nature portfolio

Peer Review File

Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation

Corresponding Author: Dr Johannes Linder

Version 0:

Decision Letter:

5th Oct 2023

Dear Dr Linder,

Your Article, "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

In brief, the reviews acknowledge the technical and conceptual advance presented by Borzoi, but also suggest there are important issues that require improvement.

Reviewer #1 is supportive, saying your work is "highly valuable", but think that the utility of Borzoi must be enhanced ("the usefulness of the model for the field is presently limited by the fact that the authors present many ways to employ it, without ever making a clear statement as to which method is to be preferred"); they provide a long and detailed list of guidance to improve this and other aspects. Furthermore, they highlight that many of your code/materials for model training/evaluation need to be shared.
Referee #2 is similarly supportive; their requests have some overlap with the first reviewer, e.g., questions about how much data is needed to train Borzoi and how this will impact on usability for less well-resourced labs.
Reviewer #3 - who we would point out is representative of the journal's broader, non-ML-specialist audience - straightforwardly finds the current presentation overly technical and dense. We note that there are likely some misunderstandings on e.g. the training/test datasets, but we are also inclined to agree with their overall point that the presentation must be majorly improved so that your message can be better-communicated to that non-expert audience.

In our reading we think all the major requests are quite doable and reasonable, and we would encourage you to respond to them all as fully as possible. In particular, we would highlight the questions about Borzoi's utility (how/when can/should it be used?), and Referee #3's viewpoint, as especially important - we would recommend asking a non-ML biologist colleague or two to read the text!

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a “Response to referees” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the

referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics

ORCiD: 0000-0003-1589-7087

Referee expertise:

Referees #1,2: machine learning, computational biology, genomics.

Referee #3: RNA-seq, splicing, molecular diagnostics.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Linder et al. describe a novel deep learning model, called "Borzoi", that can predict RNA-seq coverage in a variety of human tissues at a resolution of 32 base pairs. The model is trained on ENCODE and GTEx and presents a substantial methodological advance over the previous state-of-the-art model of human transcriptional control Enformer. By predicting RNA-seq coverage directly, the model overcomes some major limitations of previous approaches and allows interrogation of transcriptional and post-transcriptional regulation.

Overall, we believe that this paper will be highly valuable to the field and congratulate the authors on their hard work. However, currently several aspects of the manuscript are unclear and may give a wrong impression to the reader. Moreover, the usefulness of the model for the field is presently limited by the fact that the authors present many ways to employ it, without ever making a clear statement as to which method is to be preferred.

Major points

1. The Authors claim (line 188, a section title) that Borzoi “predicts the impact of genetic variation on gene expression”. This is not shown in the manuscript. What the authors show is (1) that a random forest trained on the outputs of the model can be used to classify likely eQTLs vs. matched negatives and (2) that there is a (limited) spearman correlation between the model prediction and the normalized eQTL effect size. Since the former is a classification task and the latter only measures the concordance in terms of rank, neither of these really correspond to “predicting the impact of genetic variation on gene expression”. To do this, the authors would have to demonstrate that the model can quantitatively predict fold changes due to variants. Accordingly, the section should be reworded to reflect that currently it is only shown that the method can rank or prioritize variants.

2. Related to the first point, two analyses would be valuable to better gauge how the model fares in predicting quantitative effects of variation. First, the Enformer paper benchmarked their method on saturation mutagenesis data. This data also is imperfect, of course, but nevertheless a decidedly more quantitative measure of the ability of the model to predict the effects of local variants on expression while sidestepping the problem of linkage. The authors should include this benchmark or provide a compelling reason why this should not be done.

3. Second, the benchmark of Enformer cited by the authors (citation [56]) claims that the predicted effect of (validated) enhancers decays strongly as a function of distance to the TSS. To what extent is this also the case for the RNA-seq predictions of Borzoi? Additionally, is the effect of enhancers uniform across the RNA-seq track (indicating that the model uses distal elements to modulate its overall prediction of the transcription of a gene) or are the genic bins closer to an enhancer/variant affected more heavily than those further away? Answering these questions will help practitioners gauge how to compare predicted effects at different distances.

4. Furthermore, as the authors themselves note, one of the main uses of their model likely will be in interpreting variants. However, many of the design choices by the authors make this nontrivial. Firstly, because of the ensembling strategy, there is not just one Borzoi, but four models. Secondly, sometimes attributions or predictions are averaged over forward and reverse complements. Thirdly, the authors use a variety of different attribution methods (gradients, smoothed gradients, ism, “ism-shuffle”) in a very task-specific manner, without really explaining why a specific method was chosen for a particular task. As a result, after reading the manuscript, it remains unclear to us what the recommended way to interrogate variants with Borzoi is. Especially because running this model is very slow and expensive, it would severely limit the useability if users had to try every combination of possibilities explored in the paper themselves. Accordingly, the authors should evaluate the advantages of these different approaches and determine best practices for employing their method.

5. Related to this, the many different attribution methods, Gaussian filters, cutoffs and other processing steps employed in the paper do give the authors additional degrees of freedom, which raises the concern that sometimes the analysis method is tailored to make the results look more impressive in a way that is not generalizable. Specifically, in Figure 3 C-D, the authors compare the ability to prioritize distal enhancers to Enformer. In the Enformer paper, this analysis was done using three methods: gradient*input, attention and ISM. In this paper only the gradient method was used, although the authors additionally “smooth the scores… using a Gaussian filter” (which to our knowledge was not done in the Enformer paper). How were the parameters of this Gaussian filter determined? Why was this method used and not another? Also why do the authors evaluate the gradient for Enformer based on “two 128 bp bins centered at the gene’s TSS” (line 817-819) whereas the original Enformer paper used three bins: “We note that ‘CAGE at TSS’ always means summing the absolute gradient values from three adjacent bins, as is also done in gene-focused model evaluation. The three bins include the bin overlapping the TSS and one flanking bin on each side” (Enformer paper, section “Enhancer prioritization”). Moreover, the Enformer paper specifies that “The enhancer–gene score was obtained by summing the absolute gradient × input scores in the 2-kb window centered at the enhancer.”, whereas in this paper, we did not find any information as to the size of the window around the enhancer. Given the overlap in authors between these two papers, it is difficult to understand why the details of this benchmark differ. To make the results interpretable, the authors should either run the benchmark exactly as was done previously for both Enformer and Borzoi (as 128 is a multiple of 32, the difference in bin size does not prevent this). Or the authors should explain why the new parameters they have chosen are superior to those chosen previously.

