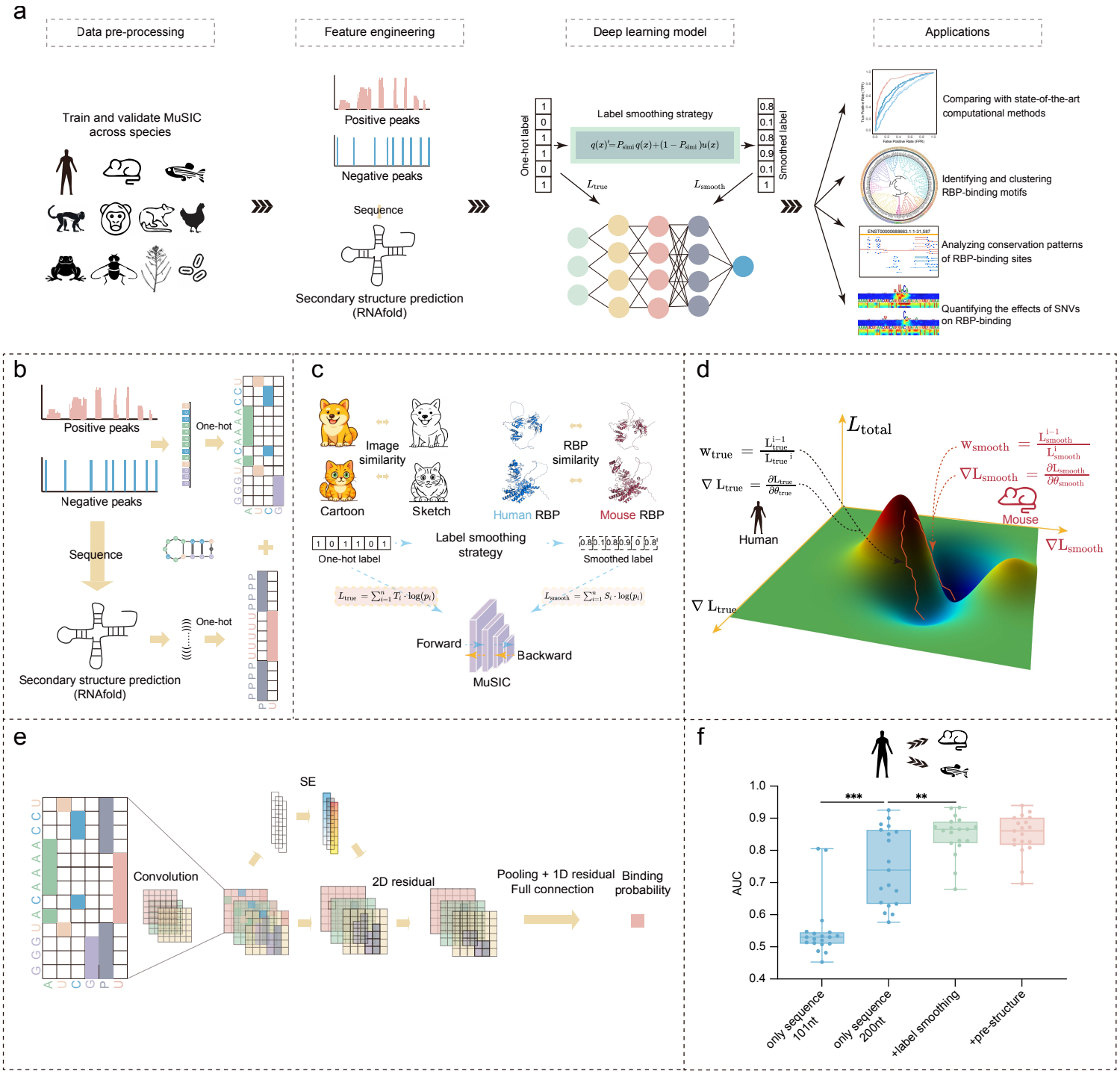

a
Data pre-processing
Feature engineering
Deep learning model
Applications
Train and validate MuSIC across species
Positive peaks
Negative peaks
Sequence
Secondary structure prediction (RNAfold)
One-hot label
Label smoothing strategy
Smoothed label
Comparing with state-of-the-art computational methods
Identifying and clustering RBP-binding motifs
Analyzing conservation patterns of RBP-binding sites
Quantifying the effects of SNVs on RBP-binding
b
One-hot
c
Image similarity
Cartoon
Sketch
RBP similarity
Human RBP
Mouse RBP
Forward
Backward
MuSIC
d
Human
Mouse
e
SE
Convolution
2D residual
Pooling + 1D residual
Full connection
Binding probability
f
AUC
only sequence 101nt
only sequence 200nt
+label smoothing
+pre-structure

Figure 3

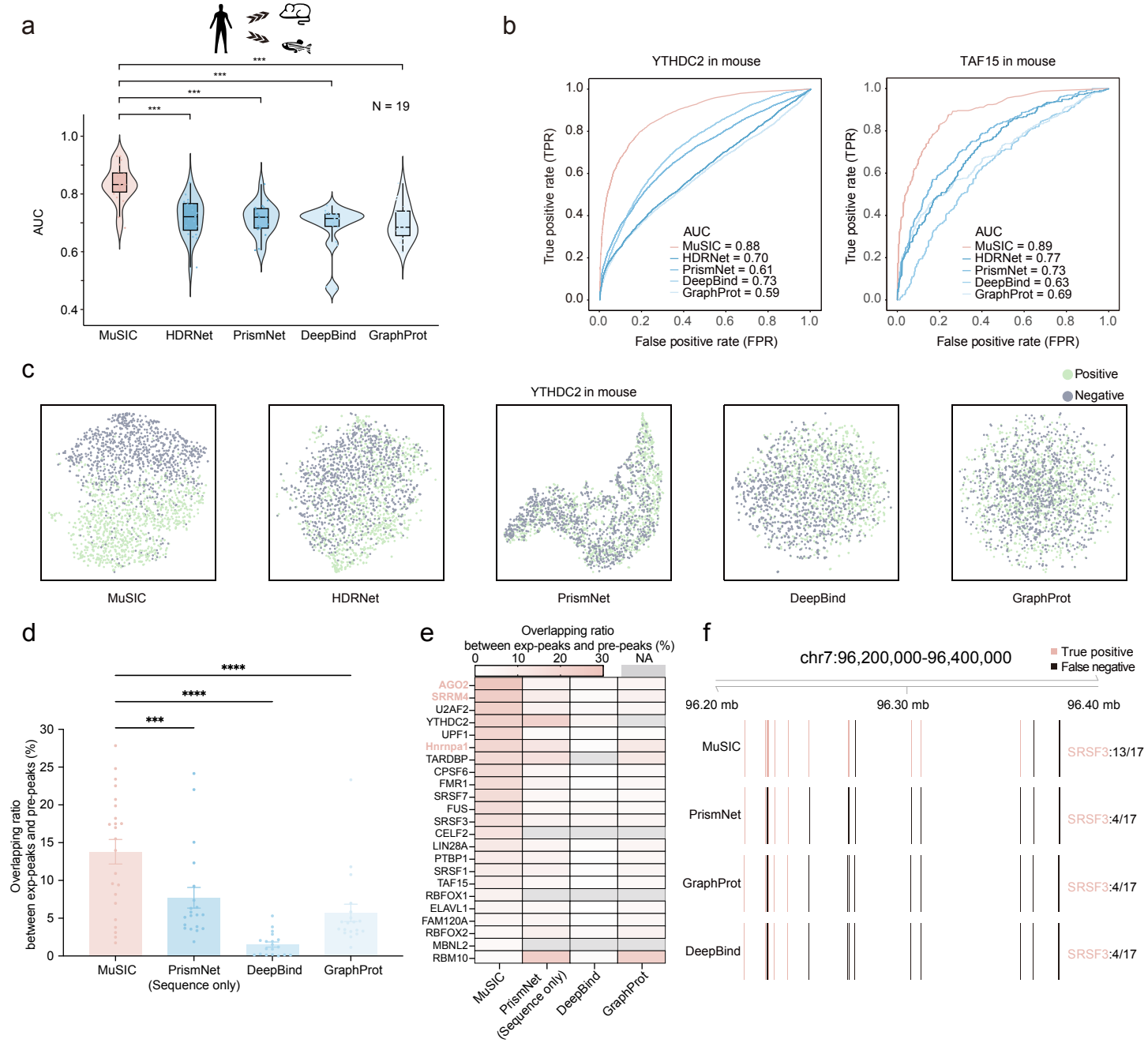

a
N = 19
AUC
MuSIC
HDRNet
PrismNet
DeepBind
GraphProt
b
YTHDC2 in mouse
TAF15 in mouse
True positive rate (TPR)
False positive rate (FPR)
AUC
MuSIC = 0.88
HDRNet = 0.70
PrismNet = 0.61
DeepBind = 0.73
GraphProt = 0.59
AUC
MuSIC = 0.89
HDRNet = 0.77
PrismNet = 0.73
DeepBind = 0.63
GraphProt = 0.69
c
YTHDC2 in mouse
Positive
Negative
MuSIC
HDRNet
PrismNet
DeepBind
GraphProt
d
Overlapping ratio between exp-peaks and pre-peaks (%)
MuSIC
PrismNet (Sequence only)
DeepBind
GraphProt
e
Overlapping ratio between exp-peaks and pre-peaks (%)
NA
AGO2
SRRM4
U2AF2
YTHDC2
UPF1
Hnrnpa1
TARDBP
CPSF6
FMR1
SRSF7
FUS
SRSF3
CELF2
LIN28A
PTBP1
SRSF1
TAF15
RBFOX1
ELAVL1
FAM120A
RBFOX2
MBNL2
RBM10
MuSIC
PrismNet (Sequence only)
DeepBind
GraphProt
f
chr7:96,200,000-96,400,000
True positive
False negative
96.20 mb
96.30 mb
96.40 mb
MuSIC
PrismNet
GraphProt
DeepBind
SRSF3:13/17
SRSF3:4/17
SRSF3:4/17
SRSF3:4/17

