

An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics

Brendan Frey (✉ frey@deepgenomics.com)

Deep Genomics

Albi Celaj

Deep Genomics <https://orcid.org/0000-0002-5888-772X>

Alice Gao

Deep Genomics

Tammy Lau

Deep Genomics

Erle Holgersen

Deep Genomics

Alston Lo

Deep Genomics

Varun Lodaya

Deep Genomics <https://orcid.org/0000-0003-2746-000X>

Christopher Cole

Deep Genomics <https://orcid.org/0000-0002-6733-633X>

Robert Denroche

Deep Genomics

Carl Spickett

Deep Genomics

Omar Wagih

Deep Genomics

Pedro Pinheiro

Deep Genomics

Parth Vora

Deep Genomics

Pedrum Mohammadi-Shemirani

Deep Genomics

Steve Chan

Deep Genomics

Zach Nussbaum

Deep Genomics

Xi Zhang

Deep Genomics <https://orcid.org/0000-0001-6485-4564>

Helen Zhu

University of Toronto

Easwaran Ramamurthy

Deep Genomics

Bhargav Kanuparthi

Deep Genomics

Michael Iacocca

Deep Genomics

Diane Ly

Deep Genomics

Ken Kron

Deep Genomics

Marta Verby

Deep Genomics

Kahlin Cheung-Ong

Deep Genomics

Zvi Shalev

Deep Genomics

Brandon Vaz

Deep Genomics

Sakshi Bhargava

Deep Genomics

Farhan Yusuf

Deep Genomics

Sharon Samuel

Deep Genomics

Sabriyeh Alibai

Deep Genomics

Zahra Baghestani

Deep Genomics

Xinwen He

Deep Genomics

Kirsten Krastel

Deep Genomics

Oladipo Oladapo

Deep Genomics

Amrudha Mohan

Deep Genomics

Arathi Shanavas

Deep Genomics

Magdalena Bugno

Deep Genomics

Jovanka Bogojeski

Deep Genomics

Frank Schmitges

Deep Genomics

Carolyn Kim

Deep Genomics

Solomon Grant

Deep Genomics

Rachana Jayaraman

Deep Genomics

Tehmina Masud

Deep Genomics

Amit Deshwar

Deep Genomics

Shreshth Gandhi

Deep Genomics

Biological Sciences - Article

Keywords:

Posted Date: September 25th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3373630/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: **Yes** there is potential Competing Interest. All listed authors are present or past employees of Deep Genomics Inc. This study received funding from Deep Genomics in the form of salary support and covering of computational costs. The founder was involved in the decision to submit for publication.

1 An RNA foundation model enables discovery of 2 disease mechanisms and candidate therapeutics

3

4 Albi Celaj, Alice Jiexin Gao, Tammy T.Y. Lau, Erle M. Holgersen, Alston Lo, Varun Lodaya,
5 Christopher B. Cole, Robert E. Denroche, Carl Spickett, Omar Wagih, Pedro O. Pinheiro,
6 Parth Vora, Pedrum Mohammadi-Shemirani, Steve Chan, Zach Nussbaum, Xi Zhang,
7 Helen Zhu, Easwaran Ramamurthy, Bhargav Kanuparthi, Michael Iacocca, Diane Ly, Ken
8 Kron, Marta Verby, Kahlin Cheung-Ong, Zvi Shalev, Brandon Vaz, Sakshi Bhargava,
9 Farhan Yusuf, Sharon Samuel, Sabriyeh Alibai, Zahra Baghestani, Xinwen He, Kirsten
10 Krastel, Oladipo Oladapo, Amrudha Mohan, Arathi Shanavas, Magdalena Bugno,
11 Jovanka Bogojeski, Frank Schmitges, Carolyn Kim, Solomon Grant, Rachana Jayaraman,
12 Tehmina Masud*, Amit Deshwar*, Shreshth Gandhi*, Brendan J. Frey*

13

14 Abstract

15 Accurately modeling and predicting RNA biology has been a long-standing challenge,
16 bearing significant clinical ramifications for variant interpretation and the formulation of
17 tailored therapeutics. We describe a foundation model for RNA biology, “BigRNA”, which
18 was trained on thousands of genome-matched datasets to predict tissue-specific RNA
19 expression, splicing, microRNA sites, and RNA binding protein specificity from DNA
20 sequence. Unlike approaches that are restricted to missense variants, BigRNA can
21 identify pathogenic non-coding variant effects across diverse mechanisms, including
22 polyadenylation, exon skipping and intron retention. BigRNA accurately predicted the
23 effects of steric blocking oligonucleotides (SBOs) on increasing the expression of 4 out
24 of 4 genes, and on splicing for 18 out of 18 exons across 14 genes, including those
25 involved in Wilson disease and spinal muscular atrophy. We anticipate that BigRNA and
26 foundation models like it will have widespread applications in the field of personalized
27 RNA therapeutics.

28 Main

29 Building machine learning models that can predict gene expression from DNA sequence
30 has been a long-standing research goal¹, and one that has seen significant strides
31 owing to recent advancements in deep learning². These models could revolutionize
32 drug discovery by pinpointing how pathogenic genetic variants alter gene expression
33 and gene processing, and by designing customized drug candidates to counteract these
34 effects³. Currently, most efforts have focused on predicting data that measures overall

35 gene expression levels^{2,4}, which are not suited to predicting regulatory interventions, for
36 example, specific transcriptional perturbations on splicing or polyadenylation.

37

38 RNA sequencing (RNA-seq) data provides a widely-available resource for measuring
39 RNA expression at high resolution and capturing complex transcriptional regulation
40 events across diverse genotypes. This includes both exome variation inherently coded
41 within RNA-seq data itself, and through extensive resources like the Genotype-Tissue
42 Expression⁵ (GTEx) project that pairs RNA-seq with Whole Genome Sequencing (WGS).
43 While building deep neural networks that directly learn from RNA-seq offers the
44 opportunity to understand how changes in DNA sequence lead to changes in complex
45 transcriptional phenotypes, this goal has remained elusive.

46

47 We introduce “BigRNA”, a deep learning model that is directly trained on RNA-seq
48 datasets. BigRNA learns from paired genotype and 128bp resolution RNA expression
49 data from many individuals, and can also be applied in a range of downstream tasks
50 such as predicting RNA-binding protein (RBP) specificity and microRNA binding sites.
51 Because BigRNA directly models RNA-seq data, it can discover a diverse set of
52 pathogenic non-coding mechanisms that would each require a specialized model, and
53 can pinpoint their effects on a transcript. We show that BigRNA can discover the
54 effects of non-coding variants on expression and splicing, and matches or exceeds the
55 performance of specialized models in recovering known pathogenic variants.

56

57 BigRNA can also help design different types of RNA based therapeutics, including steric
58 blocking oligonucleotides (SBOs). Without any additional training, BigRNA accurately
59 identifies compounds that induce a targeted splicing change, and recovers known
60 approved SBO therapies with high specificity. The ability of BigRNA to understand
61 regulatory mechanisms also allows it to design SBOs that block predicted inhibitory
62 regions to increase the expression of a disease gene. BigRNA represents a new
63 generation of massive deep learning models that can be applied to a range of different
64 personalized RNA therapeutic discovery tasks.

65 Results

66 **BigRNA accurately predicts tissue-specific RNA expression and the binding sites of** 67 **proteins and microRNAs**

68

69 To train BigRNA to predict RNA-seq data from the corresponding DNA sequence, we
70 employed a transformer-based architecture² and utilized the GTEx⁵ resource (**Methods**).
71 Given an individual's genotype, we input two potential haplotypes independently into

72 identical instances of the model, and train it to predict the observed RNA-seq data as
73 the combined output from these haplotypes ([Fig. 1a](#), [Supplementary Figs. S1 and S2](#)).
74 Each output "head" of the model predicts the expression of a single GTEx sample, so
75 that it learns to predict the outputs of 2,956 RNA-seq samples from 70 individuals,
76 covering 51 tissues in total. After training on these RNA-seq datasets, the model is
77 fine-tuned to predict the specificity of RBP and microRNA binding sites ([Fig. 1a](#)).

78

79 We first evaluated the ability of BigRNA to predict the expression of unseen genomic
80 sequences. We measured the model's ability to predict tissue-specific expression levels
81 for all genes outside of genomic regions in the training set. BigRNA exhibited strong
82 performance for predicting expression levels of unseen genes, achieving a correlation
83 coefficient (r) between 0.47 and 0.77 across all tissues (mean=0.70, [Fig. 1b](#)). We
84 observed slightly stronger performance in brain tissues than non-brain tissues (mean
85 $r=0.74$ versus 0.69, $p=5e-03$), and highlight that the model is able to accurately predict
86 expression levels in the hypothalamus ($r=0.74$, [Fig. 1c](#)). The ability to predict overall
87 expression levels and capability to accurately delineate intron/exon junctions is
88 illustrated by BigRNA's predictions for *SLC7A8*, an amino acid transporter within the test
89 set ([Fig. 1d](#)). To evaluate BigRNA on the much harder task of predicting differences
90 between pairs of tissues, we used BigRNA predictions to compute the fold-change in
91 total exonic coverage between tissue pairs and compared that to observed
92 fold-changes. Across all inter-tissue comparisons, we observed a mean correlation of
93 $r=0.4$, owing to the increased difficulty of this task ([Fig. 1e](#)). We highlight a comparison
94 between liver and the hypothalamus ($r=0.58$, $p=7e-64$, [Fig. 1f](#)) to illustrate this capability.