6. Also, why are ISM-based, and not gradient-based analyses, then used for the splicing and polyadenylation examples? Did gradient-based methods not work for these cases?
7. Why is the overall ability to predict variation in coverage between exons and introns “evaluated on the top 20% of test set genes with highest variance in coverage across the span of exons and introns” and not, say 10% or 25% or 50%? What happens beyond 20%?

8. “We clip the lower end of the 4, 778 maximum deviations at the 25th percentile (to avoid up-weighting gradients with very low magnitudes) and divide each gene’s gradient by this number.” Similar question: why did the authors choose the 25th percentile here and not the 20th as above?

9. “Finally we selected one negative SNP for each fine-mapped causal paQTL by requiring that they have identical distances to an annotated PAS and that the negative SNP occurs in a gene with expression levels at most 1.625-fold within the expression levels of the paQTL gene” (698-700). Where does 1.625-fold come from?

10. “For each type of assay (e.g. DNase), we exhaustively search for the window size W that maximizes the spearman correlation between the resulting scalar predictions and the TRIP measurements. Note that when we perform 20-fold cross-validation, we only use the training split of the current fold of the data to search for the optimal window size” (lines 805-807).

Do we understand correctly that not only for every promoter tested in trip-seq, but potentially even for each cross-validation fold of every promoter, a different window size was used? If so, the resulting metric, while not formally wrong, is difficult to interpret and not useful for practitioners, as this finetuning cannot be done for promoters which have not been experimentally measured. The “natural” window size to measure DNase at a promoter is the minimal window that covers this promoter. Why not use this? Also, why do the authors not optimize the window-size for other analyses, e.g. the enhancer prioritization analysis using Enformer (see (a)). Why is this suddenly a concern here but nowhere else?
Overall, we urge the authors to either give reasons for these choices or to benchmark different approaches.

11. A potential pitfall could be that the model predicts better exon boundaries for mRNAs, which are enriched for coding exons with characteristic sequence distributions, than for non-coding RNAs (See also the SpliceAI paper). The author should comparatively assess the performance on non-coding RNAs.

12. In the cited benchmark of Enformer (figure 5A-B of citation [56]), it was shown that Enformer can predict very well the strength of promoters measured in a STARR-seq reporter assay, but struggles to predict the enhancer effect. It was speculated that this is because the enhancer in STARR-Seq is transcribed, leading to post-transcriptional effects that Enformer does not account for, as it was trained primarily on CAGE. It may be interesting to rerun this analysis to see if Borzoi performs notably better, as it could theoretically account for such biases if they are present, although we would understand if the authors believe that the computational cost (c. 600k predictions) is not justified.

Minor points

13. The authors write that “Nevertheless, the predicted exon-to-intron coverage ratio surrounding a potential splice site discriminates between annotated sites and negatives with accuracies on par with the state of the art model Pangolin (Supp Figure S6B)” (lines 286-288). Either the figure S6B is wrong or the sentence is wrong, as the figure strongly indicates that the model is not on par with Pangolin, showing significantly worse performance for both kinds of splice donors analyzed.

14. The authors note that Borzoi outperforms Pangolin on variants close to the splice site but underperforms on far-away splice variants (lines 305-307). The authors note that “Most of these far-away SNPs are de novo splice-gain mutations”, but it is not clear to us what this implies. Intuition would suggest that, if anything, Borzoi should perform better at long-range splicing (due to the attention mechanism and its large receptive field) and worse at close-by events (due to the binning at 32 bp resolution). The fact that this relationship is reversed may warrant further comment or investigation.

15. The introduction states that “From a single RNA-seq experiment, Borzoi derives the primary cell type/state-specific TF motifs and a genome-wide map of nucleotide influence on gene structure and expression” (line 55). It is unclear to us what the authors are trying to say here, but as written, it does not correspond to the manuscript. The model is never trained on a “single RNA-seq experiment”, it is trained on hundreds of RNA-seq and other experiments simultaneously in a multi-task fashion, and can therefore pool information across experiments. It is never investigated what the model learns if it is trained on only a single experiment. Accordingly, this sentence should be clarified or removed.

16. The introduction cites a previous work by the last author which demonstrates a slight advantage to multi-species training (lines 46-47). But this was done only using a particular model design which did not use the attention mechanism. Since models have become significantly larger and more powerful in the meantime, it is not entirely clear to what extent this insight still applies. It may not be within the scope of this paper to investigate this, but in that case this should be clarified. As it stands, it is not shown that Borzoi benefits from the mouse data, nor is it even really investigated whether the mouse predictions are particularly meaningful, since almost all the benchmarks presented are in human.

17. The authors make a number of highly un-intuitive choices when preprocessing the data, such as exponentiating the bin values by 3/4 and taking a square root if the resulting value is above 384, choosing which GTEx samples to include by k-means clustering metatissues with k = 3, and processing GTEx data in an unstranded fashion. Firstly, for reasons of readability, these transformations should be presented as formulas or pseudocode, rather than as text, with all hyperparameters clearly stated. Secondly, the authors should either clearly explain why these choices of transformations, preprocessing steps and cutoffs are important, or note that they were arbitrary choices.

18. Relatedly, while we appreciate that the training data is available at all, it can only be downloaded against a substantial payment. For reproducibility and a wide accessibility, the authors should provide their preprocessing scripts that generate the training data from publicly available resources, particularly in light of the many complex steps described above.

19. Many figures showing quantitative information, e.g. 5A,E,F, 6A,C,F, do not display the scale of the y-axis. Including scales will make these plots more interpretable and will allow the reader to judge to what extent predictions and observations are quantitatively comparable.

20. Figure 3C-D: the Enformer paper helpfully noted the number of enhancers in each distance bin, as well as the number of positive examples in each bin. As the AUPRC is quite sensitive to the class balance and since these metrics generally become more shaky with smaller sample sizes, it would make the figure more interpretable if this information was included. Additionally error bars should be indicated, or some other information should be provided with which the reader can gauge whether the performance improvements vis-a-vis previous methods are statistically significant.