Figure 4

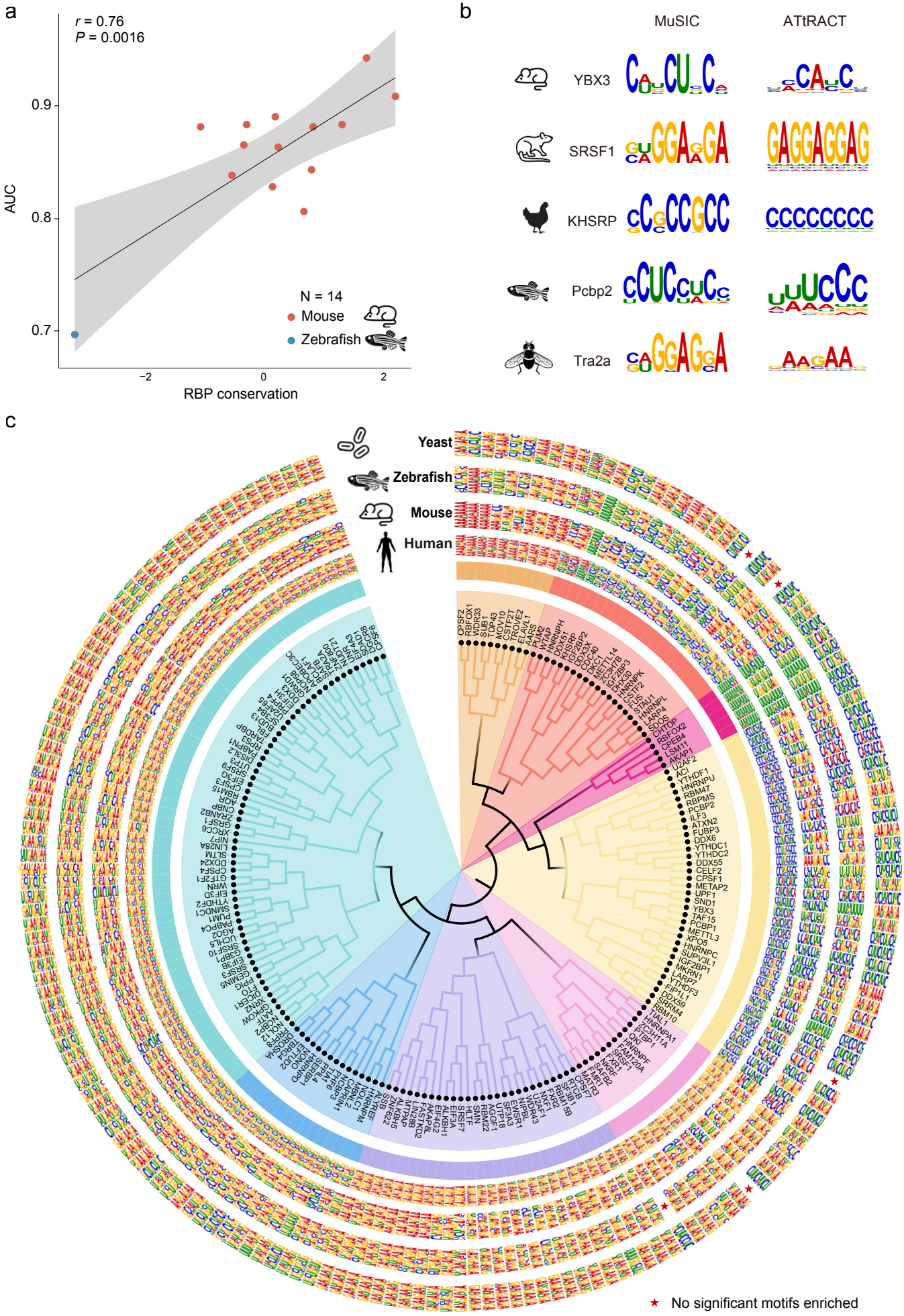

a
r = 0.76
P = 0.0016
AUC
0.9
0.8
0.7
−2
0
2
RBP conservation
N = 14
Mouse
Zebrafish
b
MuSIC
ATtRACT
YBX3
SRSF1
KHSRP
Pcbp2
Tra2a
c
Yeast
Zebrafish
Mouse
Human
No significant motifs enriched

Figure 5

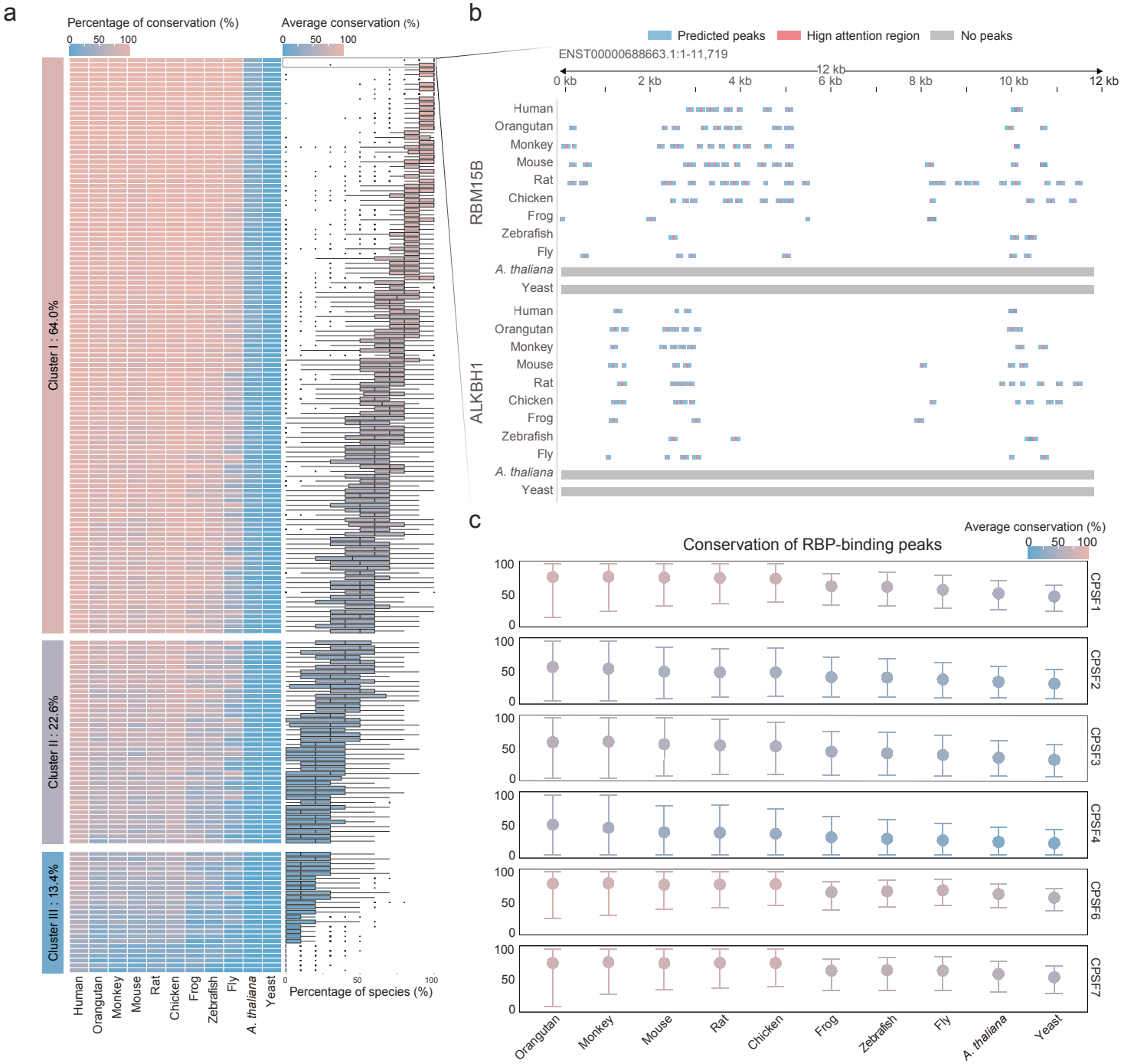

a
Percentage of conservation (%)
0 50 100
Average conservation (%)
0 50 100
Cluster I : 64.0%
Cluster II : 22.6%
Cluster III : 13.4%
Human
Orangutan
Monkey
Mouse
Rat
Chicken
Frog
Zebrafish
Fly
A. thaliana
Yeast
0 50 100
Percentage of species (%)
b
Predicted peaks
Hign attention region
No peaks
ENST00000688663.1:1-11,719
12 kb
0 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
RBM15B
ALKBH1
Human
Orangutan
Monkey
Mouse
Rat
Chicken
Frog
Zebrafish
Fly
A. thaliana
Yeast
c
Conservation of RBP-binding peaks
Average conservation (%)
0 50 100
CPSF1
CPSF2
CPSF3
CPSF4
CPSF6
CPSF7
100
50
0
Orangutan
Monkey
Mouse
Rat
Chicken
Frog
Zebrafish
Fly
A. thaliana
Yeast