95

96 Since drug discovery tasks benefit from clarity of mechanisms, we next examined how
97 well the fine-tuned BigRNA model could predict RBP binding specificity and microRNA
98 binding sites. For the RBP task, we used a large-scale resource of transcriptome-wide
99 binding profiles for 223 datasets covering 150 unique human RBPs in K562 and HepG2
100 cells⁶. We found that BigRNA achieved high average precision for many RBPs and
101 performed better than the previously-published DeepRiPe⁷ system for all 142 datasets
102 that they had in common ([Fig. 1g](#)). On predicting microRNA binding sites, BigRNA
103 achieved a median AUC of 0.84 and for all 12 cell lines that we tested, performed better
104 than a previously published method, TargetScan⁸ ([Fig. 1h](#)). These predictions are useful
105 for identifying regulatory factors that are altered by variants and SBOs (see below).

106

107 **Predicting the effects of variants on gene expression**

108

109 A key challenge in human genetics is to predict the impact of sequence variants that
110 may be found within the human population. Many deep learning models that do well on

111 unseen genes using certain metrics, such as AlphaFold⁹, struggle to predict variant
112 effects¹⁰. While some accurate methods exist for predicting the pathogenic impact of
113 rare missense variants^{11,12}, non-coding variants, such as those located within the 3' and
114 5' untranslated regions (UTRs) of genes, remain difficult to interpret.

115

116 To address this gap, we evaluate BigRNA's ability to predict the impact of a curated set
117 of pathogenic or likely pathogenic (P/LP) UTR variants from ClinVar¹³. We found that
118 BigRNA exhibited strong performance as a general pathogenicity model for variants in
119 both the 3' UTR and 5' UTR (AUC=0.95 and 0.8, [Fig. 2a](#)) by predicting their effects on
120 the expression of their associated disease genes. The weaker performance in the 5'
121 UTR may be due to a smaller proportion of P/LP variants that modulate RNA expression
122 (18/47 compared to 16/17 for the 3' UTR amongst variants with known mechanisms,
123 $p=0.046$), and a substantial proportion of mechanisms that affect translation (29/47).
124 We further investigated a known pathogenic expression-decreasing variant in the 3' UTR
125 of *NAA10*¹⁴ (NM_003491.4:c.*43A>G). This variant is known to cause syndromic
126 X-linked microphthalmia, and reduces expression by disrupting the polyadenylation site
127 (PAS) of the *NAA10* transcript. The BigRNA predictions highlight the
128 expression-decreasing effects of this variant (false positive rate, FPR <0.5%), and also
129 predicted the expected lengthening of the 3' UTR that was observed in RNA-seq
130 samples of affected patients¹⁴ ([Fig. 2b](#)). An *in-silico* saturation mutagenesis near this
131 variant highlighted the importance of the PAS, and confirmed the effects of two other
132 nearby P/LP variants (c.*39A>G, c.*40A>G)¹⁴ ([Fig. 2c](#)).

133

134 We compared BigRNA to Framepool¹⁵, a ribosomal load model, Saluki¹⁶, an RNA stability
135 model, and Enformer², an expression model that learns from CAGE-seq. We observed
136 improved performance compared to Enformer for pathogenic variants in both the 5' and
137 3' UTR ($p=0.04$ and $p=0.02$, respectively, **Supplementary Fig. S3**). Framepool, a model
138 that predicts ribosomal load¹⁵, performed similarly to BigRNA for pathogenic variant
139 classification in the 5' UTR (AUC=0.67 versus 0.78 for BigRNA, $p=0.07$, **Supplementary**
140 **Fig. S3**), but BigRNA performed better at classifying the subset of pathogenic 5' UTR
141 variants that are known to modulate RNA expression (AUC=0.61 versus 0.86 for
142 BigRNA, $p=0.002$, **Supplementary Fig. S4**). Saluki, an RNA half-life model, had similar
143 performance on the 3' UTR task (AUC=0.87 vs 0.94 for BigRNA, $p=0.27$).

144

145 Within these genes, we noted many variants of uncertain significance (VUS) in their
146 untranslated regions. Applying BigRNA to these variants at a 5% FPR yielded 12
147 potential expression-modulating variants in the 3' UTR (out a total of 139) and 23 in the
148 5' UTR (out a total of 222) ([Fig. 2d](#)). For example, the 3' UTR of *HBB* had the highest
149 number of VUSs surpassing this threshold ($n=6$). The highest scoring VUS

150 (NM_000518.5(HBB):c.*112A>T) is in the PAS of this gene, and shares the same
151 position as a known pathogenic variant (c.*112A>G). The PAS region of *HBB* also
152 contains the majority of known P/LP variants (6 of 8). The second-highest scoring VUS
153 (c.*47C>G) was outside of the PAS, and less is known about its function. Looking
154 further, we found that despite being classified as a VUS, this variant is reported to cause
155 decreased expression of *HBB*, supporting the BigRNA prediction¹⁷. We also noted that
156 three additional P/LP variants in the *HBB* PAS, which were not included in our
157 benchmark due to a lack of evidence in the ClinVar submission¹³, scored above this
158 threshold ([Fig. 2e](#)), providing computational support for their P/LP classification.

159

160 In more genetically complex diseases, it can be challenging to discover causal
161 expression-modulating variant(s) due to linkage disequilibrium (LD). For example,
162 rs705379 and rs854572 are both annotated as expression quantitative trait loci (eQTLs)
163 for Paraoxonase 1 (*PON1*) in GTEx, but a luciferase reporter assay and statistical
164 fine-mapping of the locus show that only rs705379 has an effect on expression^{18,19},
165 which is consistent with BigRNA's prediction of a much stronger effect and its direction,
166 despite the strong LD. BigRNA also assigned a stronger effect, and correct direction, for
167 two other known expression modulation variants, rs854571 and rs3735590²⁰ ([Fig. 2f](#)).

168 To benchmark BigRNA more broadly, we evaluated its ability to identify fine-mapped
169 eQTLs from negative controls matched on effector gene (eGene), distance to
170 transcription start site (TSS), and minor allele frequency. We saw considerable
171 performance for this task (AUC = 0.74, [Fig. 2g](#)), improving over Enformer (AUC = 0.70,
172 $p=4.8e-04$ for difference, **Supplementary Fig. S5**). We note that a series of
173 improvements in eQTL scoring, including matching the predictions to the eQTL tissue of
174 interest, and evaluating over the entire contiguous coding sequence rather than the
175 transcription start site made significant improvements to our performance for both
176 models (**Supplemental Note 1**). BigRNA's classification performance was similarly
177 strong for variants more than 10 kilobases from their eGene's TSS (AUC 0.73, versus
178 0.66 for Enformer, $p=8.0e-05$ for difference, **Supplementary Fig. S6**). Together, these
179 results indicate that BigRNA is able to help prioritize causal variants that mediate more
180 common diseases, which has been challenging for sequence-based deep neural
181 networks^{18,19}.

182

183 **Predicting the effects of variants on splicing and intron retention**

184

185 An important subset of pathogenic variants affect splicing, such as those which cause
186 skipping of an exon. These variants often occur in coding regions, and may be
187 incorrectly classified as benign mutations based on their amino acid substitutions,
188 despite their pathogenic splicing effects²¹. We evaluated BigRNA's ability to classify the

189 splicing impact of exonic variants that cause substantial (>50%) exon skipping, versus
190 those that do not cause any splicing changes, using results from a massively parallel
191 splicing assay (MaPSy)²¹. By predicting a change in junction coverage caused by these
192 variants, BigRNA was able to accurately predict these skipping variants (AUC = 0.89 [Fig.](#)
193 [3a](#)), and showed better performance compared to a previously published method,
194 SpliceAI²² on this task (AUC=0.80, $p < 1e-05$ for difference, **Supplementary Fig. S7**). We
195 further investigated a pathogenic variant that causes skipping of exon 6 in the *ACADM*
196 gene, leading to a potentially fatal medium-chain acyl-CoA dehydrogenase
197 deficiency^{23,24}. BigRNA predicted the exon skipping effects of this variant (FPR = 0.002,
198 [Fig. 3b](#)), and that it causes this skipping by creating a binding site for the TDP-43
199 protein²³, yielding insight into the mechanism-of-action. We further investigated a VUS
200 in *ATP7B* (c.3243+5G>A), a gene which clears copper from liver cells and causes Wilson
201 disease when it is defective²⁵. This variant was predicted by BigRNA to cause in-frame
202 skipping of *ATP7B* exon 14 (FPR=0.004, [Fig. 3c](#)), which contains the ATP site and other
203 critical elements²⁶, thus causing a pathogenic loss-of-function. We generated a
204 homozygous HepG2 line and used RT-PCR to assay the effects of this variant and
205 confirm the exon skipping predicted by BigRNA ([Fig. 3c](#)).

206

207 Another class of pathogenic splicing variants are cryptic splicing mutations that cause
208 full intron retention. We evaluated BigRNA on its ability to predict a set of reported
209 intron retention variants²⁷, using nearby common variants as the negative set. We
210 observed strong performance on classifying these mutations (AUC=0.9, [Fig. 3d](#) and
211 **Supplementary Fig. S8**), so we next investigated whether BigRNA could predict more
212 complex splicing aberrations. We focused our attention on a pathogenic non-canonical
213 splice site variant in the *ABCA4* gene (c.5714+5G>A), which had been found to induce
214 Stargardt disease by causing skipping of *ABCA4* exon 40²⁸. This variant was strongly
215 predicted to cause both the skipping of exon 40, and retention of intron 40 (FPR=0.008
216 and <0.04, respectively, [Fig. 3e](#)), but the latter had not been reported, likely due to
217 technical limitations in the assay²⁸. To test this prediction, we edited a retinoblastoma
218 cell line (WERI-Rb-1) to be homozygous for c.5714+5G>A, and performed RNA
219 sequencing to capture the full suite of splicing events. This confirmed BigRNA's
220 predictions that this variant causes a complex set of aberrations that includes partial
221 skipping of exon 40, as well as retention of intron 40.