21. The paper should state the hyperparameters used during training of the model

22. In the gradient based attribution methods, the authors always subtract the mean across nucleotides. Why is this done? If it is because of the analysis presented in the paper “Correcting gradient-based interpretations of deep neural networks for genomics”, then this should be cited.

23. The conventional abbreviation for the spearman correlation is the greek letter rho (ρ) and not R, which generally refers to the pearson correlation.

24. The plots showing RNA-seq coverage of Ref and Alt (ex.6C,F) could benefit from including a version with a measure of the difference between them in the y axis, such as the ratio between Ref and Alt coverage. It is sometimes hard to see the changes in coverage claimed by the authors.

25. In the example depicted on 6C the coverage prediction by Borzoi has substantial differences from the ground truth. It generally predicts much more coverage for the intronic regions, the second to last exon and the longest version of the last exon. Such differences are even higher than what can be observed due to the effect of a SNP. This should be commented.

26. For both panel 6C and 6F it is not mentioned why the coverage examples are in Testis and Nerve.

27. In panel 6F there is no scale labels on the attribution scores which makes it hard to understand the score differences between reference and alternate alleles

28. In 4E the legend refers to a precision recall curve when it is a ROC curve

29. In line 624, the upper bound of the summation over the layers should be 7.

30. In the caption for Figure 1, the upsampling operation followed by a convolution is referred to as a deconvolutional layer. This term, though often used, is technically confusing and should be avoided. See:
https://medium.com/@marsxiang/convolutions-transposed-and-deconvolution-6430c358a5b6

31. In Supp Figure S8, the hidden dimension (C) of the Transformer block should be 1536.

Reviewer #2:
Remarks to the Author:
This study presents a novel deep learning model, named Borzoi, which can learn to predict cell- and tissue-specific RNA-seq coverage from DNA sequence. Importantly, the trained Borzoi models are effective for understanding different layers of gene regulation, including transcription, splicing, and polyadenylation via model interpretation and variant effect scoring. It is also shown that Borzoi is competitive with, and often outperforms, state-of-the-art models trained on task-specific regulatory data. Model performance is assessed using highly appropriate measures, and caveats in performance (and hence areas for improvement, which are useful for the field to know about) are honestly presented throughout the study.

There are multiple core improvements for this Borzoi model over other models:
1. It extends previous work in the field for modeling sequencing read count features by moving from local, small windows to much-expended regions capable of incorporating enhancers and RNA elements located at distal regions of a given gene.
2. On the technical side, the model has an improved receptive field size as well as an improved transformer/attention layer context window size. These are very important for modeling the long-range dependencies. Further, its coupling with the U-net structure provides additional support for using distal sequences as well as better resolved output.
3. The showcases presented in this study combine the visualization of attention maps alongside other classic genomic data model-interpretating methods. As demonstrated, this is particularly useful for learning the mechanisms underlying multi-layered gene regulation.
4. The paper demonstrates how a deep learning model trained with RNA-seq coverage data can be used for several important, biologically-motivated, downstream tasks.

Overall, this is a very well-done study with few flaws. That said, there are several opportunities for further improvement. I have grouped them into “Moderate issues”, which I would like to see addressed, and “Minor issues”, which largely comprise suggestions for improving performance or interpretation, but are not viewed as necessary for publishing this particular study - they are possible opportunities for the future that could be either discussed or quickly tested here if feasible.

Moderate Issues
1. Understanding the driving forces behind the performance improvement. While the Borzoi model did present better utilization of distal regulatory sequence features, e.g., in Fig. 3 A-D, it’s unclear where these improvements came from. Is it because of the U-net structure (which provides support for marginally distant sequences)? Can an ablation analysis be performed to show whether this is or is not the case? On top of this, it’s hard to understand Borzoi’s extraordinary performance in Fig. 3C, where the trend of decreasing AUPR w.r.t distance is completely reversed for only the Borzoi model

only within the 130k-262k bin.

2. Training data. Were all data types used by Borzoi necessary for performance? For example, would Borzoi have performed equally well with just chromatin accessibility and RNA-seq? This is of great interest to the field because ATAC-seq and RNA-seq are routine and feasible for most labs as both bulk and single-cell assays. Is it possible to present the weight (relative importance) of each data type or dataset?

3. Use cases. How does Borzoi gene expression prediction compare for highly versus moderately versus lowly expressed transcripts? I'd like to see Spearman correlation coefficients broken down in this way. Similarly, how does Borzoi gene expression prediction break down based on number of exons or gene length? If there are indeed differences, please provide some interpretation.

Minor Issues

4. Suggestion 1 for possible technical improvement. The Borzoi model and its applications extended the frontiers of modeling and understanding multi-layered gene regulation in our field. This improvement is largely based on building its core modeling capability off of its predecessor, the Enformer model. Both models use transformer-based architectures with large receptive fields, plus convolutional kernels as sequence encoders to accommodate large genomic region of interest. Both also stick with 128bp resolved bins in their core transformer/attention layers as a compromise for balancing between feasible computational complexity/cost and better modeling capacity. In this study, the context window of the transformer/attention layers has been extended to around 4k from 1.5k in the Enformer model, enabling much needed attending to gene positions spanned in large region. I am a little concerned that the important internal attention solution is left at 128 bp resolution, thus making it harder for precisely modeling certain interactions in the genome. This was also raised by the authors in the Model section on page 14, claiming that "Conversely, mammalian exons regularly cover fewer than 128bp, and modeling the coverage patterns around these exons at such a coarse resolution can obstruct splice site learning."

There are several tools and tricks that can be used to help extend the context window beyond 4k (and therefore better resolved information for attending) - these include the use of flash attention and multi-query attention. And compared to the approximation-based approach "Scatterbrain" mentioned in the Discussion section, these directly optimized and alternative implementations might offer better performance. It would be fantastic if the authors could train a better resolved model with a larger or better resolved context window in their attention layers to show if such improvement would indeed be needed to help generate better insights and/or emerging knowledge and understanding for any of the downstream analysis.