Figure 6

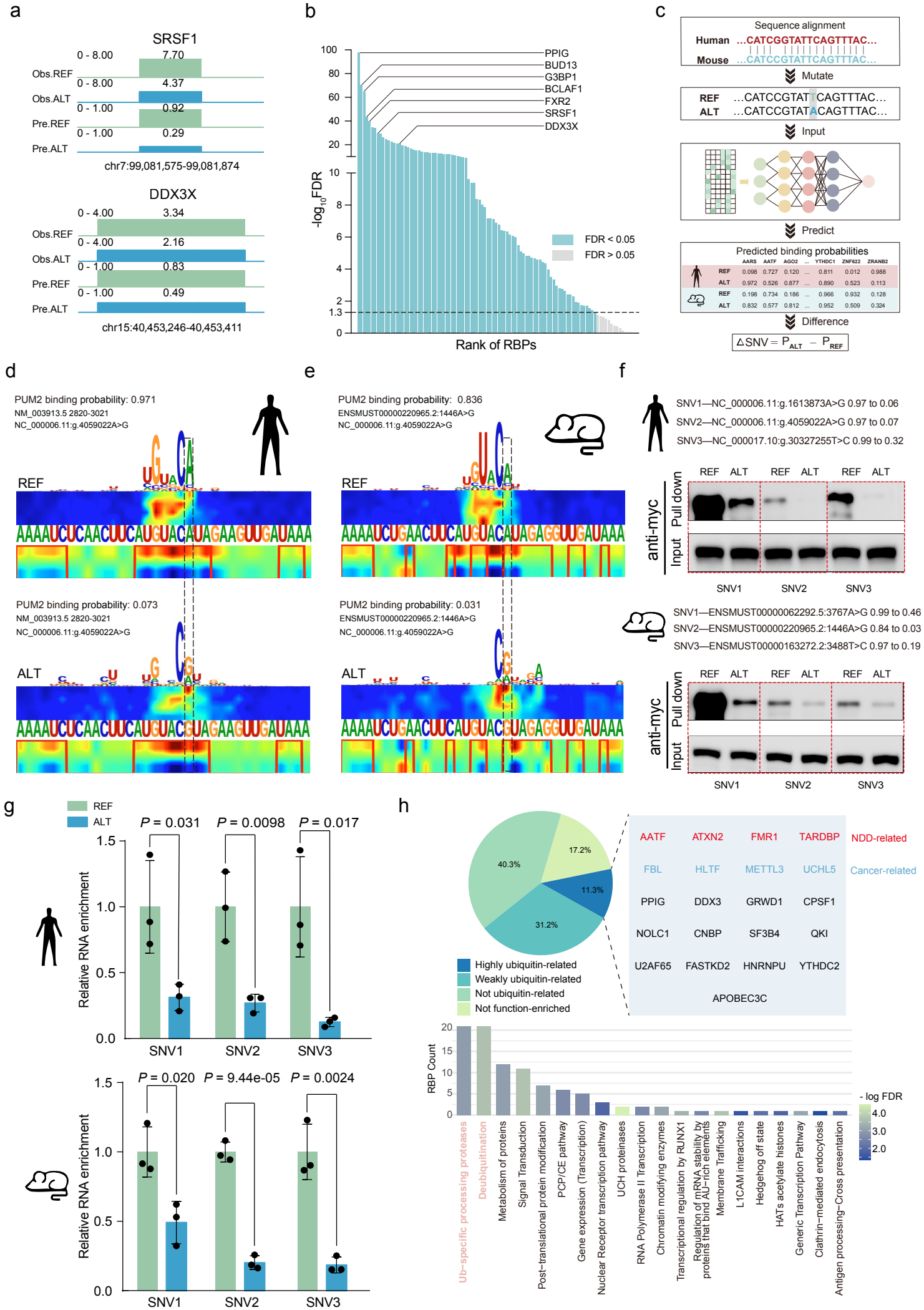

a
SRSF1
0 - 8.00
7.70
Obs.REF
0 - 8.00
4.37
Obs.ALT
0 - 1.00
0.92
Pre.REF
0 - 1.00
0.29
Pre.ALT
chr7:99,081,575-99,081,874
DDX3X
0 - 4.00
3.34
Obs.REF
0 - 4.00
2.16
Obs.ALT
0 - 1.00
0.83
Pre.REF
0 - 1.00
0.49
Pre.ALT
chr15:40,453,246-40,453,411
b
PPIG
BUD13
G3BP1
BCLAF1
FXR2
SRSF1
DDX3X
-log10FDR
FDR < 0.05
FDR > 0.05
Rank of RBPs
c
Sequence alignment
Human ...CATCGGTATTCAGTTTAC...
Mouse ...CATCCGTATTCAGTTTAC...
Mutate
REF ...CATCCGTATTCAGTTTAC...
ALT ...CATCCGTATACAGTTTAC...
Input
Predict
Predicted binding probabilities
AARS AATF AGO2 ... YTHDC1 ZNF622 ZRANB2
REF 0.098 0.727 0.120 ... 0.811 0.012 0.988
ALT 0.972 0.526 0.877 ... 0.890 0.523 0.113
REF 0.198 0.734 0.186 ... 0.966 0.932 0.128
ALT 0.832 0.577 0.812 ... 0.952 0.509 0.324
Difference
ΔSNV = P_ALT − P_REF
d
PUM2 binding probability: 0.971
NM_003913.5 2820-3021
NC_000006.11:g.4059022A>G
REF
AAAAUCUCAACUUCAUGUACAUAGAAGUUGAUAAA
PUM2 binding probability: 0.073
NM_003913.5 2820-3021
NC_000006.11:g.4059022A>G
ALT
AAAAUCUCAACUUCAUGUACGUAGAAGUUGAUAAA
e
PUM2 binding probability: 0.836
ENSMUST00000220965.2:1446A>G
NC_000006.11:g.4059022A>G
REF
AAAAUCUGAACUUCAUGUACAUAGAGGUUGAUAAA
PUM2 binding probability: 0.031
ENSMUST00000220965.2:1446A>G
NC_000006.11:g.4059022A>G
ALT
AAAAUCUGAACUUCAUGUACGUAGAGGUUGAUAAA
f
SNV1—NC_000006.11:g.1613873A>G 0.97 to 0.06
SNV2—NC_000006.11:g.4059022A>G 0.97 to 0.07
SNV3—NC_000017.10:g.30327255T>C 0.99 to 0.32
REF ALT REF ALT REF ALT
anti-myc
Pull down
Input
SNV1 SNV2 SNV3
SNV1—ENSMUST00000062292.5:3767A>G 0.99 to 0.46
SNV2—ENSMUST00000220965.2:1446A>G 0.84 to 0.03
SNV3—ENSMUST00000163272.2:3488T>C 0.97 to 0.19
REF ALT REF ALT REF ALT
anti-myc
Pull down
Input
SNV1 SNV2 SNV3
g
REF
ALT
P = 0.031 P = 0.0098 P = 0.017
Relative RNA enrichment
SNV1 SNV2 SNV3
P = 0.020 P = 9.44e-05 P = 0.0024
Relative RNA enrichment
SNV1 SNV2 SNV3
h
40.3%
17.2%
11.3%
31.2%
Highly ubiquitin-related
Weakly ubiquitin-related
Not ubiquitin-related
Not function-enriched
AATF ATXN2 FMR1 TARDBP NDD-related
FBL HLTF METTL3 UCHL5 Cancer-related
PPIG DDX3 GRWD1 CPSF1
NOLC1 CNBP SF3B4 QKI
U2AF65 FASTKD2 HNRNPU YTHDC2
APOBEC3C
RBP Count
- log FDR
Ub-specific processing proteases
Deubiquitination
Metabolism of proteins
Signal Transduction
Post-translational protein modification
PCP/CE pathway
Gene expression (Transcription)
Nuclear Receptor transcription pathway
UCH proteinases
RNA Polymerase II Transcription
Chromatin modifying enzymes
Transcriptional regulation by RUNX1
Regulation of mRNA stability by proteins that bind AU-rich elements
Membrane Trafficking
L1CAM interactions
Hedgehog off state
HATs acetylate histones
Generic Transcription Pathway
Clathrin-mediated endocytosis
Antigen processing-Cross presentation

Figure S1

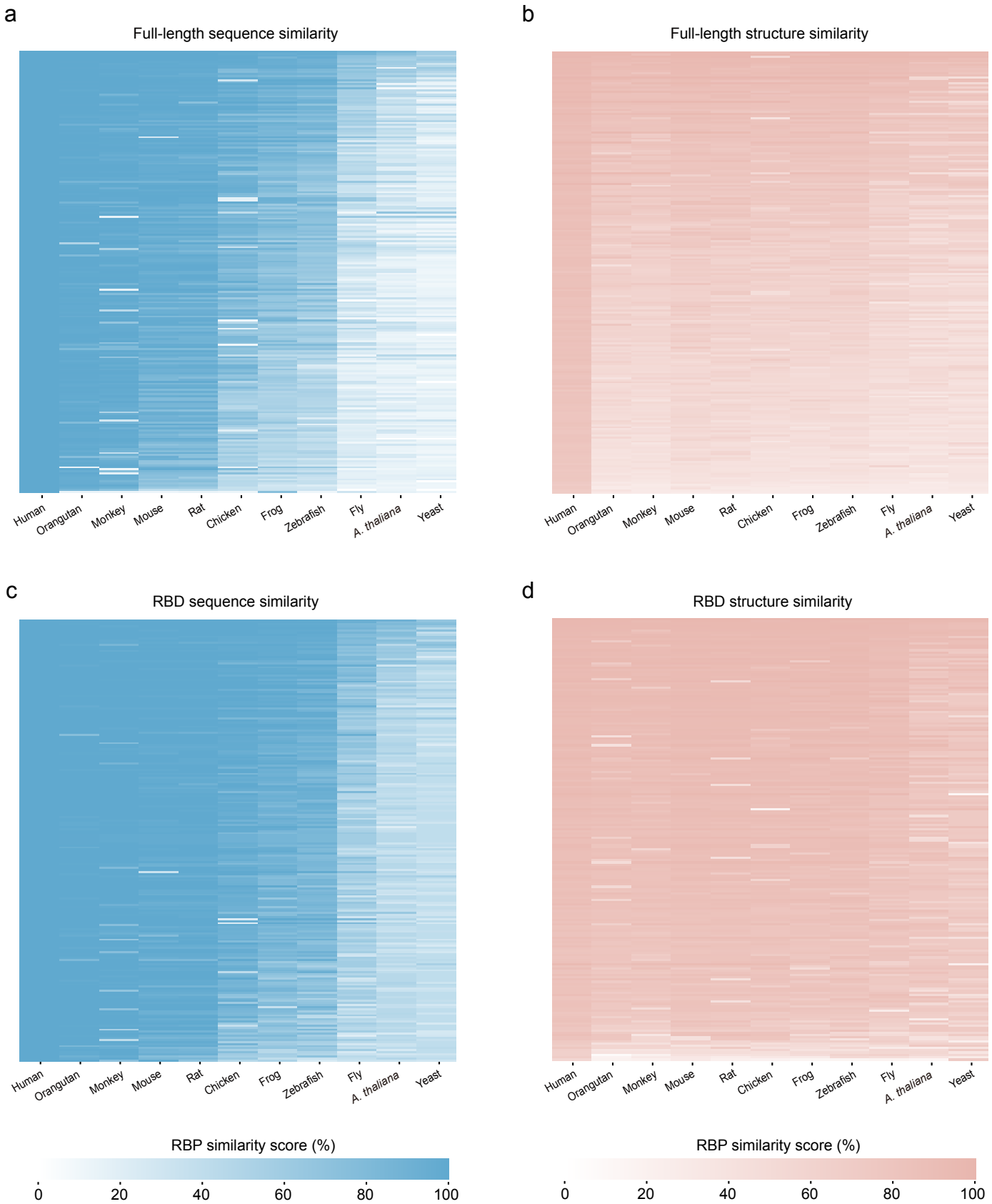

a
Full-length sequence similarity
b
Full-length structure similarity
c
RBD sequence similarity
d
RBD structure similarity
Human
Orangutan
Monkey
Mouse
Rat
Chicken
Frog
Zebrafish
Fly
A. thaliana
Yeast
RBP similarity score (%)
0
20
40
60
80
100