222

223

224 **Designing splice-switching and expression-increase molecules**

225 The ability of BigRNA to understand regulatory mechanisms affecting splicing and gene
226 expression may allow it to design therapeutic interventions that rescue pathogenic
227 variant effects. For this application, we evaluated whether BigRNA could reverse

228 splicing defects by designing steric blocking oligonucleotides (SBOs) – short,
229 chemically-modified synthetic nucleic acid strands purposed to bind specific RNA
230 targets to modulate splicing and gene expression. For example, Nusinersen, an FDA
231 approved SBO, treats spinal muscular atrophy by reversing the skipping of exon 7 in
232 *SMN2*²⁹, thus restoring SMN protein levels and mitigating motor neuron loss and
233 muscular atrophy. One way to predict the effect of an SBO is to hide the complementary
234 binding site from the model's input (**Methods**). This approach is an instance of
235 'zero-shot learning', because no additional task-specific SBO data is used when making
236 the prediction.

237

238 To evaluate the utility of zero-shot learning for virtual screening, we first evaluated the
239 ability of BigRNA to re-discover Nusinersen amongst the set of all possible SBOs within
240 200 base-pairs of *SMN2* exon 7. Strikingly, BigRNA ranked Nusinersen within the top 3
241 of 437 compounds (**Fig. 4a**). To more systematically evaluate the effectiveness of this
242 approach, we treated 15 exons in 12 genes with a total of 620 SBOs, and observed a
243 strong and statistically significant correlation with the predicted and
244 experimentally-measured exon inclusion levels in all cases ($r=0.41-0.77$, $p=7e-12$ to $2e-2$,
245 **Fig. 4b**). For comparison, SpliceAI correlated with experiments in 11/15 exons and the
246 correlation was lower than BigRNA for 13/15 exons.

247

248 We then used BigRNA to design a novel splice-switching SBO that rescues a pathogenic
249 splicing defect. Previously, we had reported that a missense variant in the *ATP7B* gene
250 (c.1934T>G, Met645Arg) leads to Wilson disease by promoting skipping of exon 6, thus
251 resulting in lowered levels of functional protein and subsequent copper accumulation in
252 liver cells²⁵. We created a disease model of the Met645Arg variant in HepG2 cells, and
253 used this system to test a set of SBOs targeting the skipped exon (**Methods**). We
254 observed a strong relationship between the predicted and measured splicing changes
255 ($r=0.91$, $p=4.7e-22$, **Fig. 4c**). The top compound from this assay was predicted to be in
256 the top 7 of 458 possible compounds by BigRNA. To summarize, BigRNA predicted both
257 the exon skipping caused by Met645Arg (FPR=0.007) and the restorative effect of the
258 top experimentally-validated compound (**Fig. 4d**).

259

260 BigRNA's ability to score SBOs has utility in developing therapeutic candidates targeting
261 extremely rare variants within a constrained budget. First, we evaluated BigRNA's ability
262 to score SBOs that target a pseudo-exon in the *ATM* gene caused by the rare
263 c.5763-1050A>G mutation, leading to ataxia-telangiectasia³⁰. We observed significant
264 correlation between the predictions and experimentally observed splicing efficiencies
265 ($r=0.64$, $p=3.3e-04$, **Supplementary Fig. S9**), and ranked the lead therapeutic candidate
266 in the top 7 of 516 possible compounds. We sought to explore whether similar

267 therapeutic candidates could be developed for other rare splicing diseases. After
268 curating a set of extremely rare, so-called “N=1”, pathogenic variants from ClinVar
269 (**Methods**), we used BigRNA to predict which ones are likely to act through exon
270 skipping while not affecting the core splice donor or acceptor site (**Fig. 4e**), thus
271 potentially being eligible for SBO remediation. This included synonymous variants,
272 non-synonymous variants predicted to be tolerated³¹, and variants near splice sites.
273 One such variant was in intron 22 of *MYO1E*, which is associated with
274 glomerulosclerosis³². While no published mechanism exists for this variant, it was
275 predicted to cause skipping of exon 23, and the top SBO was predicted to completely
276 rescue this skipping defect, suggesting that this variant is amenable to personalized
277 SBO treatment (**Fig. 4f**).

278

279 Owing to BigRNA’s striking ability to help design splice-switching SBOs, we turned to the
280 more challenging problem of designing SBOs that amplify gene expression. This
281 requires the model to rank all possible compounds targeting any part of the transcript,
282 again without any additional training and additionally with no prior knowledge of
283 inhibitory regions. Due to the greatly increased search space, we first developed a
284 method to score a large number of compounds in a computationally efficient manner.
285 For this, we applied a combination of established saliency mapping techniques^{33,34} to
286 evaluate the contribution of each base pair in a transcript on its expression in a given
287 tissue, and took the minimum contribution score at the SBO binding region as the
288 ‘inhibitory score’ of each compound (**Methods**). We again benchmarked this scoring on
289 Nusinersen, reasoning that the skipping of exon 7 and subsequent nonsense-mediated
290 mRNA decay is a major expression bottleneck. Considering all 26,901 SBOs of length
291 18, Nusinersen ranked in the top 2.28% (**Supplementary Fig. S10**), suggesting that
292 BigRNA’s inhibitory scores can be used to identify inhibitory regions, and that this
293 strategy could have recovered Nusinersen within a tractable screening budget.

294

295 We then sought to systematically assess how well BigRNA could be used to discover
296 novel therapeutically beneficial expression-increasing SBOs. An example is
297 Paraoxonase 1 (*PON1*), where variants that decrease expression of the gene or catalytic
298 activity of the protein have been associated with an increased risk of atherosclerotic
299 cardiovascular disease^{35,36} (ASCVD). In murine models, modulation of *PON1* expression
300 has been shown to directionally influence the risk of ASCVD and related phenotypes^{37–40},
301 thus presenting a compelling opportunity for expression-increasing therapeutics. We
302 used BigRNA to perform large-scale SBO design, experimentally tested the predicted
303 SBOs in primary human hepatocytes, and identified 10 compounds that showed activity
304 for increasing *PON1* expression (**Methods**). By using a liver-specific score to rank all
305 positive compounds, BigRNA showed a strong ability to prioritize expression increase

306 compounds (AUC=0.818, [Fig. 4h](#)). To expand this study, we screened
307 expression-increasing compounds for *ATP7B* (to benefit a broader population beyond
308 Met645Arg), as well as *PRRT2* and *SERPING1*, which may confer therapeutic benefits for
309 benign familial infantile epilepsy⁴¹ and hereditary angioedema⁴². For all three genes,
310 BigRNA's predictions successfully prioritized expression-increasing SBOs without
311 requiring any additional training (AUC=0.72-0.85, [Fig. 4h](#)).

312 Discussion

313 The rapid evolution of computational models in genomics has enabled the use of
314 methods that can learn from large-scale genomics data to predict RNA expression from
315 DNA sequence. Using deep learning to model RNA-seq data and take into account
316 individual genomic sequence variation, we can enable novel and accelerated discovery
317 on several drug discovery tasks.

318

319 When we adapted previously published deep learning systems to the drug discovery
320 tasks that we evaluated, we found that BigRNA performed substantially better overall. It
321 improved significantly over specialized models like TargetScan⁸ and DeepRiPe⁷ for
322 predicting microRNA and RBP binding sites, and was more accurate than SpliceAI²² at
323 identifying exon skipping variants as well as designing splice-switching SBOs. BigRNA
324 could accurately predict pathogenic variants in untranslated regions, matching
325 specialized models for the 5' and 3' UTRs^{15,16}, and improved upon the general-purpose
326 Enformer model². In cases where BigRNA's performance matched an existing model,
327 direct modeling of RNA-seq data had distinct advantages. For example, unlike a
328 previously described ribosomal loading model¹⁵, BigRNA could predict all classes of
329 pathogenic mutations in the 5' UTR, and unlike a model of RNA half life¹⁶, it could predict
330 that a pathogenic variant acts by changing the polyadenylation site, which reduces the
331 half-life. Existing methods for predicting splice donor and acceptor strength²² are
332 unable to identify correlated splicing events, such as intron retention, but we found that
333 BigRNA is able to do so. For complex traits, in contrast to traditional fine-mapping
334 methods that do not provide insight into the mechanistic impact of causal mutations⁴³,
335 BigRNA can make predictions for complex trait heritability contributions from many
336 different mechanisms that do not exert their effect through a change in protein
337 structure.

338

339 The ability of BigRNA to learn mechanisms of RNA regulation is reflected by the fact
340 that it was able to accurately design SBOs that counteract the effects of pathogenic
341 variants or that increase gene expression, without being provided with a single training
342 case of an SBO and its effect. Nonetheless, a further avenue of work would include

343 fine-tuning BigRNA by learning from SBO treatment data, such as from the rich
344 information encoded by SBO-treated RNA-seq samples⁴⁴. Similar approaches can be
345 used for other therapeutic modalities such as predicting the phenotypic effects of
346 induced ADAR (adenosine deaminase acting on RNA) editing so that they confer a
347 similar compensatory effect on splicing or expression⁴⁵, or designing mRNAs that have
348 increased half-life and translation efficiency.

349

350 Several avenues exist to improve the predictive abilities of BigRNA. The 128bp
351 resolution of the model can be improved with additional training resources².
352 Improvements in the speed and scalability of the transformer architecture⁴⁶, coupled
353 with the use of parameterized upsampling⁴⁷ may allow the model to retain a high
354 context size while producing predictions at single base-pair resolution. Training on
355 more individuals could improve generalization across genotypes. While the training
356 procedure takes into account variation from 70 individuals, WGS-paired RNA-seq data is
357 available for many more GTEx samples, and can be supplemented with additional
358 datasets⁴⁸. To take into account such a large amount of data, methods have been
359 developed to prioritize the most informative training points⁴⁹, allowing the training
360 procedure to scale and effectively learn from extremely large datasets. To explore
361 improved prediction of differences between individuals, a contrastive training objective
362 can be used^{50,51,52} and predictions can be made for the difference in expression between
363 two haplotypes⁵³.