5. Suggestion 2 for possible technical improvement. Although deep neural network models can and do learn complex relationships from data, there might be a trade-off between (1) multi-task learning of all cell types and data types together (e.g., where the model must learn that HepG2 H34me3 is valuable for HepG2 RNA-seq prediction and less so to K562 RNA-seq, etc.) versus (2) single-task learning of RNA-seq using only relevant data types for each cell type. With the addition of CLIP-seq, could a single-task learning model (e.g., in K562, HepG2, or some other cell type with abundant CLIP-seq) outperform the multi-task? Relatedly, are "singleton data tracks" (RNA-seq from a cell type with no other epigenome data, etc.) helpful to model performance?

6. Additional application/discussion topic. It would be nice for the authors to discuss their views of the opportunity for applying this kind of approach to multiome data (scATAC-seq+scRNA-seq), which in principle would enable the direct tying of enhancers to function. It would be even better to see how the model worked using transfer learning with application to scRNA-seq + scATAC-seq data.

7. Missing details. It would be helpful to include # of positive and negative examples for the prediction tasks in Fig. 3C, D in main text or figure legend, in addition to Methods.

8. Model interpretation. In Fig. 3A, 3B, true positive Borzoi examples are interpreted. It would be interesting to know what features contributed to false positives. (Do the four model folds attend different sequence regions or disagree in some other way…? Are they attending signal in the wrong cell type? Is there any intuition to be gained?) What lessons can be learned from these mistakes?

Reviewer #3:

Remarks to the Author:

SUMMARY OF KEY RESULTS

A very technically dense manuscript. While I hold expertise in disease-relevant RNA splicing and ML, I struggled greatly with the complexity of information in this manuscript. Training and Test data sets for ML and deep learning are an embodiment of critical thinking. If explained well and clearly, any scientist can appraise the critical thinking.

Borzoi is a deep learning neural network trained on the raw DNA sequence, methylation data from different tissues, as well as adjunct genomic data from a variety of assays – that each inform inference of binding sites of transcription factors, promotor machinery and transcriptional activators, splicing regulators, polyadenylation machinery.

The authors claim Borzoi can: 1) predict RNA isoform expression in different cells and tissues based on the DNA sequence alone; 2) identify key transcription factor binding motifs with purported likelihood to influence tissue-specific isoform

expression of genes, though rigorous evaluation of precision and recall using a high confidence dataset is not conducted; to 3) interpret the effects of intragenic SNVs upon RNA isoform expression, with purported application to help prioritise/identify of SNVs associated with human disorders (monogenic disorders, polygenic risk alleles etc).

Whether Borzoi is an incremental advancement or transformative was difficult to discern, due largely to limited explanation of the critical thinking behind, and the specific content of, the training and test datasets for each Figure. An algorithm is only as good as its Training and Test Data. Most Test Datasets used are big, genome-wide datasets. Assignment of True Positives versus True Negatives within the Test datasets was not defined. Further, I was unable to discern if the authors re-trained the Neural Network from scratch without any data from these genes, or whether Borzoi was fed all data other than the empirical RNA-Seq data - or otherwise how sources of predictive bias was mitigated in the Training Data for each Test Set.

Consequently, I am unable to weight Borzoi's predictive accuracy and innovation over other tools. Some Pearson coefficients were low @ 0.53 and some higher @ 0.83. In many instances I cannot grasp exactly what metrics are being compared by Pearson correlation. Discrepancies with other DL tools was glossed over.

REVIEWER RESPONSES TO NG EVALUATION CRITERIA

ORIGINALITY: Borzoi is based on the Enformer algorithm. While some differences are clearly articulated (e.g. training set window set at 32 nt rather than 128 nt), other points of difference/novelty are hard to discern. Figure 4 shows an incremental improvement in Borzoi (AUC 0.794) over Enformer (AUC 0.747).

SIGNIFICANCE: Demonstrating a capability of deep-learning to perceive patterns in raw DNA sequence that can RELIABLY inform tissue-specific isoform expression is highly significant to transformative. However, a lack of clarity around training and test datasets, critical thinking behind the methods, made it very difficult to weight the strength of evidence supporting conclusions made.

DATA AND METHODOLOGY were opaque. Extremely difficult to understand and beyond the comprehension of most Nature Genetics readers. Every approach is technically complex. Nevertheless, it is possible to present a high-level explanation of the critical thinking behind the training and test datasets.

Of particular importance is lack of clarity around Test and Training datasets and what, exactly, is compared or presented in figures. Figure 1: Why EGFR? Without being 100% sure that Borzoi was not trained on any EGRF RNA-Seq data, and without being plainly informed of a significant variation in tissue-specific EGFR isoform expression that you were stress-testing Borzoi's ability to spot, one cannot weight Figure 1B or most of Figure 1. If the algorithm was trained on any EGRF RNA-Seq data, and there is only one main EGFR isoform expressed similarly in all tissues, getting the isoform prediction right is itself predictable, right?

CONCLUSIONS: Though transparently declaring I struggled with this review, I hold reservations about most test datasets. I could not understand what was being shown or compared in many figures. Therefore, I am circumspect and uncertain about most conclusions. I feel Borzoi may well identify key transcription factor binding motifs likely to influence tissue-specific isoform expression of genes. I am not convinced Borzoi can reliably predict RNA isoform expression based on DNA sequence alone. I feel somewhat convinced Borzoi can assist with prioritising SNVs that may affect RNA isoform expression and annotated versus unannotated (erroneous) polyadenylation. I am unsure if Borzoi is superior to other tools.

SUGGESTED IMPROVEMENTS

1. Define acronyms and explain in plain language what each assay does that Borzoi is trained on. For each technique explain simply what you hope the algorithm might 'learn' from this dataset. Clarify the specific nature of RNA-Seq data used from ENCODE i.e. whether PolyA enriched, length of reads etc. These details should be provided for all training data in Methods.

2. Provide high-level explanations of the critical thinking and content of each TRAINING and TESTING dataset. Explain and justify how you mitigated sources of bias in your TRAINING set for each TESTING set.

3. Limit field-specific jargon. Explain terms/techniques like: "saliency scores", "attention matrix". I remain unsure exactly what you are correlating in a "Gene Level Pearson correlation" – relative splice-junction usage across the predicted isoform, akin to a sort of Anova?