Figure S2

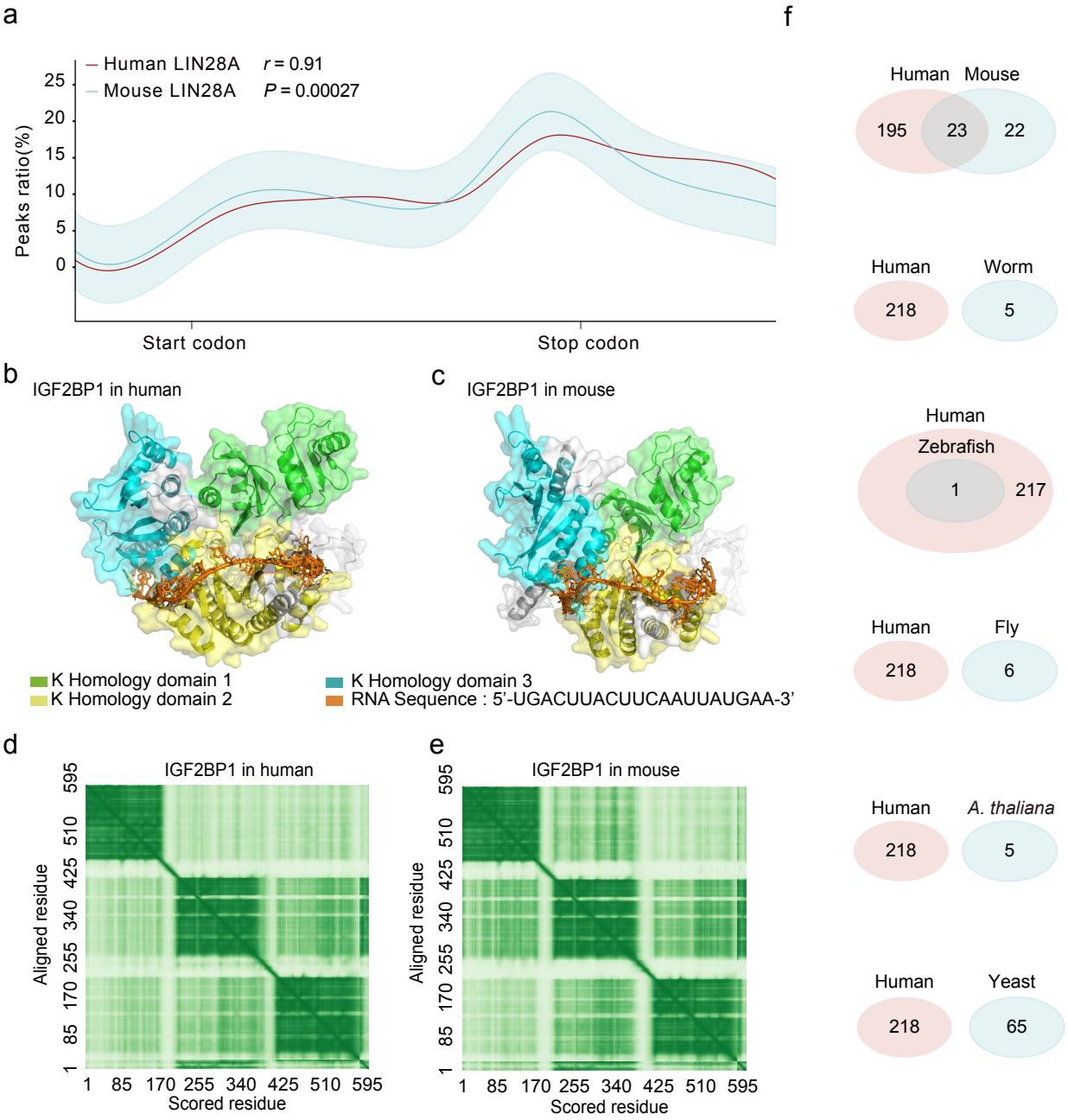

a
Human LIN28A r = 0.91
Mouse LIN28A P = 0.00027
Peaks ratio(%)
25
20
15
10
5
0
Start codon
Stop codon
b
IGF2BP1 in human
c
IGF2BP1 in mouse
K Homology domain 1
K Homology domain 2
K Homology domain 3
RNA Sequence : 5'-UGACUUACUUCAAUUAUGAA-3'
d
IGF2BP1 in human
Aligned residue
Scored residue
e
IGF2BP1 in mouse
Aligned residue
Scored residue
1 85 170 255 340 425 510 595
f
Human Mouse
195 23 22
Human Worm
218 5
Human
Zebrafish
1 217
Human Fly
218 6
Human A. thaliana
218 5
Human Yeast
218 65

Figure S3

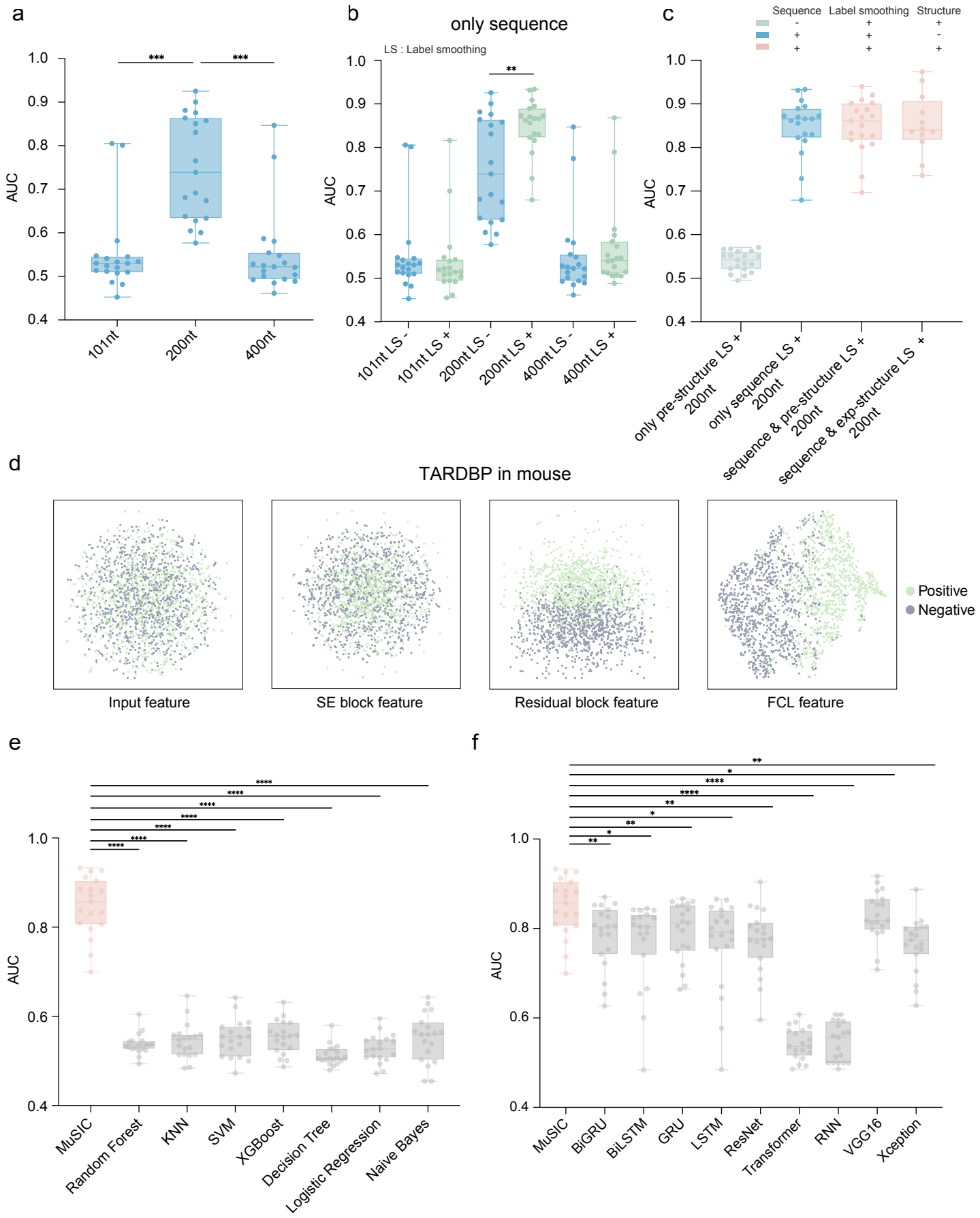

a
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
***
***
101nt
200nt
400nt
b
only sequence
LS : Label smoothing
**
101nt LS -
101nt LS +
200nt LS -
200nt LS +
400nt LS -
400nt LS +
c
Sequence
Label smoothing
Structure
only pre-structure LS + 200nt
only sequence LS + 200nt
sequence & pre-structure LS + 200nt
sequence & exp-structure LS + 200nt
d
TARDBP in mouse
Input feature
SE block feature
Residual block feature
FCL feature
Positive
Negative
e
MuSIC
Random Forest
KNN
SVM
XGBoost
Decision Tree
Logistic Regression
Naive Bayes
f
MuSIC
BiGRU
BiLSTM
GRU
LSTM
ResNet
Transformer
RNN
VGG16
Xception

Figure S4

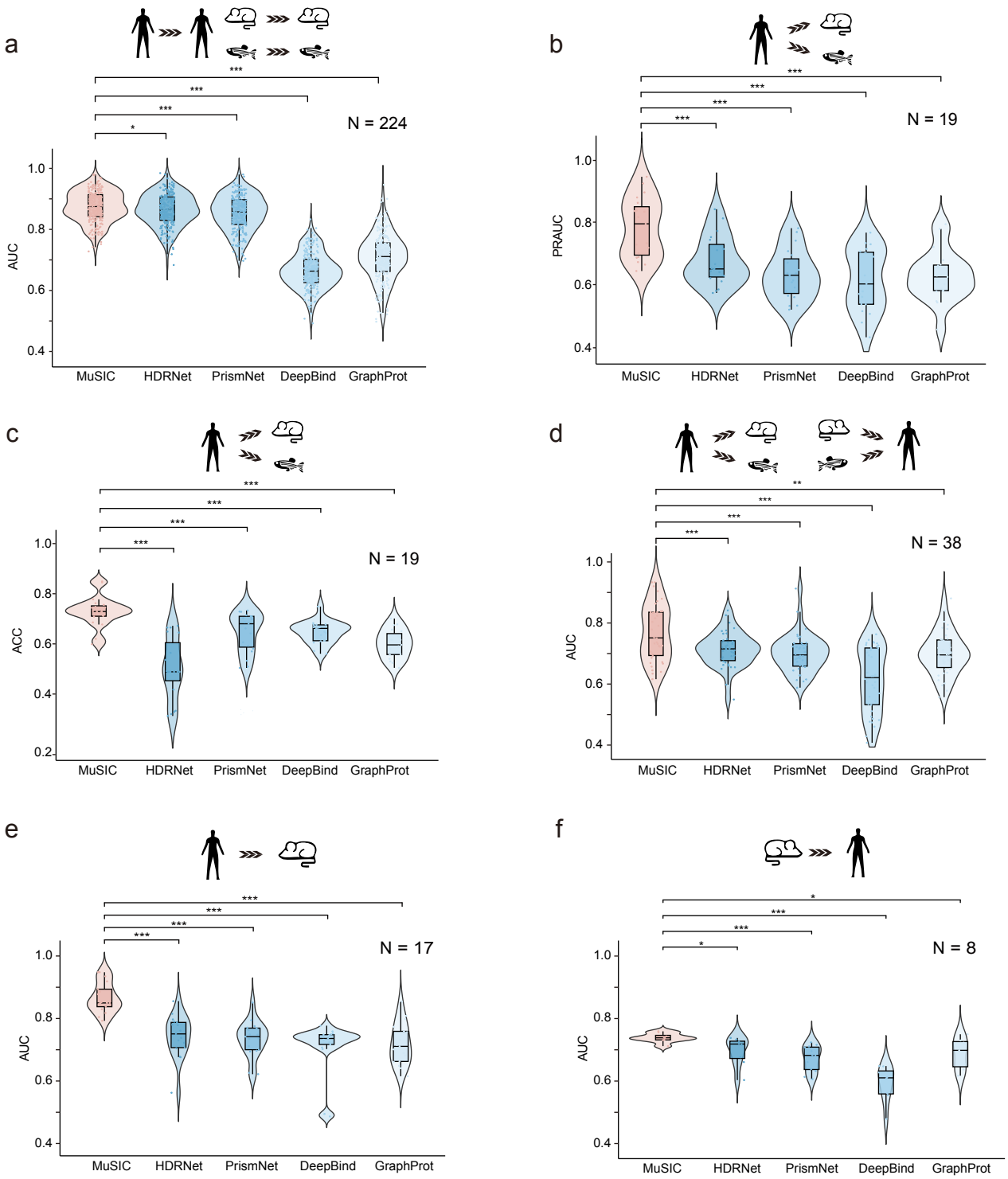
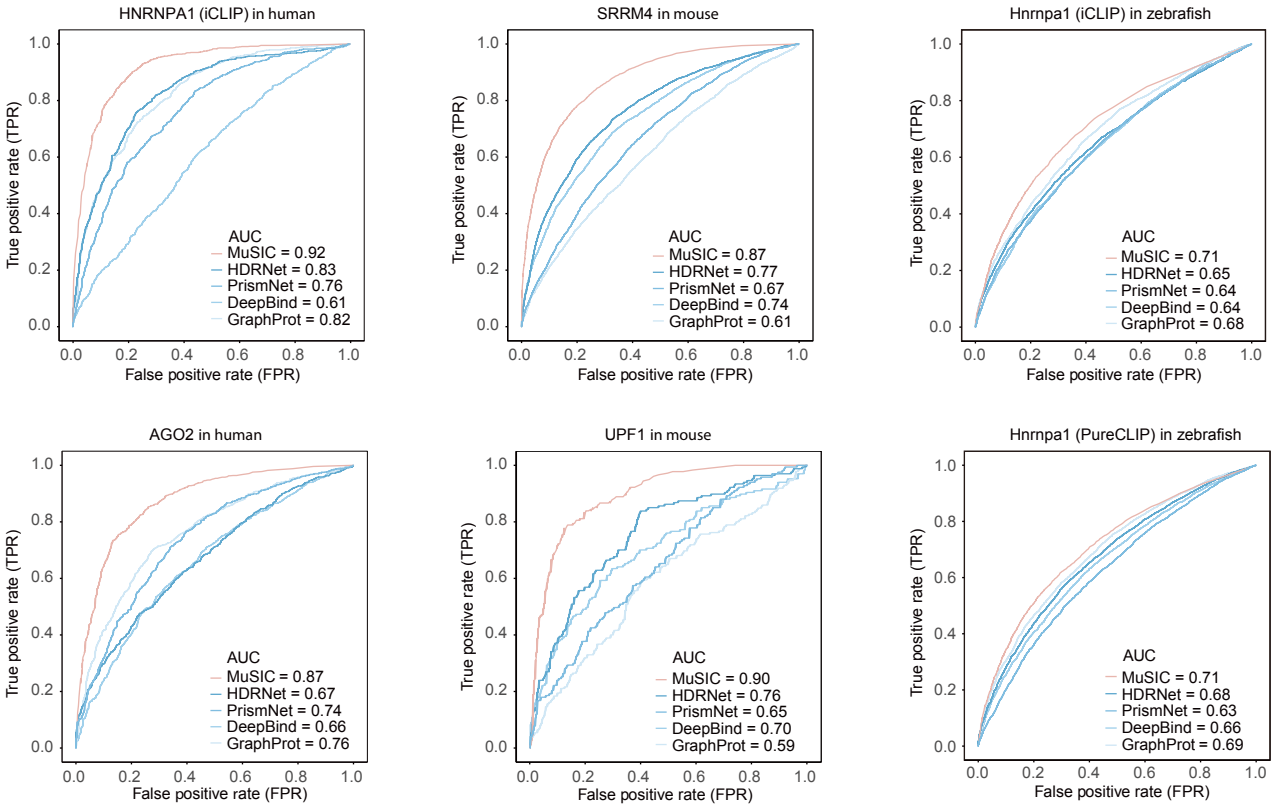