364

365 Our results show that different drug discovery tasks can be assisted by deep learning.
366 We believe that BigRNA and deep learning systems like it have the potential to
367 transform the field of RNA therapeutics.

368 Methods

369

370 RNA-seq model training

371 We downloaded and aligned RNA-seq data from the GTEx consortium⁵ V6 release,
372 processing all available data from the set of 70 individuals with the most tissue
373 availability (data from a total of 51 tissues are available, but data availability varies
374 between individuals). Data was processed using an in-house pipeline (Supplementary
375 Information 1.2). Each RNA-seq sample was processed into two data tracks: coverage
376 and junction, where the junction track contains a subset of read counts at splice
377 junctions. To make the data compatible with the 128bp resolution of the model's
378 architecture², we applied 128bp-window average-pooling on coverage tracks, and
379 128bp-window sum-pooling on junction tracks. To incorporate genomic variants from
380 each individual, we re-aligned the RNA-seq data to match the insertions and deletions
381 introduced by each individual's haplotype (Supplementary Information 1.2). BigRNA was
382 trained with a separate output for each sample, so that each output can be
383 independently learned. We trained BigRNA by minimizing differences between
384 prediction from both haplotypes and the observed coverage and junction tracks from
385 RNA-seq (Supplementary Information 1.3, Supplementary Equation S2). In addition to
386 the individual-specific outputs, we also added individual-agnostic per-tissue outputs to
387 encourage the model to learn a mapping from genotype to expected expression (where
388 the expectation is taken across all individuals). Description of all output heads can be
389 found in Supplementary Data 1. Fig. 1a shows the training pipeline. The same
390 procedure was used to train an ensemble of 7 models, varying learning rate, degree of
391 gradient clipping, and the pre-training strategy for each model in the ensemble
392 (Supplementary Information 1.3, Supplementary Table S1). At inference time, to predict
393 on a genomic interval, we used shifted intervals to increase the prediction resolution to
394 64 base pairs, and averaged predictions from both strands (Supplementary Information
395 1.4, Supplementary Fig. S1, Supplementary Fig. S2).

396

397 Fine-tuning on RBP and microRNA datasets

398 After training models on RNA-seq dataset, we further fine-tuned models on RBP and
399 microRNA datasets. The RBP dataset was constructed by downloading eCLIP data⁶
400 from ENCODE⁵⁴ (Supplementary Information 1.2.2). The microRNA dataset was
401 generated by processing CLIP-Seq data from 12 cell lines (Supplementary Information
402 1.2.3). We fine-tuned the model by first updating weights of the last layer for 10 epochs,
403 then updating weights of the entire model for another 30 epochs (Supplementary
404 Information 1.3). Description of all output heads can be found in Supplementary Data 2.

405

406 Held-out performance on gene expression and differential gene expression

407 We selected protein coding genes that are completely outside the training and validation
408 set, and which overlap at least one interval in the test set. Predictions and targets were
409 mean-aggregated over all exons for each gene to yield one value per gene
410 (Supplementary Information 2.1). For each tissue, we compute the correlation between
411 prediction and target across all genes. To evaluate performance on differential gene
412 expression, we constructed all pairwise comparisons between tissues, and computed
413 the \log_2 fold-change using the predicted and target coverage data (Supplementary
414 Information 2.2). For each tissue pair we computed the correlation between the
415 predicted and target \log_2 fold-changes across all genes.

416

417 Visualizing prediction on SLC7A8

418 Sequence of SLC7A8 gene was obtained from hg38 genome build with Gencode v29
419 annotation. We averaged output heads that correspond to coverage in the “Brain -
420 Hypothalamus” tissue to obtain BigRNA prediction for visualization (Supplementary
421 Information 2.3).

422

423 Held-out performance on RBP

424 Processed RBP peaks were obtained from ENCODE⁵⁴, and processed into low
425 resolution binary labels by taking into account noise in the data [Supplementary
426 Information 1.2.2, Supplementary Equation S8]. We selected protein coding genes that
427 are completely outside the training and validation set, and made predictions using
428 BigRNA and DeepRiPe⁷. Both BigRNA and DeepRiPe predictions were averaged within
429 each 128-bp window (Supplementary Information 2.4). Fig. 1g shows the average
430 precision performance of BigRNA and DeepRiPe.

431

432 Held-out performance on microRNA

433 The microRNA dataset was generated by processing CLIP-Seq data from 12 cell lines
434 (Supplementary Information 1.2.3). The called peaks were further processed into low
435 resolution binary labels by taking into account noise in the data (Supplementary
436 Equation S9). We selected protein coding genes that are completely outside the training
437 and validation set, and made predictions using BigRNA and TargetScan⁸. Both BigRNA
438 and TargetScan predictions were averaged within each 128-bp window (Supplementary
439 Information 2.5). Fig 1h shows the au-ROC performance of BigRNA and TargetScan.

440

441 Benchmarking variant effect predictions on pathogenic variants

442 Pathogenic or likely pathogenic (P/LP) UTR SNVs were obtained from Bohn et al¹³.
443 Putative benign SNVs located in the same UTR were obtained from ClinVar, if they were
444 classified as benign or likely benign (B/LB), and gnomAD v3 if their global allele

445 frequency was greater than 0.001⁵⁵ (Supplementary Information 3.1.1). For the 5' UTR
446 benchmark, we predicted the effect of the variant using BigRNA, Enformer, and
447 FramePool and took the absolute value of the variant effect scores. For the 3' UTR
448 benchmark, we evaluated BigRNA, Enformer, and Saluki and again, took the absolute
449 values of the variant effect scores (Supplementary Information 3.1.2, Supplementary
450 Equation S10-12). In addition to Fig. 2d, Supplementary Fig. S3 shows the ROC curve
451 and PRC of classification performance of all models. To compare models, we
452 performed permutation tests with 10000 permutations (Supplementary Information
453 3.1.3). Variants of uncertain significance (VUS) in the UTRs of the genes that were in the
454 benchmark were extracted as described in Supplementary Information 3.1.4.

455

456 Predicting the impact of disrupting polyadenylation sites

457 To evaluate BigRNA's ability to predict poly(A) sites, we conducted an in-silico 11 bp
458 N-mask tiling analysis across each poly(A) region. Poly(A) sites (PAS) from 200 genes
459 were obtained from PolyASite 2.0⁵⁶ (Supplementary Information 3.3). For each PAS, we
460 expanded the site by ± 100 bp to cover proximal regulatory elements, resulting in 206 bp
461 regions. We subsequently N-masked 11 bp tiles across the region and compared
462 BigRNA predictions for the N-masked sequences (mutant) and the poly(A) signal
463 sequence (wildtype). The BigRNA predictions were based on the mean of the individual
464 sample RNA-seq coverage heads across all tissue types (Supplementary Information
465 3.3). For the *NAA10* PAS and its surrounding 100 bp context, we performed saturation
466 mutagenesis by point-mutating every reference nucleic acid base to every other nucleic
467 acid base. Similar to the poly(A) site analysis, we carried out predictions using the
468 BigRNA model to assess the impact of these mutations on gene expression.

469

470 Expression quantitative trait loci (eQTLs) and linkage disequilibrium (LD) estimation for 471 *PON1* variants

472 The four variants with known expression effect were rs705379 (chr7:95324583:G:A),
473 rs854571 (chr7:95325307:T:C), rs854572 (chr7:95325384:C:G) and rs3735590
474 (chr7:95298183:G:A). The eQTL and normalized effect size of these variants on *PON1*
475 liver tissue expression were obtained from the GTEx eQTL Calculator . The LD R^2 values
476 between variants was calculated using the NIH LDmatrix tool with the GBR population
477 selected.

478

479 Classifying expression quantitative trait loci (eQTLs) versus matched controls

480 To construct a benchmark dataset from confidently fine-mapped eQTLs, variants with a
481 posterior inclusion probability of 0.5 or greater (indicating that they are the most likely
482 causal variant in the credible set) were selected from eQTLGen statistical fine-mapping
483 of expression modulating variants in GTEx v8¹⁹. eQTLs within 50kbp of the transcription

484 start site of the primary or most highly expressed transcript for the reported eGene were
485 selected to ensure that deep learning models would have sufficient genomic context to
486 accurately predict changes in expression. For each eQTL we selected a matched
487 negative control variant from the same effector gene (eGene) which was not associated
488 with its expression ($P > 0.05$) in any tissue and within 10% of the eQTL's minor allele
489 frequency and 10kbp of the eQTL's genomic position. This resulted in a dataset of 1374
490 eQTL variants and 1162 matched negative controls.

491

492 Classifying variants that cause intron retention

493 Variants that cause full intron retention were manually curated from splicing variants
494 downloaded from the SPCards database²⁷. A matching set of variants that do not cause
495 intron retention were processed from gnomAD⁵⁵ (Supplementary Information 4.1.1). For
496 each variant, we use BigRNA to predict the relative coverage between intron and the two
497 flanking exons, and compute the score as the ratio between wild-type and mutant-type,
498 aggregated across models in ensemble (Supplementary Information 4.1.2,
499 Supplementary Equation S14-16).