4. Explain key results plainly. I am unable to decipher: "To study predictions at the gene-level, we sum coverage in bins overlapping exon annotations and compute log2. Transforming the experimental and predicted RNA-seq coverage similarly, we observe a mean 0.89 Pearson R across genes (Figure 1D). Finally, by quantile-normalizing the gene-level predictions and subtracting the average expression value taken over coverage tracks, we observe a mean 0.62 Pearson R (Figure 1E), indicating that the model explains a significant amount of the variation observed between tracks (such as tissue- and cell type-specific differences)."

5. Consider detailed evaluation of higher confidence, albeit smaller, TESTING sets to enable a non-expert to more readily critically evaluate Borzoi's predictive abilities and accuracy. For example, the ability to understand Figure 1 would be transformed if it focussed upon a 'Hold-out TESTING set' of ~30 genes from among the ~1000 genes that show the most

tissue-specific differences in expression level and isoform expression. Strategically include genes that stress test different aspects of Borzoi's predictive capabilities. Any scientist can understand critical thinking. For example:
A) ~5 'Test Genes' with expression tightly restricted to a specific tissue. Most organs (and particularly skeletal muscle, kidney, eye, heart, brain) have genes expressed predominantly or exclusively in only those tissues. Show Borzoi predicts expression of rod and cone genes in the eye, and nowhere else. This is easy to understand.
B) 5-10 'Test Genes' expressed in a couple of tissues, not in others, and further, select example genes with empirical evidence in RNA-Seq data that confirms tissues-specific isoform expression in the tissues they are expressed in. Show Borzoi can pick this up (or not).
C) 10-15 Test Genes that may be expressed ubiquitously though use alternative TSS in different tissues, some with different patterns of alternative splicing of exons within the gene body, some with annotated use of different PolyA/3'UTRs.
D) Go deep into these genes. Articulate there is no data relevant to these 20-30 genes in the TRAINING set, mitigating possible sources of predictive bias. Expand your analyses to hundreds or thousands of genes if you must – but if so, explain how you have controlled for this by not training your algorithm on any of the data you are asking to offer a prediction for. I can't figure out how you removed all of the relevant data from the Training Set using such large Testing sets.

6. Define exactly what is being measured or evaluated in each Figure. As another example, for Figure 6, what qQTLs were selected in the Test set. Why? How did you control for bias in your Training Set. How exactly did you compare Borzoi with Pangolin? How did you assign each qQTL variant in your Test Set as a True Positive or True Negative. Just like polyadenylaton can occur erroneously at motifs that mimic genuine adenylation signals, cryptic splice-site usage also occurs erroneously. I am an expert in splice-site recognition and usage, hold deep expertise in SpliceAI and Pangolin, and genuinely haven't any idea what you have done or shown in Figure 6

7. If the purpose of developing the Borzoi tool was to assist identification/interpretation of human genetic variation associated with genetic conditions or susceptibility risk factors, I cannot understand rationale for including murine RNA-Seq data. The vast body of scientific literature indicates that variation in gene/isoform expression is what underpins most variation over mammalian evolution. Please justify inclusion of murine RNA-Seq data in the training set.

OTHER COMMENTS:

BOLD STATEMENTS THAT DON'T MAKE SENSE:

"Borzoi accurately predicts RNA-seq and other sequencing assays"
"Borzoi - A neural network for predicting RNA-seq coverage from sequence"

Do you mean Borzoi accurately predicts the pattern of alternative splicing of known alternatively spliced genes from the DNA sequence alone?

To support these statements, one must show more robust case-level data. It is not clear to me why EGFR was selected (known alternative splicing in different tissues?). It is not explicitly clear that all EGRF data was removed from the TRAINING set.

OVERSTATEMENTS:

Borzoi outperformed CADD (mean AUROC 0.54 vs 0.53). Do the authors consider this significant? What variants comprised the Test Dataset? Why? How did you mitigate sources of bias from your Training Set? How did you assign each Test Variant as a True Positive or True negative to event derive a AUROC?

CAVEATS WITH TEST DATASETS.

The eQTL and common versus rare gnomAD variants is encouraging, but there caveats with these Test Datasets (and again is not clear how data was held-out from Training Set). The zygosity of rare variants can be important re inferred likely impact on RNA expression levels or isoform. Similarly, common variants can exert an effect upon RNA expression levels or isoform that is functionally benign. I appreciate the value in a genome-wide approach. However, surely it is more robust and rigorous to evaluate (a smaller) TESTING set of variants empirically confirmed to alter isoform expression - and prove conclusively via precision-recall analyses that Borzoi picks these easy and better than any other tool.

For example, re the ability of Borzoi to interpret SNVs that may affect isoform expression. One could readily curate 50-100 True Positive variants from literature in UTRs or non-coding regions that are classified Pathogenic in ClinVar, which don't affect an annotated splice-site, for which empirical data confirms an affect upon isoform expression levels or processing. Recurrently identified Start loss or Stop loss pathogenic variants from Clinvar might be OK to also include in this mix. Combine these True positives with an equal number of similarly positioned (with respect to location in exons, introns, promoter regions, UTRs) common variants in gnomAD (for the reasons you describe). Sure it's a smaller test dataset, but MUCH higher confidence. This would give compelling evidence for sensitivity and specificity of Borzoi versus other tools.

CLARITY

Numerous abbreviations neither defined nor explained. TF, TF CHiP, ATAC, CAGE assays are not understandable terms. Many clinician geneticists will not be aware what these assays are or show.

The clear explanation in Methods why 32 nt resolution is better than the 128 nt of the Enformer algorithm was welcomed. However, rationale why 32 nt was selected and not another window length is not provided.

Line 376: The authors note “We processed the data slightly differently relative to prior analyses” – but do not explain to which ‘other studies’ they refer.
Similarly Line 394: The authors note “In contrast to previous efforts in which a single train/validation/test split is chosen”, but do not define to which ‘previous efforts’ they refer.

How is one supposed to understand what the authors have done and how it is different or better to other studies?

A lot of important information and critical thinking is buried in high technical detail or is omitted.

Version 1:

Decision Letter:

20th Aug 2024

Dear Johannes,

Your Article, "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

In brief, the reviewers all acknowledge the improvements made in this revision but highlight a number of outstanding issues that require further addressing.

Reviewer #1 asks a number of questions on model performance.
Referee #2 is satisfied and has no further comments.
Reviewer #3 - who is very much representative of the broader, potential user group for Borzoi - has a range of questions, both technical as well as those that would help guide an interested user in applying the model for their work.