a
N = 224
AUC
MuSIC
HDRNet
PrismNet
DeepBind
GraphProt
b
N = 19
PRAUC
c
N = 19
ACC
d
N = 38
AUC
e
N = 17
AUC
f
N = 8
AUC

Figure S5

a


HNRNPA1 (iCLIP) in human
AUC
MuSIC = 0.92
HDRNet = 0.83
PrismNet = 0.76
DeepBind = 0.61
GraphProt = 0.82
SRRM4 in mouse
AUC
MuSIC = 0.87
HDRNet = 0.77
PrismNet = 0.67
DeepBind = 0.74
GraphProt = 0.61
Hnrnpa1 (iCLIP) in zebrafish
AUC
MuSIC = 0.71
HDRNet = 0.65
PrismNet = 0.64
DeepBind = 0.64
GraphProt = 0.68
AGO2 in human
AUC
MuSIC = 0.87
HDRNet = 0.67
PrismNet = 0.74
DeepBind = 0.66
GraphProt = 0.76
UPF1 in mouse
AUC
MuSIC = 0.90
HDRNet = 0.76
PrismNet = 0.65
DeepBind = 0.70
GraphProt = 0.59
Hnrnpa1 (PureCLIP) in zebrafish
AUC
MuSIC = 0.71
HDRNet = 0.68
PrismNet = 0.63
DeepBind = 0.66
GraphProt = 0.69
True positive rate (TPR)
False positive rate (FPR)

b

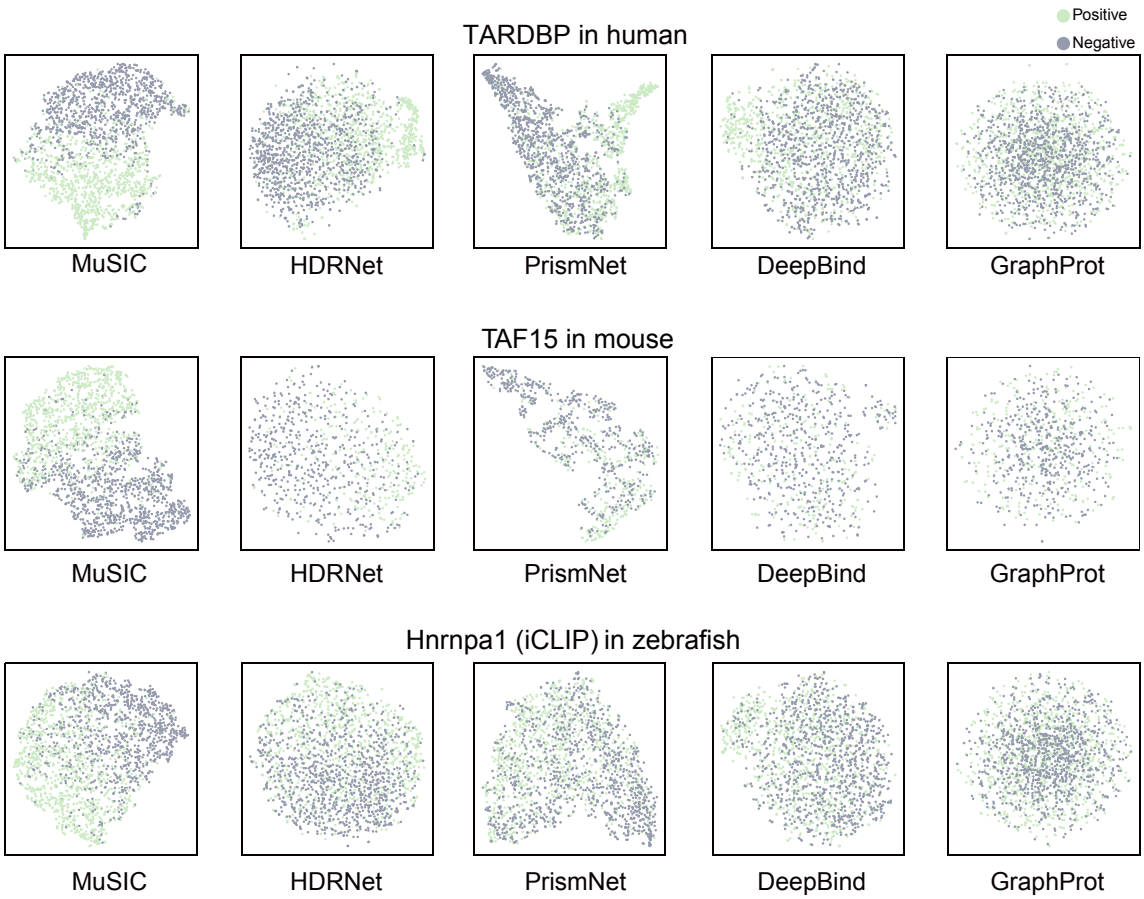

Positive
Negative
TARDBP in human
MuSIC
HDRNet
PrismNet
DeepBind
GraphProt
TAF15 in mouse
MuSIC
HDRNet
PrismNet
DeepBind
GraphProt
Hnrnpa1 (iCLIP) in zebrafish
MuSIC
HDRNet
PrismNet
DeepBind
GraphProt

Figure S6

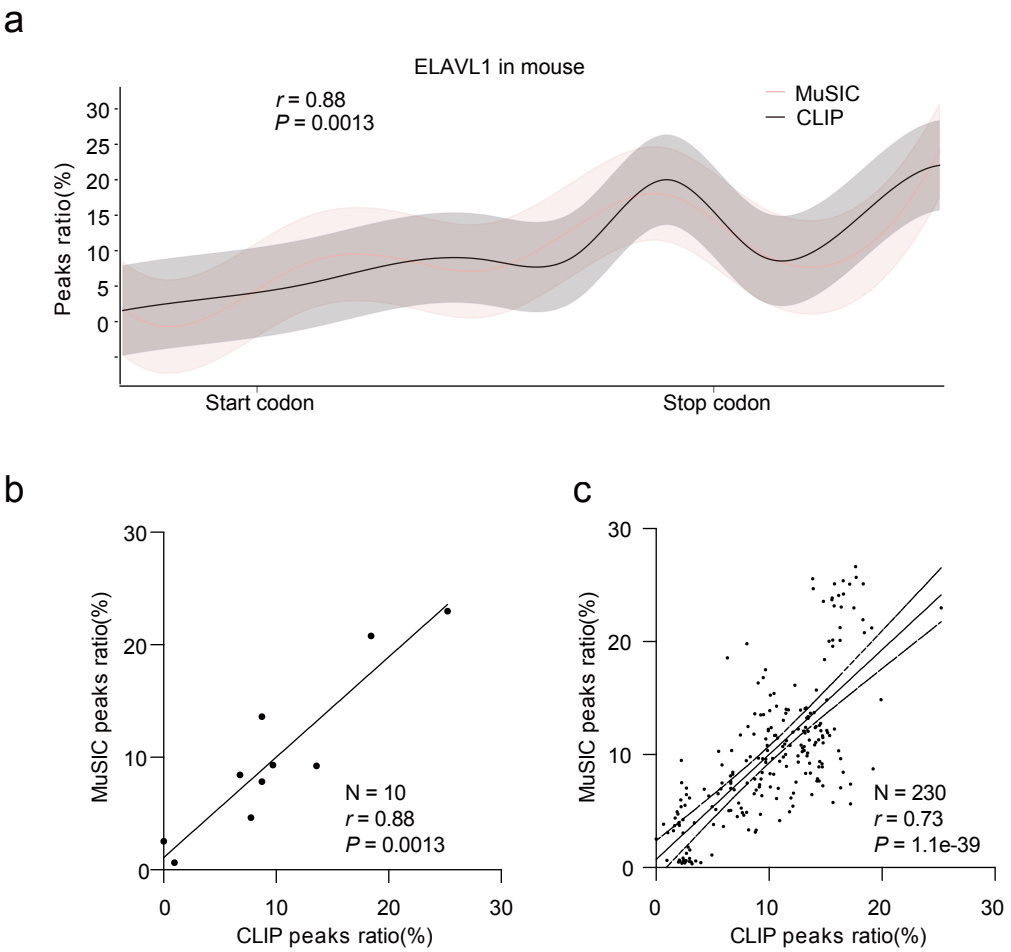

a
ELAVL1 in mouse
r = 0.88
P = 0.0013
MuSIC
CLIP
Peaks ratio(%)
30
25
20
15
10
5
0
Start codon
Stop codon
b
MuSIC peaks ratio(%)
CLIP peaks ratio(%)
N = 10
r = 0.88
P = 0.0013
0
10
20
30
c
MuSIC peaks ratio(%)
CLIP peaks ratio(%)
N = 230
r = 0.73
P = 1.1e-39
0
10
20
30