500

501 Classifying variants that cause exon skipping

502 For each mutation in the MaPSy dataset²¹, we computed the splicing odds ratio and
503 confidence interval using the reported readout from both in-vitro and in-vivo assays, to
504 create a high confidence binary label on skipping versus non-skipping at splicing levels
505 ranging from 50% to 10% (Supplementary Information 4.2.1, Supplementary Equation
506 S17-18). For each mutation, we used BigRNA to predict the difference in junction counts
507 between wild-type and mutant-type, normalized by exon, and aggregate across models
508 in ensemble (Supplementary Information 4.2.2, Supplementary Equation S19). Fig 3a
509 shows ROC curve of classification performance on skipping versus non-skipping at 50%
510 splicing level. For model comparison (Supplementary Fig. S7), we performed
511 permutation tests with 100,000 permutations.

512

513 Predicting the effect of splice-switching SBOs

514 To obtain the relative ranking of Nusinersen, we ranked all possible SBOs of length 18
515 within 200 base pairs of exon 7 of SMN2 (Supplementary Information 5.1). We used
516 RT-PCR to measure the Percentage Spliced In (PSI) values for 15 exons in the HEK293T
517 cell line, and compared the measured PSI with the predicted SBO effect of SpliceAI and
518 BigRNA using the Spearman Correlation metric (Supplementary Information 5.2). We
519 repeated the above evaluation for SBOs targeting Met645Arg; here we edited HepG2
520 cells to introduce the c.1934T>G Met645Arg variant, and screened a library of 55 SBOs
521 by qPCR. Spearman correlation was computed between BigRNA predictions and the
522 experimentally observed *ATP7B* expression levels (Supplementary Information 5.3). The

523 same evaluation was carried out on published data of SBOs designed to skip a
524 pseudo-exon created by the c.5763-1050A>G variant in ATM³⁰.

525

526 The set of “N=1” variants was created by selecting pathogenic or likely pathogenic
527 variants (ClinVar) from genes that are exclusively associated with autosomal recessive
528 disorders (OMIM). BigRNA predictions were made for SNVs with very low estimated
529 worldwide prevalence (n=1582, GnomAD) and we curated synonymous, tolerated
530 missense (SIFT) and intronic variants (excluding the core dinucleotides) for their
531 mechanisms of pathogenicity (Supplementary Information 5.5). All possible 20-mers
532 within 200 bp of *MYO1E* exon 23 were scored for their ability to remedy the effect of the
533 c.2481-12A>G variant, and we visualized the predictions for the highest ranked SBO.

534

535 Predicting the effect of expression increase SBOs

536 Expression increase can occur through a variety of mechanisms, and SBOs can be
537 designed anywhere in the gene. By applying a combination of established saliency
538 mapping techniques^{33,34}, we evaluated the contribution of each base pair in a transcript
539 to the expression of the related gene in the relevant tissue, yielding a sensitivity score
540 for each base pair's impact on gene expression levels, called the Inhibitory Score
541 (Supplementary Information 1.5). This per-base-pair score was then used to rank SBOs
542 by taking the minimum score of any overlapping base-pair (Supplementary Information
543 5.6.2). For the Nusinersen ranking evaluation, we used the BigRNA Inhibitory Score to
544 score all candidate SBOs of length 18 targeting the entirety of the gene body of *SMN2*.
545 The same process was applied to expression increase SBOs identified from screens of
546 *PON1*, *ATPB*, *PRRT2*, and *SERPING1*. Scores between hit SBOs and the background of all
547 candidate SBOs were compared with a Mann-Whitney U-Test.

548

549 **Data Availability**

550 Data and code to be made available upon peer review.

551

552 **References**

553 1. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198

554 (2004).

555 2. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range

556 interactions. *Nat. Methods* **18**, 1196–1203 (2021).

557 3. Wainberg, M., Merico, D., DeLong, A. & Frey, B. J. Deep learning in biomedicine. *Nat.*

- 558 *Biotechnol.* **36**, 829–838 (2018).
- 559 4. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level
560 mammalian expression atlas. *Nature* **507**, 462–470 (2014).
- 561 5. The GTEx Consortium *et al.* The GTEx Consortium atlas of genetic regulatory effects across
562 human tissues. *Science* **369**, 1318–1330 (2020).
- 563 6. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding
564 proteins. *Nature* **583**, 711–719 (2020).
- 565 7. Ghanbari, M. & Ohler, U. Deep neural networks for interpreting RNA-binding protein target
566 preferences. *Genome Res.* **30**, 214–226 (2020).
- 567 8. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites
568 in mammalian mRNAs. *eLife* **4**, e05005 (2015).
- 569 9. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,
570 583–589 (2021).
- 571 10. Buel, G. R. & Walters, K. J. Can AlphaFold2 predict the impact of missense mutations on
572 structure? *Nat. Struct. Mol. Biol.* **29**, 1–2 (2022).
- 573 11. Wu, Y. *et al.* Improved pathogenicity prediction for rare human missense variants. *Am. J.*
574 *Hum. Genet.* **108**, 2389 (2021).
- 575 12. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human
576 genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
- 577 13. Bohn, E., Lau, T. T. Y., Wagih, O., Masud, T. & Merico, D. A curated census of
578 pathogenic and likely pathogenic UTR variants and evaluation of deep learning models for
579 variant effect prediction. *Front. Mol. Biosci.* **10**, 1257550 (2023).
- 580 14. Johnston, J. J. *et al.* NAA10 polyadenylation signal variants cause syndromic
581 microphthalmia. *J. Med. Genet.* **56**, 444–452 (2019).
- 582 15. Karollus, A., Avsec, Ž. & Gagneur, J. Predicting mean ribosome load for 5'UTR of any
583 length using deep learning. *PLOS Comput. Biol.* **17**, e1008982 (2021).

- 584 16. Agarwal, V. & Kelley, D. R. The genetic and biochemical determinants of mRNA
585 degradation rates in mammals. *Genome Biol.* **23**, 245 (2022).
- 586 17. Hino, M. *et al.* The +1,506 (A>C) Mutation in the 3' Untranslated Region Affects β -Globin
587 Expression. *Hemoglobin* **36**, 399–406 (2012).
- 588 18. Leviev, I. & James, R. W. Promoter Polymorphisms of Human Paraoxonase *PON1* Gene
589 and Serum Paraoxonase Activities and Concentrations. *Arterioscler. Thromb. Vasc. Biol.* **20**,
590 516–521 (2000).
- 591 19. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and
592 splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
- 593 20. Liu, M.-E. *et al.* A functional polymorphism of *PON1* interferes with microRNA binding to
594 increase the risk of ischemic stroke and carotid atherosclerosis. *Atherosclerosis* **228**,
595 161–167 (2013).
- 596 21. Soemedi, R. *et al.* Pathogenic variants that alter protein code often disrupt splicing. *Nat.*
597 *Genet.* **49**, 848–855 (2017).
- 598 22. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning.
599 *Cell* **176**, 535-548.e24 (2019).
- 600 23. Grønning, A. G. B. *et al.* DeepCLIP: predicting the effect of mutations on protein–RNA
601 binding with deep learning. *Nucleic Acids Res.* gkaa530 (2020) doi:10.1093/nar/gkaa530.
- 602 24. Waddell, L. *et al.* Medium-chain acyl-CoA dehydrogenase deficiency:
603 Genotype–biochemical phenotype correlations. *Mol. Genet. Metab.* **87**, 32–39 (2006).
- 604 25. Merico, D. *et al.* ATP7B variant c.1934T > G p.Met645Arg causes Wilson disease by
605 promoting exon 6 skipping. *NPJ Genomic Med.* **5**, 16 (2020).
- 606 26. Loudianos, G. *et al.* Abnormal mRNA splicing resulting from consensus sequence
607 splicing mutations of ATP7B: ATP7B ABNORMAL SPLICING IN WILSON DISEASE. *Hum.*
608 *Mutat.* **20**, 260–266 (2002).
- 609 27. Li, K. *et al.* Performance evaluation of differential splicing analysis methods and splicing

- 610 analytics platform construction. *Nucleic Acids Res.* **50**, 9115–9126 (2022).
- 611 28. Sangermano, R. *et al.* ABCA4 midigenes reveal the full splice spectrum of all reported
612 noncanonical splice site variants in Stargardt disease. *Genome Res.* **28**, 100–110 (2018).
- 613 29. Wurster, C. D. & Ludolph, A. C. Nusinersen for spinal muscular atrophy. *Ther. Adv.*
614 *Neurol. Disord.* **11**, 1756285618754459 (2018).
- 615 30. Kim, J. *et al.* A framework for individualized splice-switching oligonucleotide therapy.
616 *Nature* **619**, 828–836 (2023).
- 617 31. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids*
618 *Res.* **31**, 3812–3814 (2003).
- 619 32. Domingo-Gallego, A. *et al.* Clinical utility of genetic testing in early-onset kidney disease:
620 seven genes are the main players. *Nephrol. Dial. Transplant.* **37**, 687–696 (2022).
- 621 33. Majdandzic, A., Rajesh, C. & Koo, P. K. Correcting gradient-based interpretations of
622 deep neural networks for genomics. *Genome Biol.* **24**, 109 (2023).
- 623 34. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing
624 noise by adding noise. Preprint at <http://arxiv.org/abs/1706.03825> (2017).
- 625 35. James, R. W. *et al.* Promoter polymorphism T(-107)C of the paraoxonase PON1 gene is
626 a risk factor for coronary heart disease in type 2 diabetic patients. *Diabetes* **49**, 1390–1393
627 (2000).
- 628 36. Wang, M., Lang, X., Zou, L., Huang, S. & Xu, Z. Four genetic polymorphisms of
629 paraoxonase gene and risk of coronary heart disease: A meta-analysis based on 88
630 case–control studies. *Atherosclerosis* **214**, 377–385 (2011).
- 631 37. Shih, D. M. *et al.* Mice lacking serum paraoxonase are susceptible to organophosphate
632 toxicity and atherosclerosis. *Nature* **394**, 284–287 (1998).
- 633 38. Litvinov, D., Mahini, H. & Garelnabi, M. Antioxidant and anti-inflammatory role of
634 paraoxonase 1: implication in arteriosclerosis diseases. *North Am. J. Med. Sci.* **4**, 523–532
635 (2012).