In our reading, these requests all seem reasonable and should not require further major expansion of your work to be successfully addressed. We would particularly highlight Referee #3's comments - we think these are some thoughtful requests that would help a reader understand how to apply Borzoi for their analyses.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a “Response to referees” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer

review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics
ORCiD: 0000-0003-1589-7087

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The authors have addressed most of our concerns. Moreover, they have spotted and addressed a train-test leak we had not noticed. We are now raising two major points related to train vs. test performance which we think are important to be addressed.

Major points
---
1. We found a substantial discrepancy between train set performance (with profile Pearson R > 0.85 across tracks, not mentioned in the manuscript) and the test set performance (approximately 0.75 Pearson R, Fig 1C), consistent with the drop in performance between the original manuscript and the revision. Is this gap indicative of overfitting, and could it negatively impact variant analyses or other downstream evaluations? Perhaps, the choice of the smallest test and validation folds (3,4) likely exacerbates this problem. The authors should include a discussion on this topic to highlight potential issues in downstream analyses. For comparison, train vs. test performances should be provided for Enformer.

2. The authors mention (lines 88-90) that test set prediction performance improved with the use of more than one model replicate. Given the quadruple runtime, the benefit of using multiple replicates for a minute improvement (0.74 to 0.75 Pearson R across tracks) should be critically evaluated. Which replicate showed the strongest performance? Single replicate performance should be presented with standard deviation to facilitate a proper study of performance variance, as claimed by the authors (line 83).

Minor points
---

3. In the "Model ablation experiments" section (lines 578-579), the authors refer to a "baseline model" trained on 4 folds. Is this model still available, or is it based on a single fold with 4 replicates? Have the ablation experiments been conducted against the correct common fold? This should be clarified.

4. The authors state that model performance is challenging to compare to Enformer due to differences in data processing.

However, since the data processing for ChIP, DNase, ATAC, and CAGE is similar (except for different binning as noted in line 82), it is unclear why Borzoi underperforms on every common assay compared to Enformer, except for CAGE, even when adjusting for 128bp resolution. This statement should be considerably weakened, and a discussion should be included on which modalities Borzoi is suitable for, and where Enformer still outperforms. The authors may discuss why a more locality-enforcing architecture like the U-Net used in Borzoi performs worse than the Enformer on more local assays.

Reviewer #2:
Remarks to the Author:
The authors have done a fine job of addressing my concerns, no further issues on my end.

Reviewer #3:
Remarks to the Author:
I acknowledge the author's investment of time and energy to carefully and evenly address the concerns and suggestions raised by Reviewers. I also appreciate the author's patient and detailed rebuttal, which is a little easier for a non-expert to understand. For the authors, I declared transparently to the editors that I held no technical expertise in deep-learning. However, it was felt a generalised opinion from a possible future user at the research-clinical interface may be helpful.

Your rebuttal and revisions helps clarify what Borzoi is and does, at least I hope I have it right this time. Trained on small windows of gDNA sequence, uniformly processed RNA-seq from ENCODE and GTEx, paired with training datasets from the Enformer model, the authors trained four Borzoi deep-learning models. I am convinced that Borzoi has a general capacity to identify:
gene loci within the genome; and
promoter and enhancer elements regulating gene expression; and
exons demarked by flanking splice-sites within gene loci; and
transcription start sites, termination sites and polyadenylation signals; to predict
relative levels of transcript expression arising from all gene loci in different tissues; which is a transformative advance for the deep-learning field.

This explains why the Borzoi model works better when also trained on murine data – because core motifs/elements regulating gene expression and transcript content are evolutionarily conserved in mammals. It is the finer tuning of gene expression that diverges between humans and rodents.

In summary, I feel reviewer comments and author responses and emendations collectively support Borzoi as a major advancement in deep-learning. However, I am not yet persuaded Borzoi can "unravel the full transcriptomic consequences of a variant". Trying to interpret one clinical variant as a possible cause of a patient's condition is different to assessing large QTL datasets, which have caveats, as every dataset does. Being able to identify variants affecting gene regulatory motifs proximal and distal to gene loci has potential transformative application for clinical genomics, though the authors don't yet have hard evidence it can do this. A Pearson co-efficient of 0.70-0.85 shows Brozoi is doing a pretty good job at spotting expression QTLs and splicing QTLs but at the single variant level there remains significant uncertainty. I think where I land is that Borzoi may provide a good variant prioritization tool for cases unsolved after genome sequencing to identify variants that may alter transcription of the target gene. A comment below relates to exactly how to use Borzoi for this purpose, which score?

I am grateful for the opportunity to review this impressive piece of work. I have learned a great deal from the authors and Reviewers 1 and 2, who offered important and careful critique of technical aspects of the deep learning. I hope using Borzoi maybe within the grasp of keen users like myself who are not technical deep-learning experts – such that we can put its abilities to good use for clinical genetics, quickly. If the authors would like stress test Borzoi on a large panel of challenging clinical variants with rigorous RNA assay data showing the exact consequences upon splicing and/or expression (incl. UTR, miRNA binding sites, adenylation, TSS and promoter variants) levels then please reach out.

I have a few additional questions and comments:

Q1: Training Set: Can you clarify if the ENCODE datasets included developmental/pediatric specimens from humans and mice. Arguably some of the most significant changes to transcript expression levels and composition is developmentally regulated. I checked methods, Lines 491-511, and could not find details.

Q2: Thank you for providing more clarity around Test and Training Sets. Reading your methods, I understand only one model fold is used for the test set (rather than all four). However, I can't glean the exact nature of hold-out Test Set for Figs 1 and 2. How many genes, why these genes were selected. Specifically with respect to predicting exon coverage for EGFR, can you clarify if all sources of data concerning EGFR were depleted from the Training Set (RNA-Seq + CAGE, DNase, ATAC, ChIP-seq) or was only the RNA-Seq data "held out".