Figure S7

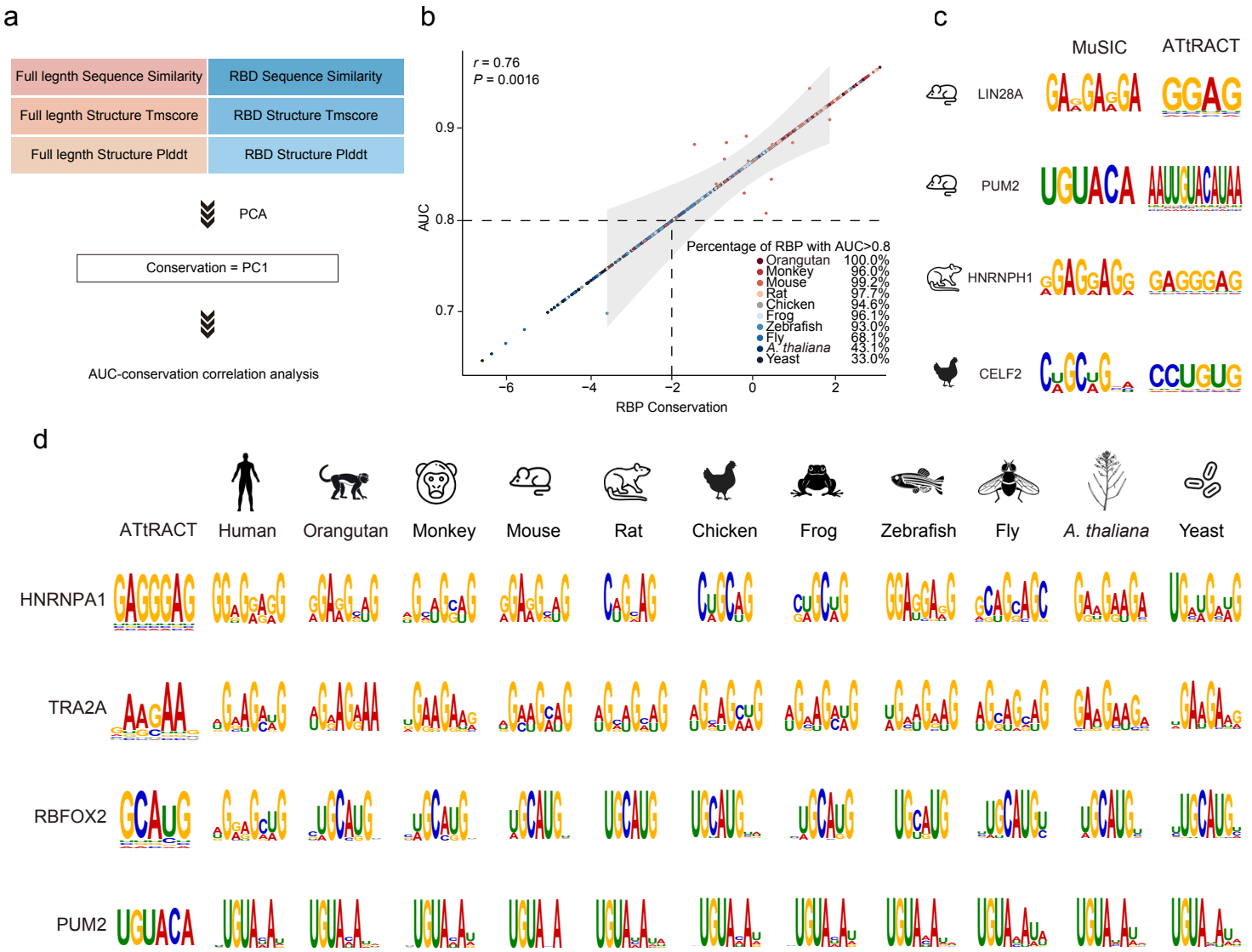

a
Full legnth Sequence Similarity
RBD Sequence Similarity
Full legnth Structure Tmscore
RBD Structure Tmscore
Full legnth Structure Plddt
RBD Structure Plddt
PCA
Conservation = PC1
AUC-conservation correlation analysis
b
r = 0.76
P = 0.0016
AUC
RBP Conservation
Percentage of RBP with AUC>0.8
Orangutan 100.0%
Monkey 96.0%
Mouse 99.2%
Rat 97.7%
Chicken 94.6%
Frog 96.1%
Zebrafish 93.0%
Fly 68.1%
A. thaliana 43.1%
Yeast 33.0%
c
MuSIC
ATtRACT
LIN28A
PUM2
HNRNPH1
CELF2
d
ATtRACT
Human
Orangutan
Monkey
Mouse
Rat
Chicken
Frog
Zebrafish
Fly
A. thaliana
Yeast
HNRNPA1
TRA2A
RBFOX2
PUM2