- 636 39. Tward, A. *et al.* Decreased atherosclerotic lesion formation in human serum
637 paraoxonase transgenic mice. *Circulation* **106**, 484–490 (2002).
- 638 40. Mackness, B., Quarck, R., Verreth, W., Mackness, M. & Holvoet, P. Human
639 Paraoxonase-1 Overexpression Inhibits Atherosclerosis in a Mouse Model of Metabolic
640 Syndrome. *Arterioscler. Thromb. Vasc. Biol.* **26**, 1545–1550 (2006).
- 641 41. Ebrahimi-Fakhari, D., Saffari, A., Westenberger, A. & Klein, C. The evolving spectrum of
642 *PRRT2* -associated paroxysmal diseases. *Brain* **138**, 3476–3495 (2015).
- 643 42. Ponard, D. *et al.* SERPING1 mutation update: Mutation spectrum and C1 Inhibitor
644 phenotypes. *Hum. Mutat.* **41**, 38–57 (2020).
- 645 43. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic
646 variation. *Nature* **526**, 68–74 (2015).
- 647 44. Holgersen, E. M. *et al.* Transcriptome-Wide Off-Target Effects of Steric-Blocking
648 Oligonucleotides. *Nucleic Acid Ther.* **31**, 392–403 (2021).
- 649 45. Merkle, T. *et al.* Precise RNA editing by recruiting endogenous ADARs with antisense
650 oligonucleotides. *Nat. Biotechnol.* **37**, 133–138 (2019).
- 651 46. Dao, T. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning.
652 (2023) doi:10.48550/ARXIV.2307.08691.
- 653 47. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical
654 Image Segmentation. Preprint at <http://arxiv.org/abs/1505.04597> (2015).
- 655 48. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human
656 epigenomes. *Nature* **518**, 317–330 (2015).
- 657 49. Mindermann, S. *et al.* Prioritized Training on Points that are Learnable, Worth Learning,
658 and Not Yet Learnt. (2022) doi:10.48550/ARXIV.2206.07137.
- 659 50. Chopra, S., Hadsell, R. & LeCun, Y. Learning a Similarity Metric Discriminatively, with
660 Application to Face Verification. in *2005 IEEE Computer Society Conference on Computer*
661 *Vision and Pattern Recognition (CVPR'05)* vol. 1 539–546 (IEEE, 2005).

- 662 51. Huang, C. *et al.* *Personal transcriptome variation is poorly explained by current genomic*
663 *deep learning models*. <http://biorxiv.org/lookup/doi/10.1101/2023.06.30.547100> (2023)
664 doi:10.1101/2023.06.30.547100.
- 665 52. Sasse, A. *et al.* *How far are we from personalized gene expression prediction using*
666 *sequence-to-expression deep neural networks?*
667 <http://biorxiv.org/lookup/doi/10.1101/2023.03.16.532969> (2023)
668 doi:10.1101/2023.03.16.532969.
- 669 53. Castel, S. E. *et al.* *A vast resource of allelic expression data spanning human tissues.*
670 *Genome Biol.* **21**, 234 (2020).
- 671 54. Luo, Y. *et al.* *New developments on the Encyclopedia of DNA Elements (ENCODE) data*
672 *portal.* *Nucleic Acids Res.* **48**, D882–D889 (2020).
- 673 55. Chen, S. *et al.* *A genome-wide mutational constraint map quantified from variation in*
674 *76,156 human genomes.* <http://biorxiv.org/lookup/doi/10.1101/2022.03.20.485034> (2022)
675 doi:10.1101/2022.03.20.485034.
- 676 56. Herrmann, C. J. *et al.* *PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3'*
677 *end sequencing.* *Nucleic Acids Res.* gkz918 (2019) doi:10.1093/nar/gkz918.

678 Acknowledgements

679 We thank David Kelley for advice, Janine Truong for editing figures and reviewing the
680 manuscript, and Laurence MacPhie for reviewing the manuscript. We also thank Daniele Merico
681 for help in conceiving of the appropriate benchmarks, and Tim Yu for suggesting we include
682 results for ultra-rare variants.

683 Author Information

684 Authors and Affiliations

685 Deep Genomics, Toronto, Canada

686 Albi Celaj, Alice Jiexin Gao, Tammy T.Y. Lau, Erle M. Holgersen, Alston Lo, Varun Lodaya,
687 Christopher B. Cole, Robert E. Denroche, Carl Spickett, Omar Wagih, Pedro O. Pinheiro,

688 Parth Vora, Pedrum Mohammadi-Shemirani, Steve Chan, Zach Nussbaum, Nicole Zhang,
689 Helen Zhu, Easwaran Ramamurthy, Bhargav Kanuparthi, Michael Iacocca, Diane Ly, Ken
690 Kron, Marta Verby, Kahlin Cheung-Ong, Zvi Shalev, Brandon Vaz, Sakshi Bhargava,
691 Farhan Yusuf, Sharon Samuel, Sabriyeh Alibai, Zahra Baghestani, Xinwen He, Kirsten
692 Krastel, Oladipo Oladapo, Amrudha Mohan, Arathi Shanavas, Magdalena Bugno,
693 Jovanka Bogojeski, Frank Schmitges, Carolyn Kim, Solomon Grant, Rachana Jayaraman,
694 Tehmina Masud, Amit Deshwar, Shreshth Gandhi, Brendan J. Frey

695 **Contributions**

696 A.C., A.J.G., and B.F. initiated the project. A.C., A.J.G., T.T.Y.L., E.M.H., and C.B.C. conceived of
697 the study and designed analyses. A.C., P.O.P. and Z.N. designed the model. A.J.G., A.L., V.L., and
698 S.C. helped implement and improve the model. A.C., X.Z., P.V., and H.Z. processed the training
699 data. A.J.G., T.T.Y.L., E.M.H., A.L., V.L., C.B.C., R.E.D., O.W., P.V., E.R., and B.K. performed
700 benchmarking and downstream analyses. P.V., P.M.S, M.I., E.R., C.K., and S.Gr. aided with
701 benchmarking. E.M.H., V.L., P.V., and R.J. processed experimental data. C.S., D.L., K.K., M.V.,
702 K.C.O, Z.S., B.V., S.B., F.Y., S.S., S.A., Z.B., X.H., K.K., O.O., A.M., A.S., M.B., J.B., and F.S. performed
703 validation experiments; T.M., A.D., S.Gh., and B.F. supervised the study

704 **Ethics Declaration**

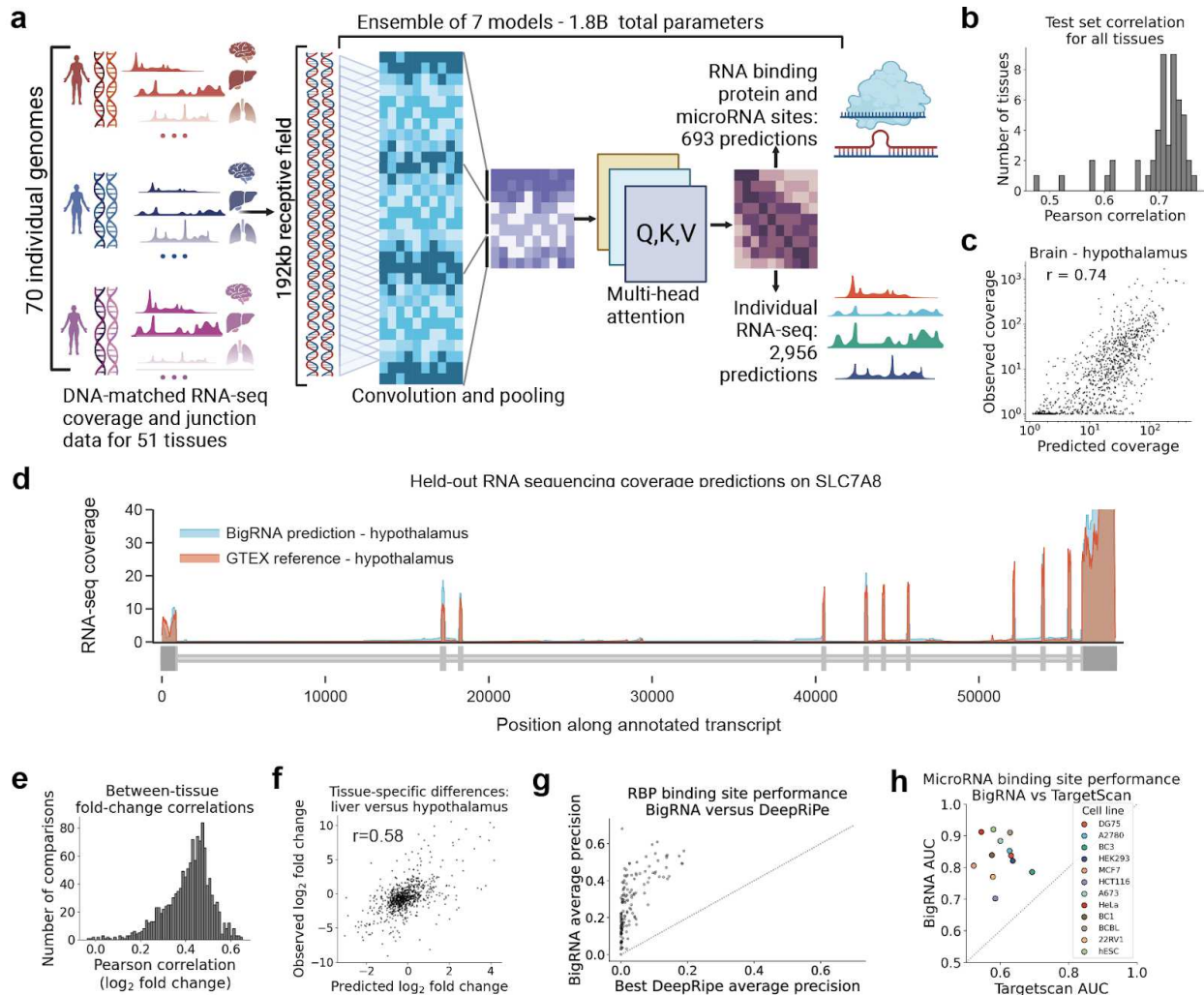
705 **Competing interests**

706 All listed authors are present or past employees of Deep Genomics Inc. This study received
707 funding from Deep Genomics in the form of salary support and covering of computational
708 costs. The founder was involved in the decision to submit for publication.