Q4: Lines 512-514. "We fragmented the human (hg38) and mouse (mm10) chromosomes and randomly divided these fragments into eight roughly evenly sized partitions, pairing orthologous regions into the same partition. One partition was held out for validation and testing each, and the remainder of the data (~ 75%) was used for training. This does not clarify

which chromosomal regions are in the Training Data and which are held-out in the Testing Data. SpliceAI (understandably) has blind spots for genes/splice-sites on chromosomes that it was not trained on. I need to know if a deep-learning model is trained on data from the clinically relevant gene/variant I am scrutinizing clinically. Please can you clarify in the methods exactly which human chromosomal regions the model is or is not trained on (if the latter is easier/shorter).

Q5: Lines 125-132. For Fig 1 H and I - what is the real coverage for the variants affecting the splice donor and PolyA signal?
a) Are these real or synthetic variants? If synthetic, why use a synthetic variant?
b) Please provide the HGVS annotation of the donor and adenylation site variants being modelled/assessed? I can't check if exon-skipping is likely because I don't know exactly which variant is being modelled.
c) The real insight is whether Borzoi can accurately predict exactly HOW the mRNA will be altered from loss of the GT. Exon-skipping is only one possible outcome: many donor variants cause activation of one or more cryptic donors and/or intron retention. 50% of all donor variants create more than one aberrant transcript. Do you have the corresponding RNA-Seq data to show that Borzoi got the prediction right – or didn't?
d) Similarly, not all cryptic AATAA/AC PolyA cleavage sites are usable and some PolyA variants just cause loss of transcripts because either adenylation does not occur properly and/or negative feedback to the RNA Pol and abnormal transcription termination. Did Borzoi get it right? Was the alternate PolyA site used as predicted?

Comment 6: Lines 386 – 387 is an overstatement. Moreover, the authors state Pangolin offers the same prediction and therefore the word unique is not appropriate.
a) The RNA-Seq coverage plot shown in Fig 7C shows high relative levels of intron retention – is this reflective of the natural expression pattern in the relevant specimen? Leaky expression of genes in testis often doesn't mirror the clinically relevant specimen. It may be more prudent for the authors to choose a sQTL to highlight that holds clinical relevance to a particular tissue and show RNA-Seq from this tissue, rather than leaky expression in testis.
b) The alternate acceptor activated in testis by the sQTL is annotated. Can the authors confirm Borzoi was not trained on any RNA-Seq or CAGE data for this gene?
c) It also may be helpful for readers trying to evaluate the variants being assess to provide the full HGVS annotated that details the gene, variant and chromosomal position.

Q7: It is not clear from reading the manuscript how Borzoi can be deployed to screen whole genome sequencing data to prioritise variants with increased likelihood of altering gene expression. It is possible for the authors to provide a plain explanation regarding how to use Borzoi for this purpose? For example, does Borzoi return a predictive score of some kind and there is a threshold score the author's recommend as a 'high confidence prediction'? It would also be helpful to have a rough idea how many candidate variants Borzoi identifies from genome sequencing for a heterogenous condition with ~100 associated genes (e.g. epilepsy), whether it identifies 10 or 10,000?

Comment 8: As a final remark, the title is not intuitively understandable for me and therefore may not be for other Nature Genetics readers. Though the decision re title wording rests with the authors. The phrase 'RNA-Seq coverage' is inherently related to RNA-Seq read-depth which can vary from 20M reads to 1 billion transcriptomic reads and is not entirely the same thing as relative transcript expression levels.

Version 2:

Decision Letter:

Our ref: NG-A63358R1

2nd Oct 2024

Dear Johannes,

Thank you for submitting your revised manuscript "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation" (NG-A63358R1). It has now been seen by the original referees #1 and #3 and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics

ORCiD: 0000-0003-1589-7087

Reviewer #1 (Remarks to the Author):

The authors have addressed all my open points.

Reviewer #3 (Remarks to the Author):

The authors have addressed my remaining comments.

I had surmised all data sets must have been held out from the Training Set - but could not be certain. Equally important is specifying the donor and PolyA variants modelled in Fig1 are synthetic. This was not obvious. Being synthetic variants, and with no way of proving if the Borzoi prediction is accurate, I agree these examples hold little utility for presentation in Figure 1. A better option would be to find a published article describing RNA assay data for pathogenic splicing variant causing a Mendelian disorder in a gene within your Testing Set.

I leave the decisions about which figures to go where to the authors and editorial team.

Inclusion of overview statements related to which scores to assess for different classes of variant is also welcomed. The best outcome is if deep learning tools can be harnessed by both fundamental and clinical researchers alike - with the latter needing user-friendly interfaces and simplified guidelines how to use it, including author recommendations regarding what constitutes a high confidence score from a low confidence score.

Version 3:

Decision Letter:

In reply please quote: NG-A63358R2 Linder

4th Dec 2024

Dear Johannes,

I am delighted to say that your manuscript "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation" has been accepted for publication in an upcoming issue of Nature Genetics.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Genetics style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf, and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: https://www.nature.com/authors/policies/embargo.html

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release,

but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that *Nature Genetics* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we strongly recommend that you upload the step-by-step protocols used in this manuscript to protocols.io. protocols.io is an open online resource that allows researchers to share their detailed experimental know-how. All uploaded protocols are made freely available and are assigned DOIs for ease of citation. Protocols can be linked to any publications in which they are used and will be linked to from your article. You can also establish a dedicated workspace to collect all your lab Protocols. By uploading your Protocols to protocols.io, you are enabling researchers to more readily reproduce or adapt the methodology you use, as well as increasing the visibility of your protocols and papers. Upload your Protocols at https://protocols.io. Further information can be found at https://www.protocols.io/help/publish-articles.

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics
ORCiD: 0000-0003-1589-7087

Response to Referees - Borzoi Manuscript, Nature Genetics (NG-A63358)

Reviewers' Comments:

(Author responses = blue), (Added responses since May 28, 2024 = red)

We thank the reviewers for their helpful comments and suggestions to improve the manuscript. Below we provide a summary of major updates and additional analyses performed in response to feedback (from reviewers and from peers & colleagues in the field). We also give detailed responses to each reviewer comment further below.