Figure S8

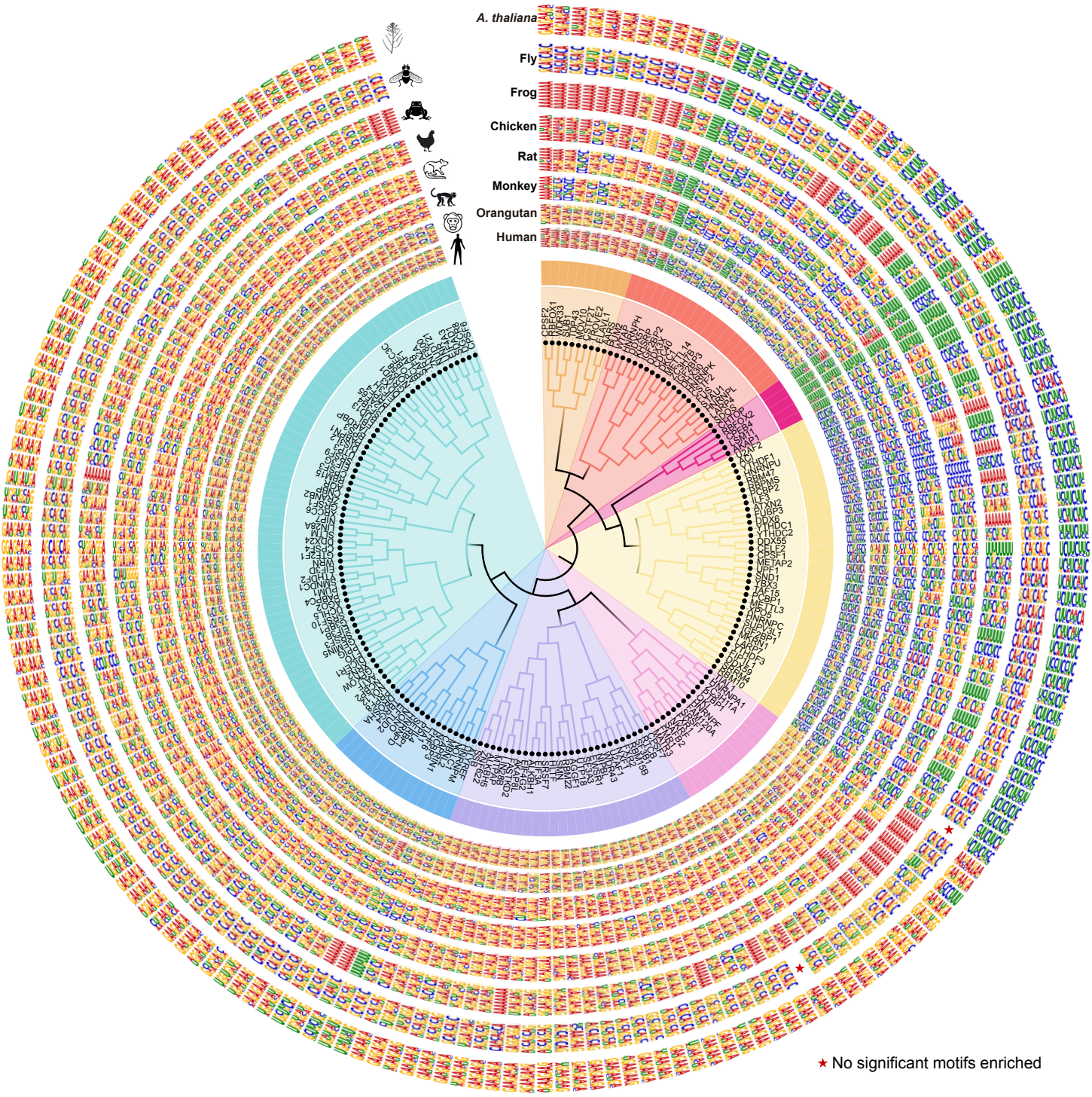

A. thaliana
Fly
Frog
Chicken
Rat
Monkey
Orangutan
Human
No significant motifs enriched

Figure S9

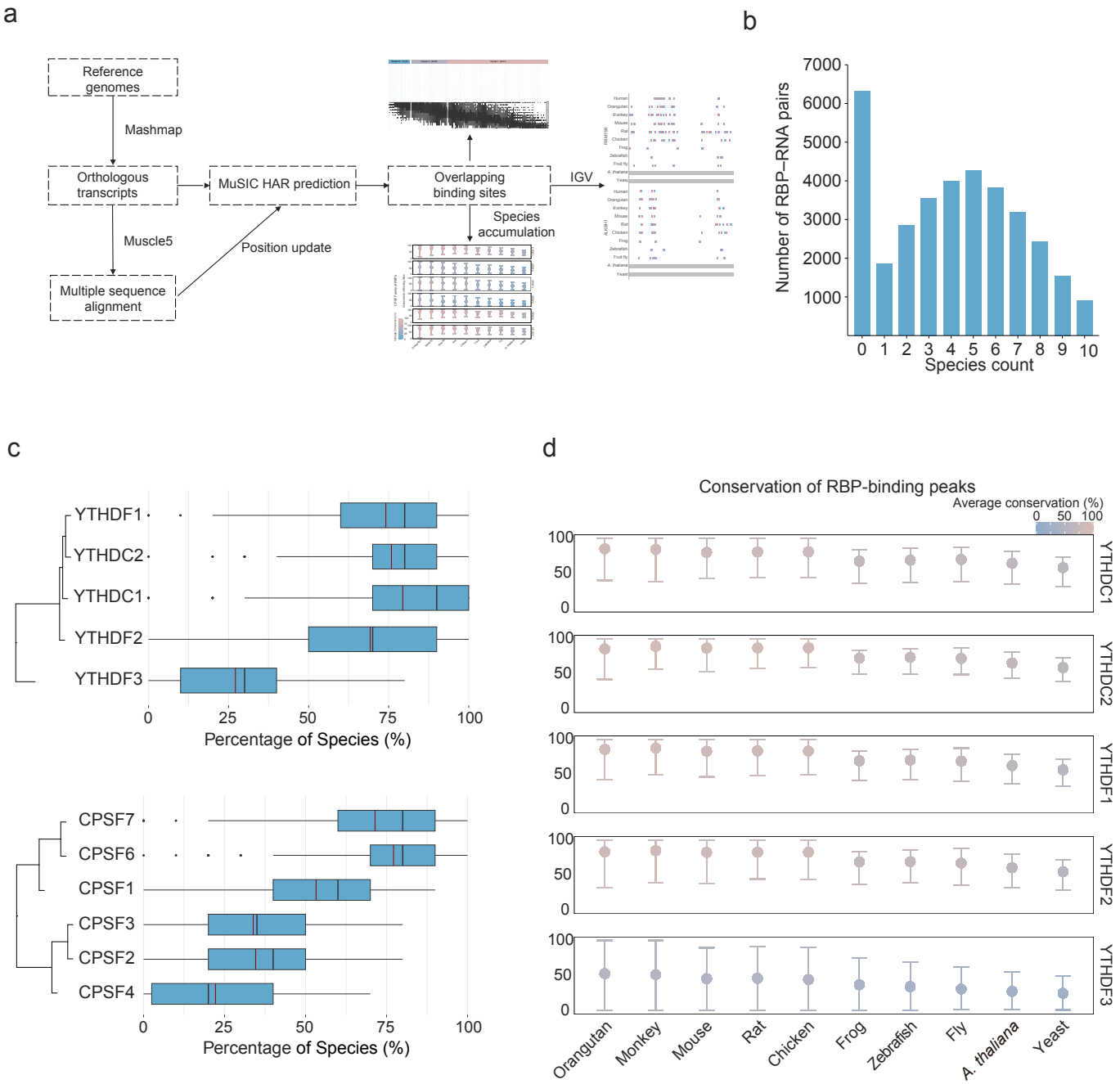
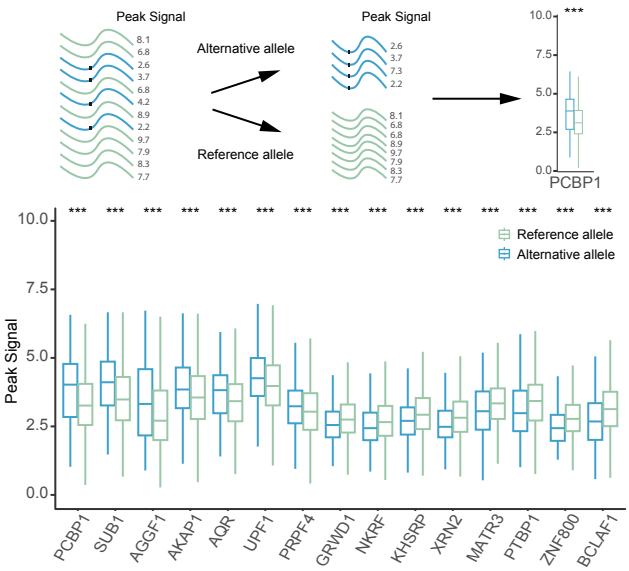

a
Reference genomes
Mashmap
Orthologous transcripts
MuSIC HAR prediction
Overlapping binding sites
IGV
Muscle5
Position update
Multiple sequence alignment
Species accumulation
b
Number of RBP–RNA pairs
7000
6000
5000
4000
3000
2000
1000
0 1 2 3 4 5 6 7 8 9 10
Species count
c
YTHDF1
YTHDC2
YTHDC1
YTHDF2
YTHDF3
0 25 50 75 100
Percentage of Species (%)
CPSF7
CPSF6
CPSF1
CPSF3
CPSF2
CPSF4
0 25 50 75 100
Percentage of Species (%)
d
Conservation of RBP-binding peaks
Average conservation (%)
0 50 100
YTHDC1
YTHDC2
YTHDF1
YTHDF2
YTHDF3
100
50
0
Orangutan
Monkey
Mouse
Rat
Chicken
Frog
Zebrafish
Fly
A. thaliana
Yeast

Figure S10

a



b

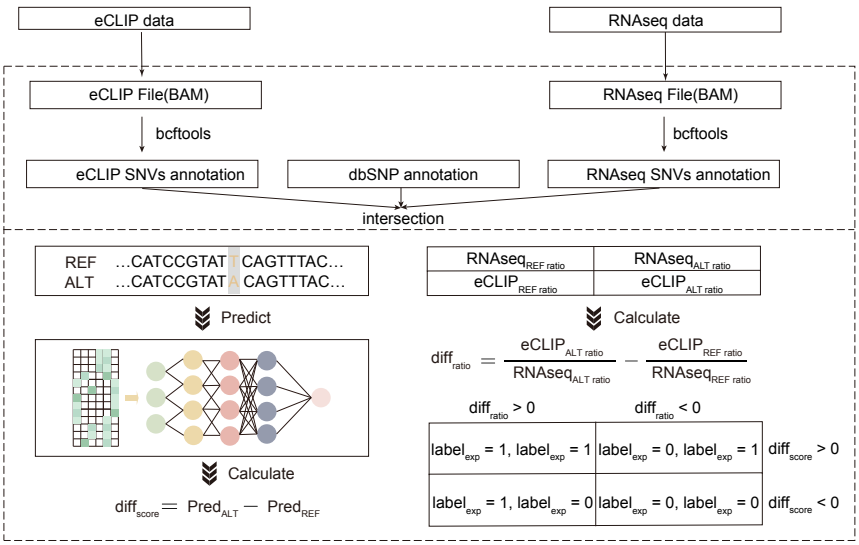


Figure S11

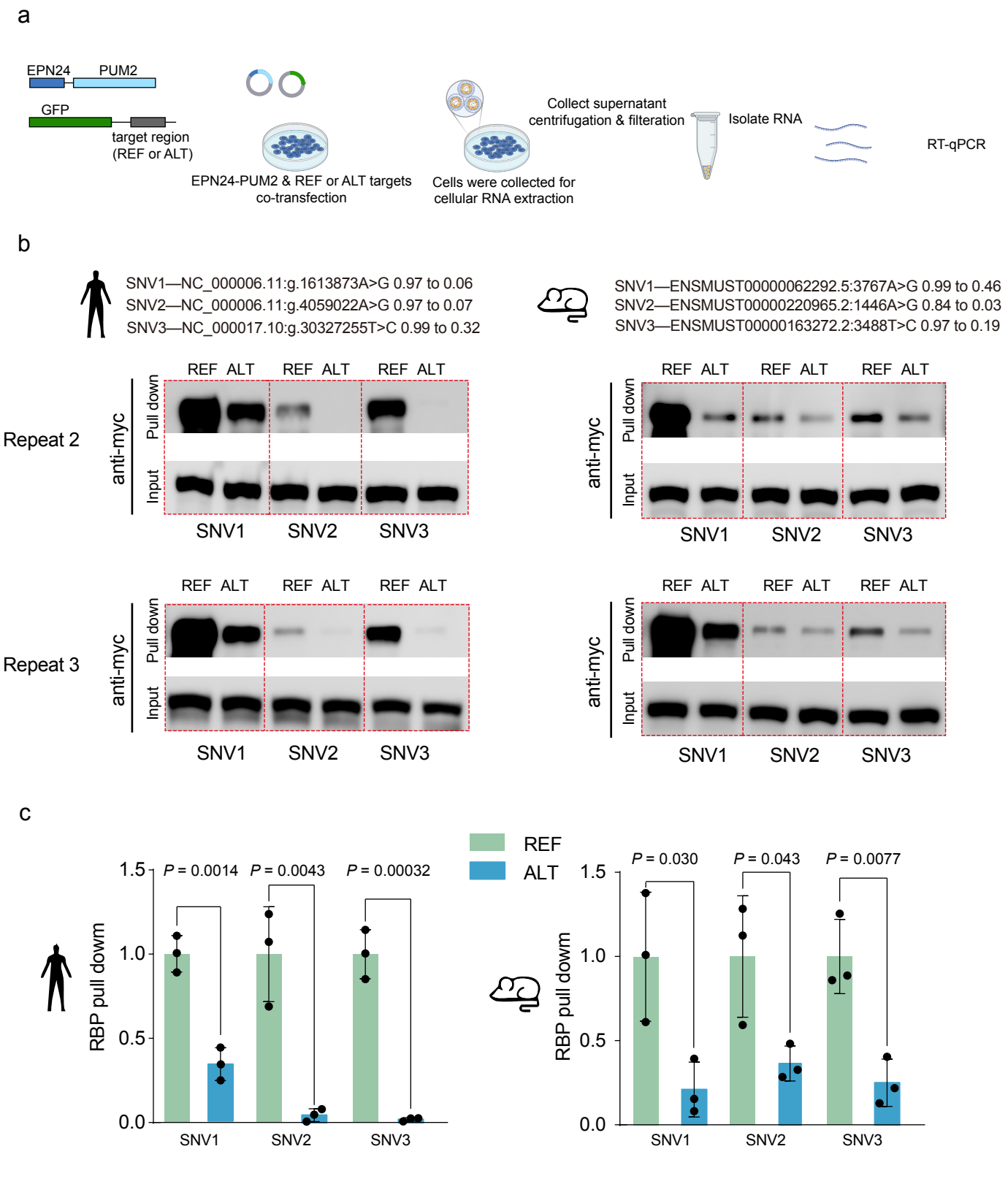

a
EPN24
PUM2
GFP
target region
(REF or ALT)
EPN24-PUM2 & REF or ALT targets
co-transfection
Cells were collected for
cellular RNA extraction
Collect supernatant
centrifugation & filteration
Isolate RNA
RT-qPCR
b
SNV1—NC_000006.11:g.1613873A>G 0.97 to 0.06
SNV2—NC_000006.11:g.4059022A>G 0.97 to 0.07
SNV3—NC_000017.10:g.30327255T>C 0.99 to 0.32
SNV1—ENSMUST00000062292.5:3767A>G 0.99 to 0.46
SNV2—ENSMUST00000220965.2:1446A>G 0.84 to 0.03
SNV3—ENSMUST00000163272.2:3488T>C 0.97 to 0.19
Repeat 2
Repeat 3
anti-myc
Pull down
Input
REF ALT
SNV1
SNV2
SNV3
c
REF
ALT
RBP pull down
P = 0.0014
P = 0.0043
P = 0.00032
P = 0.030
P = 0.043
P = 0.0077
1.5
1.0
0.5
0.0