709 **Supplementary Information**

710 In a separate document.

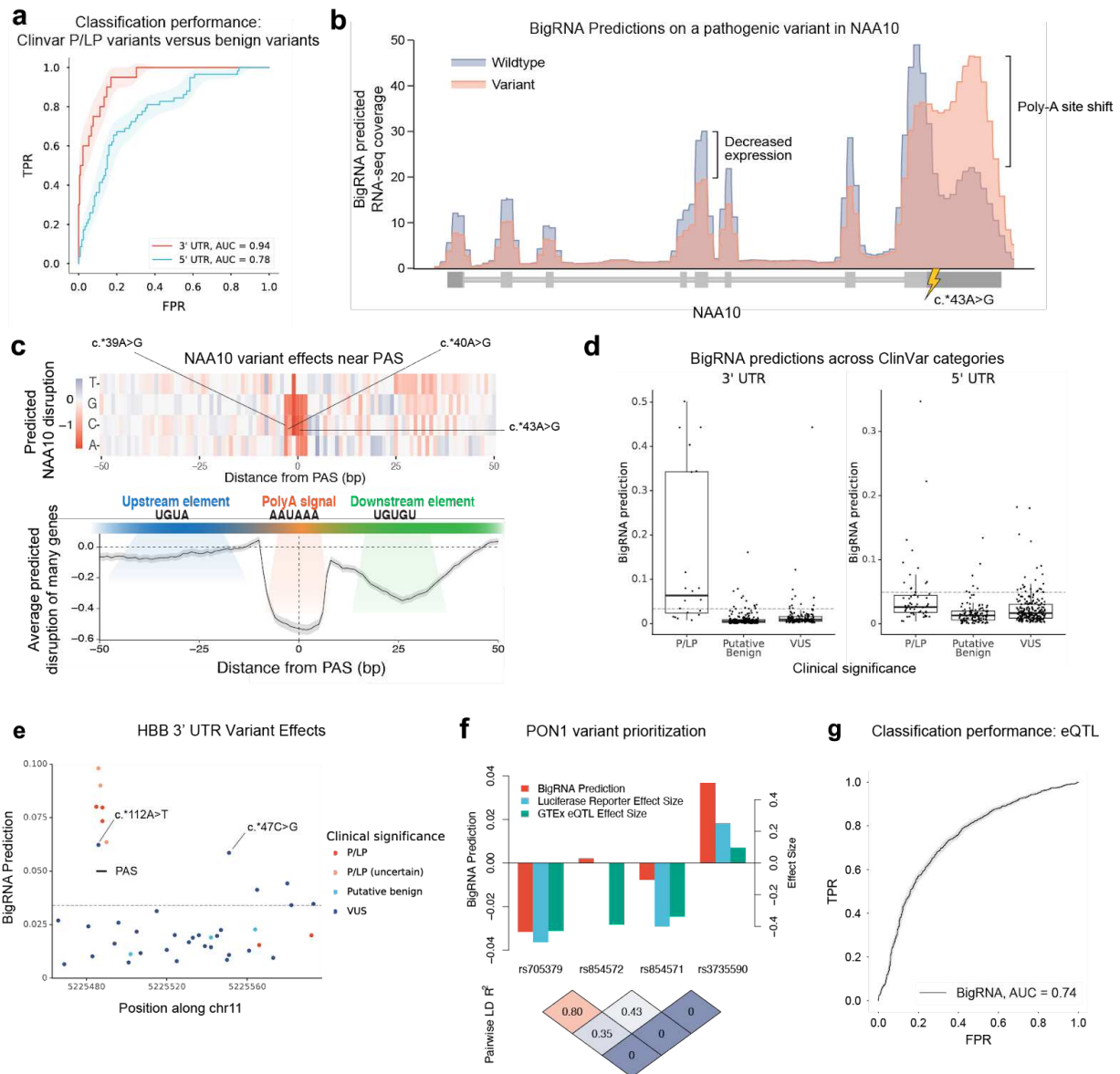
711 Figures



712

713 **Figure 1. BigRNA accurately predicts tissue-specific RNA expression of unseen sequences.**

714 **a.** A schematic of BigRNA's training. BigRNA was trained on the genomes of 70 individuals, to
 715 predict a total of 2,956 RNA-seq datasets over 51 tissues, plus 693 datasets corresponding to
 716 RNA binding protein and microRNA sites. **b.** Distribution of correlations between predicted and
 717 measured RNA-seq coverage in exonic regions for genes held-out during training (averaged
 718 across individuals). **c.** Correlation between predicted and measured RNA-seq coverage for the
 719 hypothalamus samples. **d.** Predicted versus measured coverage for *SLC7A8*, averaged across
 720 hypothalamus samples for all individuals. **e.** Distribution of correlations between predicted and
 721 measured fold-change (pearson r) for all pairwise comparisons across 51 tissues. **f.**
 722 Fold-change in gene coverage between liver and hypothalamus. **g.** Comparison of BigRNA and
 723 a previously published method, DeepRiPe, for predicting the binding sites of 98 RNA binding
 724 proteins across 2 cell lines (142 total experiments). **h.** Comparison of BigRNA and a previously
 725 published method, TargetScan, for predicting microRNA binding sites for 12 cell lines.

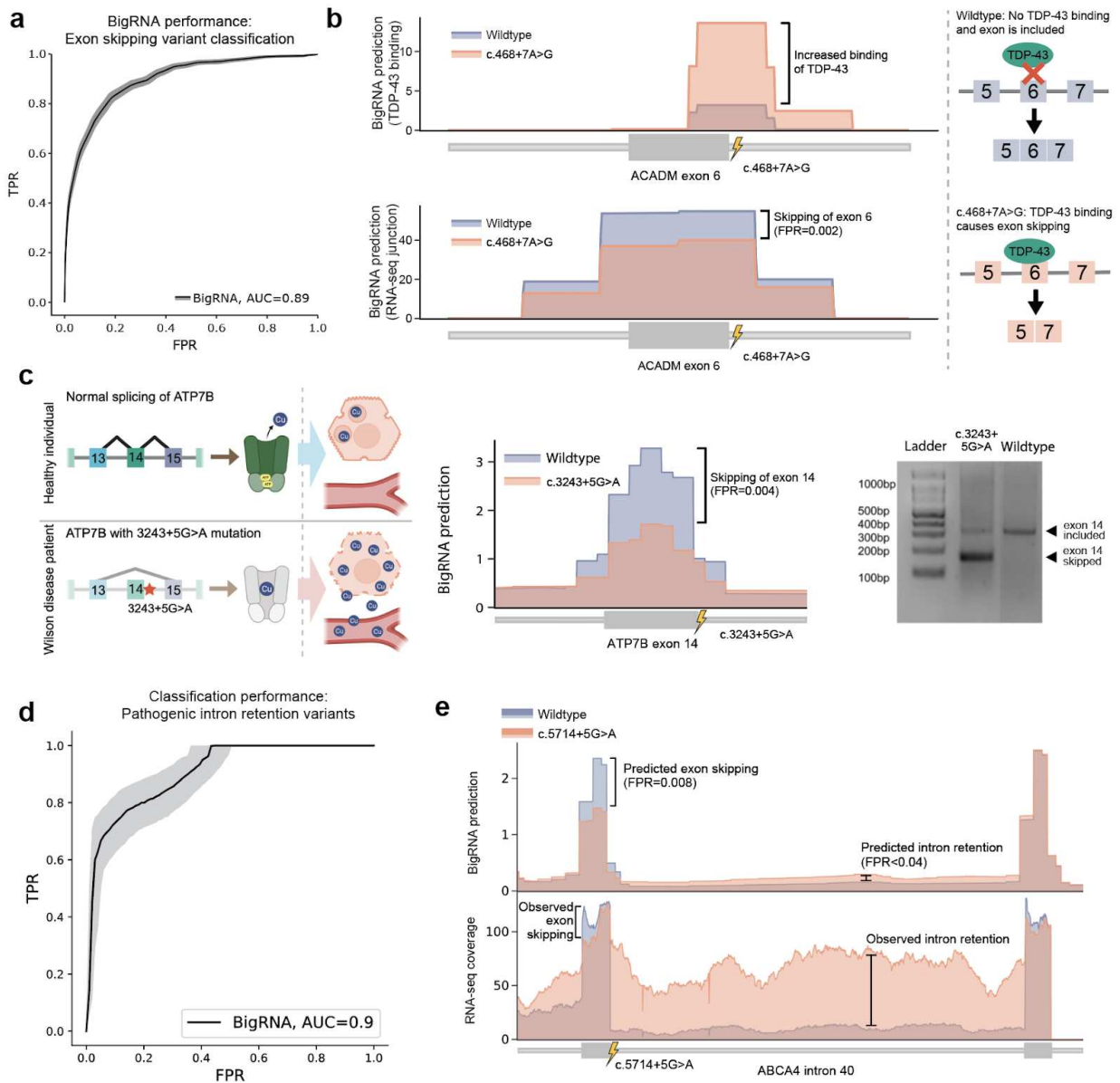


726

727 Figure 2. BigRNA predicts the effects of pathogenic expression-modulating variants

728 **a.** Performance of BigRNA on classifying P/LP variants from putative benign variants in the 3'
 729 UTR and 5' UTR. **b.** RNA-seq coverage predictions for the effects of a pathogenic variant in the 3'
 730 UTR of NAA10 (NM_003491.4), averaged across all individuals and all tissue types. **c.** Top:
 731 BigRNA predictions showing the change in expression for all possible point mutations around
 732 the polyadenylation site (PAS) of NAA10. Three variants previously identified as impacting the
 733 PAS are labeled. Bottom: Relationship between the change in expression predicted by BigRNA
 734 from ablating regions around the PAS relative to the distance from the PAS for 200 human
 735 poly(A) signal sequences selected from PolyASite 2.0. **d.** The distribution of BigRNA scores for
 736 P/LP variants, putative benign variants, and VUS variants from ClinVar for genes included in the
 737 UTR benchmarks. The dashed line in both plots (left, $y = 0.0341$; right, $y = 0.0494$) represents

738 the threshold of classifying P/LP from putative benign variants at an FPR of 5% in each of the
739 benchmark datasets. **e.** BigRNA predictions for variants of varying clinical significance in *HBB*.
740 The dashed line represents the threshold of classifying P/LP from putative benign variants at a
741 5% FPR in the 3' UTR ($y = 0.0341$). The two highest scoring VUS variants in this gene are
742 annotated. **f.** Top: Comparing BigRNA predicted effects to GTEx eQTL effect size and results of
743 a luciferase reporter assay for four variants suspected to impact PON1 expression. Bottom:
744 Estimated linkage disequilibrium between variants. **g.** Performance of BigRNA at distinguishing
745 fine-mapped expression quantitative trait loci (eQTLs) from controls matched by effector gene
746 (eGene), distance to the transcription start site of the eGene, and minor allele frequency.

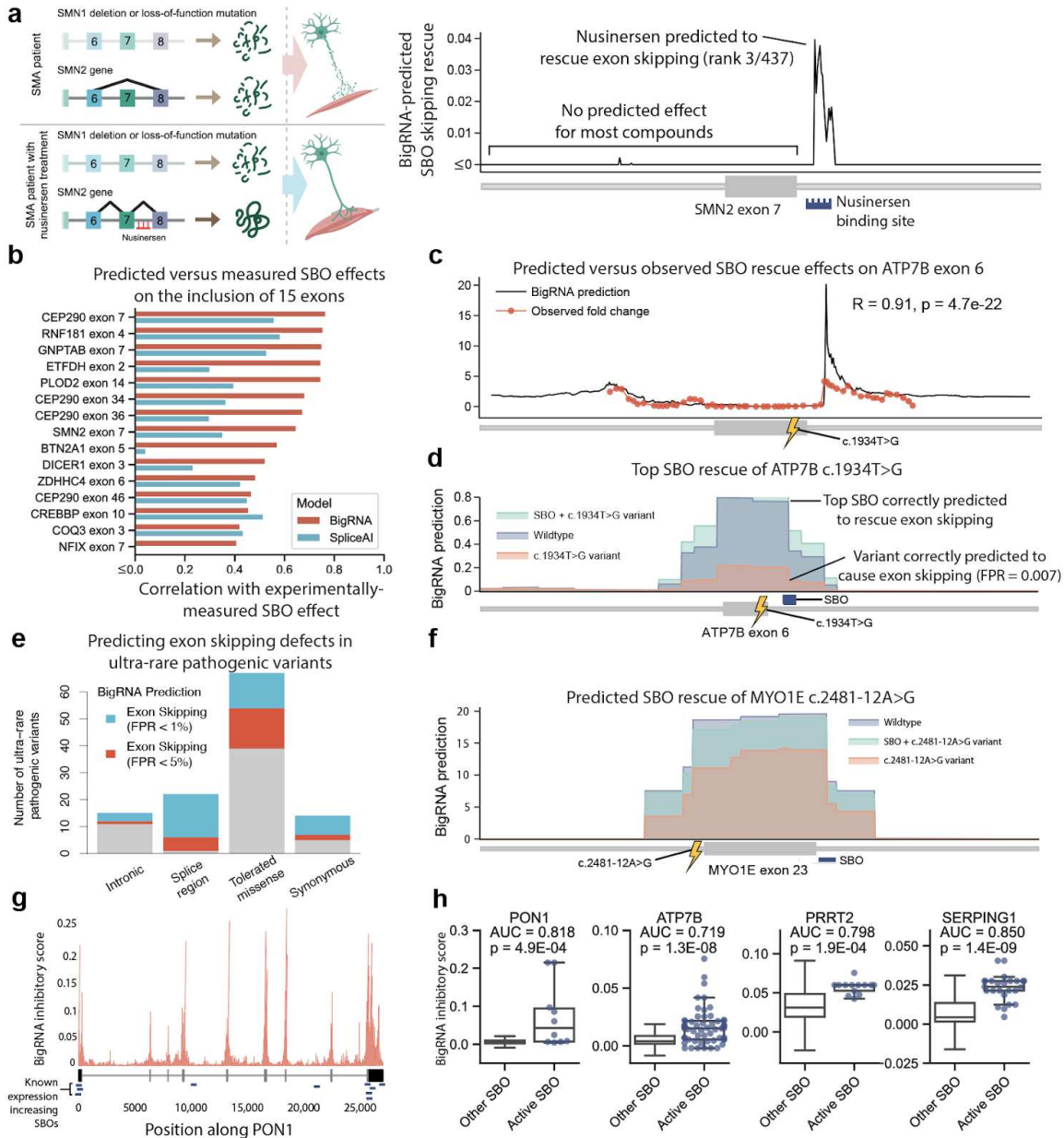


747

748 **Figure 3. BigRNA captures the effect of variants on splicing.**

749 **a.** BigRNA performance on classifying exonic variants that result in exon skipping by at least
 750 50%, from exonic variants that do not cause skipping, both obtained from MaPSy. **b.** BigRNA
 751 predicts that the c.468+7A>G variant will result in increased TDP-43 binding and skipping of
 752 *ACADM* exon 6. **c.** The *ATP7B* VUS c.3243+5G>A is predicted by BigRNA to cause in-frame
 753 skipping of exon 14. This results in reduced levels of functional ATP7B protein, leading to
 754 copper buildup in the cell. Right: An RT-PCR in HepG2 cells edited to be homozygous for
 755 c.3243+5G>A confirms the expected fragment from exon skipping. **d.** BigRNA performance on
 756 classifying variants that cause intron retention (n = 25) from a set of matched variants that do
 757 not impact splicing (n = 63). **e.** Top: BigRNA coverage predictions of the c.5714+5G>A variant in

758 *ABCA4*. Bottom: RNA-seq of wildtype WERI cells and WERI cells edited to be homozygous for
759 the variant confirm both exon skipping and intron retention effects.



761

762 Figure 4. BigRNA predicts the effects of steric blocking oligonucleotides. a. Mechanism of
763 action of the splice-switching oligonucleotide Nusinersen, an approved treatment for spinal
764 muscular atrophy (SMA). BigRNA predictions are shown for the exon-restoring effects of all
765 18-mer SBOs within 200 bp of *SMN2* exon 7. The blue bar shows the position of Nusinersen.
766 Predictions were truncated at zero for the plot. **b.** Spearman correlation between experimentally
767 observed exon-inclusion levels and predictions generated by BigRNA and SpliceAI. A negative
768 correlation for *NFIX* exon 7 versus SpliceAI ($r = -0.13$) was truncated to zero. **c.** BigRNA
769 predictions of SBO effects on *ATP7B* exon 6 inclusion. 55 SBOs were screened by qPCR to
770 measure total *ATP7B* expression relative to control (fold change), and the Spearman correlation
771 was computed between the BigRNA predictions and observed fold changes. **d.** BigRNA

772 predictions for wildtype, Met645Arg (c.1934T>G) variant, and Met645Arg variant with treatment
773 (lead SBO targeting *ATP7B* exon 6). The junction count tracks pertaining to individual samples
774 of the liver tissue are averaged for plotting. **e.** Proportion of ultra-rare pathogenic variants
775 associated with AR disorders with BigRNA exon skipping predictions above the 1% and 5% FPR
776 thresholds. Intronic (>8bp from splice site), splice region (<8bp from splice site excluding the
777 core dinucleotides), tolerated missense (SIFT score > 0.05) and synonymous variants are
778 shown. **f.** BigRNA predictions for wildtype, c.2481-12A>G variant and the variant with treatment
779 (lead SBO targeting *MYO1E* exon 23). **g.** BigRNA predicts expression increase SBOs in *PON1*.
780 BigRNA inhibitory scores are plotted by region of the gene. The transcript structure is shown
781 under the scores, and the locations of the 10 dose-response hits are shown with blue bars. The
782 distribution of BigRNA inhibitory scores for the 10 dose-response hits is significantly different
783 from the distribution for other length-matched SBOs targeting *PON1* **h.** BigRNA scores of
784 screening hits compared to background of all possible SBOs of same length for *PON1*, *ATP7B*,
785 *PRRT2*, and *SERPING1*.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementaryinformation.pdf](#)