Figure S12

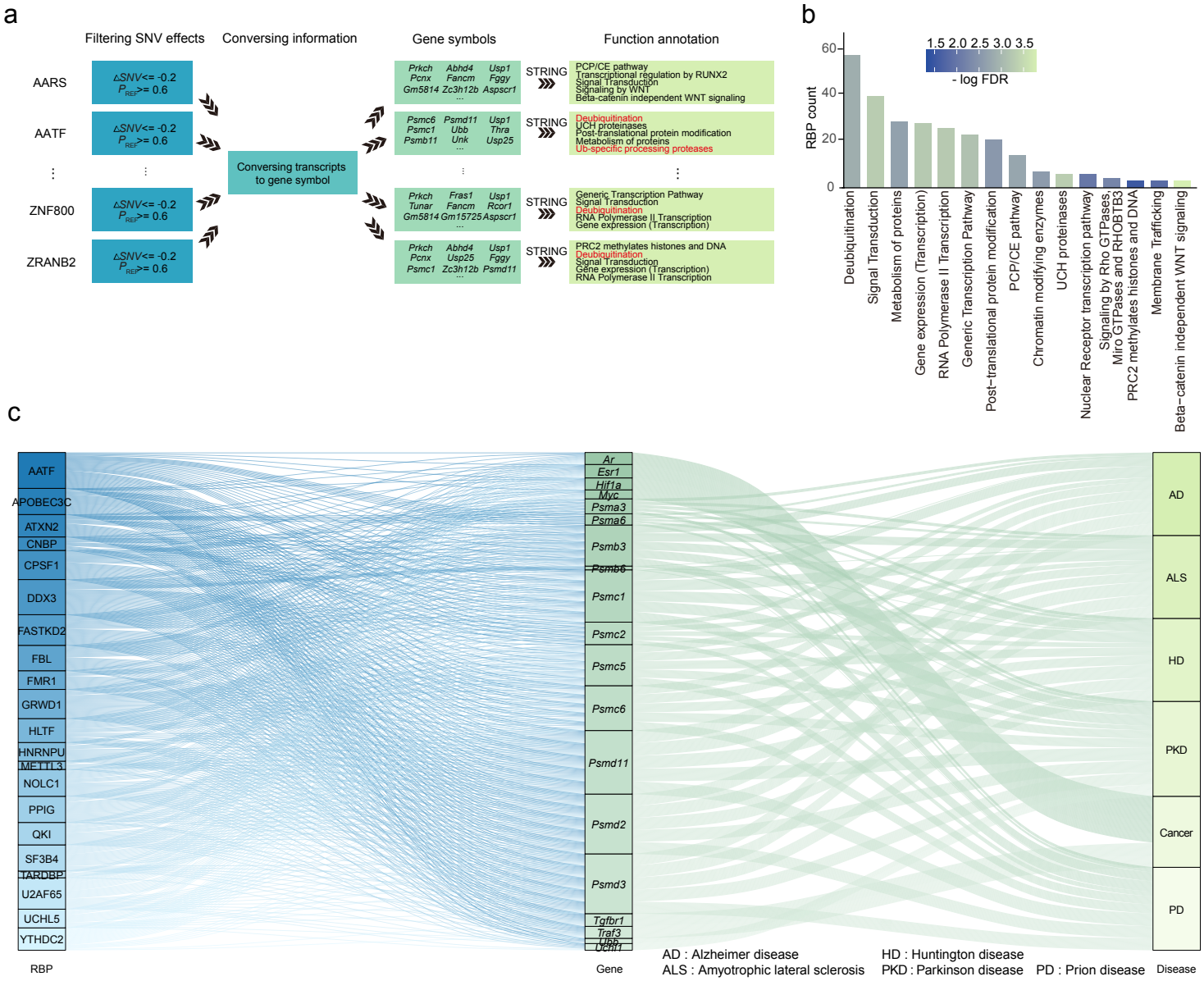
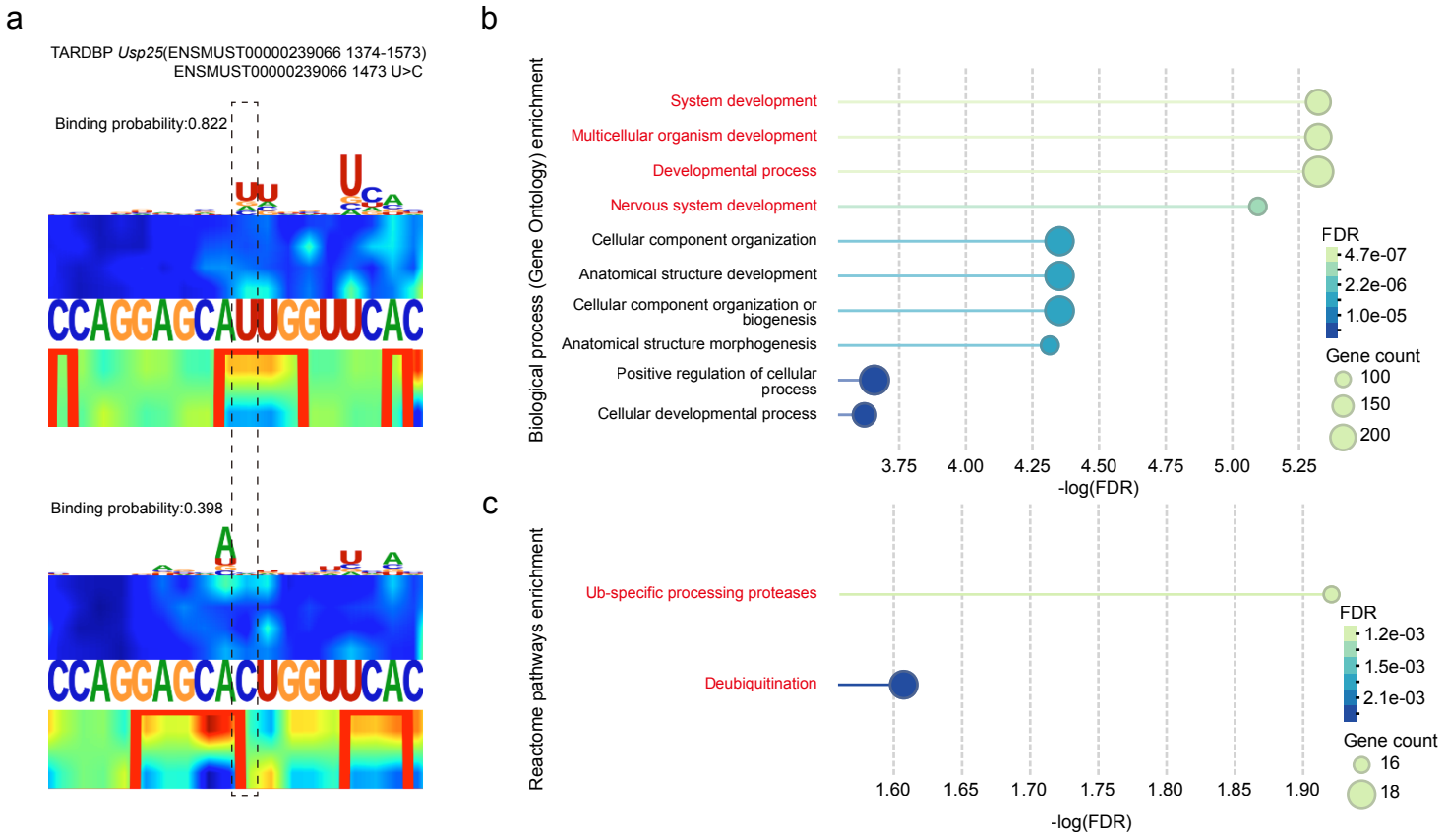

a
Filtering SNV effects
Conversing information
Gene symbols
Function annotation
AARS
AATF
ZNF800
ZRANB2
ΔSNV<= -0.2
P_REF>= 0.6
Conversing transcripts to gene symbol
Prkch Abhd4 Usp1 Pcnx Fancm Fggy Gm5814 Zc3h12b Aspscr1 ...
STRING
PCP/CE pathway
Transcriptional regulation by RUNX2
Signal Transduction
Signaling by WNT
Beta-catenin independent WNT signaling
Psmc6 Psmd11 Usp1 Psmc1 Ubb Thra Psmb11 Unk Usp25 ...
Deubiquitination
UCH proteinases
Post-translational protein modification
Metabolism of proteins
Ub-specific processing proteases
Prkch Fras1 Usp1 Tunar Fancm Rcor1 Gm5814 Gm15725 Aspscr1 ...
Generic Transcription Pathway
Signal Transduction
Deubiquitination
RNA Polymerase II Transcription
Gene expression (Transcription)
Prkch Abhd4 Usp1 Pcnx Usp25 Fggy Psmc1 Zc3h12b Psmd11 ...
PRC2 methylates histones and DNA
Deubiquitination
Signal Transduction
Gene expression (Transcription)
RNA Polymerase II Transcription
b
RBP count
1.5 2.0 2.5 3.0 3.5
- log FDR
Deubiquitination
Signal Transduction
Metabolism of proteins
Gene expression (Transcription)
RNA Polymerase II Transcription
Generic Transcription Pathway
Post−translational protein modification
PCP/CE pathway
Chromatin modifying enzymes
UCH proteinases
Nuclear Receptor transcription pathway
Signaling by Rho GTPases, Miro GTPases and RHOBTB3
PRC2 methylates histones and DNA
Membrane Trafficking
Beta−catenin independent WNT signaling
c
AATF
APOBEC3C
ATXN2
CNBP
CPSF1
DDX3
FASTKD2
FBL
FMR1
GRWD1
HLTF
HNRNPU
METTL3
NOLC1
PPIG
QKI
SF3B4
TARDBP
U2AF65
UCHL5
YTHDC2
RBP
Ar
Esr1
Hif1a
Myc
Psma3
Psma6
Psmb3
Psmb6
Psmc1
Psmc2
Psmc5
Psmc6
Psmd11
Psmd2
Psmd3
Tgfbr1
Traf3
Ubb
Uchl1
Gene
AD
ALS
HD
PKD
Cancer
PD
Disease
AD : Alzheimer disease
ALS : Amyotrophic lateral sclerosis
HD : Huntington disease
PKD : Parkinson disease
PD : Prion disease

Figure S13


a
TARDBP Usp25(ENSMUST00000239066 1374-1573)
ENSMUST00000239066 1473 U>C
Binding probability:0.822
Binding probability:0.398
b
Biological process (Gene Ontology) enrichment
System development
Multicellular organism development
Developmental process
Nervous system development
Cellular component organization
Anatomical structure development
Cellular component organization or biogenesis
Anatomical structure morphogenesis
Positive regulation of cellular process
Cellular developmental process
3.75
4.00
4.25
4.50
4.75
5.00
5.25
-log(FDR)
FDR
4.7e-07
2.2e-06
1.0e-05
Gene count
100
150
200
c
Reactome pathways enrichment
Ub-specific processing proteases
Deubiquitination
1.60
1.65
1.70
1.75
1.80
1.85
1.90
-log(FDR)
FDR
1.2e-03
1.5e-03
2.1e-03
Gene count
16
18