Summary of major updates in the revised manuscript:

In response to reviewer comments:

1. We have performed a comprehensive ablation study where we retrain instances of the Borzoi model and hold out various training data modalities (e.g. remove all modalities except RNA-seq, or remove all mouse data), or where we change certain aspects of the architecture (e.g. no U-net). The performance of the models is evaluated on the CRISPRi gene-enhancer datasets (Supp Figure S4J) as well as GTEx RNA-seq gene-level test predictions and eQTL coefficient prediction (Supp Figure S5D). Overall, we find that training on additional assays (e.g. DNase, ATAC) helps learn salient features which improves performance for RNA-seq-based predictions, and that multi-species training further boosts performance. This analysis was done in response to:
    a. Reviewer #1 - Comment #16
    b. Reviewer #2 - Comment #1 + #2 + #5
    c. Reviewer #3 - Comment #7

2. We conducted a more in-depth analysis and visualization of genes that exhibit highly tissue-specific regulation, ranging from overall differential expression to tissue-specific 3' UTR isoforms or alternative transcription start sites. We evaluated Borzoi's performance at predicting this tissue-specificity for each regulatory mechanism on test genes that were held out during training. We also moved the (negative) alternative splicing results earlier in the manuscript alongside these analyses (it previously resided in Supp Figure S6 of the original). We hope that this new analysis, which is presented in Figure 2 and related Supp Figure S2, brings clarity to what Borzoi can and cannot predict in terms of tissue-specific regulatory patterns from RNA-seq. This analysis is in response to:
    a. Reviewer #3 - Comment #5

3. There were questions regarding the use of multiple attribution methods for different tasks in the original manuscript. We have now clarified that we generally believe ISM to be a more robust and well-calibrated method for variant interpretation than Input x Gradients (for a number of reasons, see the updated Discussion on Line 454). However, there are exceptions. First, it is prohibitively computationally expensive to run ISM on the entire

524 kb sequence for the purpose of de novo motif discovery, which is why we resort to gradient-based attributions for analysis with MoDISco (for TFs as well as 3' UTR and splicing motif discovery). Second, we noted extensive buffering effects when we used ISM to interpret polyadenylation signals, which is why we use ISM Shuffle as the primary method of interpretation for the 3' UTR. We have included a supplementary analysis in Supp Figure S6B which exemplifies this buffering phenomenon. For splicing, there is less buffering (see the corresponding analysis in Supp Figure S7A). We hope that these new results will guide readers through our rationale for when and why we use different attribution methods. This is in response to:

   a. Reviewer #1 - Comment #4 + #6

4. We have included a new benchmark comparing Borzoi and Enformer on the saturation mutagenesis MPRA of Kircher et al. (2019), which is presented in Supp Figure S5G-J. We find that Borzoi outperforms Enformer on the majority of promoters tested, while the accuracies are near-equal on enhancers. We also note that the ensemble approach improves Borzoi's accuracy, even though a single replicate still outperforms Enformer on most promoters (see Supp Figure S5J - only the full ensemble recovers all salient motifs measured in the MPRA). This analysis is in response to:
   a. Reviewer #1 - Comment #2

5. We have made the scripts used to process the training data, as well as scripts for training Borzoi, available at 'https://github.com/calico/borzoi/tree/main/src/scripts/data'. We have also made available the code relevant for replicating the evaluations of the paper, located at: 'https://github.com/calico/borzoi-paper'. This is in response to:
   a. Reviewer #1 - Comment #18

Other changes (e.g. in response to feedback from peers):

6. We have removed the example attributions using SmoothGrad in Figure 1, seeing as that interpretation method was not critical or used in any subsequent analyses.

7. We have changed the distance bins in the CRISPRi benchmark (Figure 4). Previously, the second furthest distance bin was 60kb - 130kb. Since Enformer only predicts up to 98kb, it is cleaner to place a bin boundary at this distance. We now set the bin to 45kb - 98kb, which has full receptive field support for both Enformer and Borzoi and represents Enformer's most distant examples.

8. There was an incorrect calculation in the original gene-level test set evaluations (Figure 1D-E) where the predictions (and measurements) were scaled to be 32x too large. This has now been fixed, which resulted in updated Pearson R metrics in the revised manuscript. There was another bug that affected some of the example variant visualizations in Figure 5-7, where the annotated ref/alt coverage values were incorrect. This has also been fixed, resulting in slightly different numbers annotated in the visualizations. These fixes did not significantly change the main results or conclusions.

9. During our continued analysis of the Borzoi models described in the manuscript, we identified a bug in the python training scripts. The bug caused all models to have been trained on the same train/test split, in contrast to our original intention of training each replicate model on a different train/test split. Thus, the variation between models arises solely from random weight initialization and random sequence batch processing order. We have revised the Methods description to reflect this. Furthermore, we have revised some reported metrics and visualizations in Figure 1 and 2 (which were the only analyses looking at held-out test sequences of hg38) to ensure that they all make use of a correct test set evaluation. Specifically, the following figures have been updated:
   - Figure 1B-E, and Supplementary Figure S1A-J.
   - Figure 2A-F, and Supplementary Figure S2A-M.
   - Supplementary Figure S3D.

   The only affected comparisons to other models are those with Enformer in Supplementary Figure 1A-D, in which we scatter plot Pearson correlation achieved by Borzoi and Enformer on each model's test set. Our revised analysis indicates that Enformer, a larger model asked to do less, predicts some chromatin accessibility assays with slightly greater accuracy than Borzoi, while Borzoi outperforms Enformer at CAGE gene expression prediction. We also show that these metrics are influenced by differences in data processing (such as bin resolution), making exact comparisons hard. Importantly, Borzoi further adds the many capabilities unlocked by RNA-seq training described throughout the manuscript. All analyses of genetic variants and CRISPR perturbations were unaffected by this bug, and Borzoi still outperforms Enformer on these downstream tasks.

Additional notes:

1. In response to reviewer suggestions, we have added a large number of supplementary analyses to the paper. These new analyses have resulted in very long supplementary figures & captions related to each main figure, and we acknowledge that these supplementary figures are unwieldy in their current format. However, for the sake of not altering the structure of the supplementary material too much (and causing confusion by major reformatting) while revising the content together with the reviewers, we decided to keep the structure (one supplementary figure per main figure). However, if and when the paper is accepted, we will work together with the editor to split the supplementary figures into smaller, more manageable subfigures and/or extended data figures.
2. All references to figures (e.g. Figure 2) or line numbers (e.g. Line 454) refer to the updated numbers in the revised manuscript, unless otherwise stated, in the responses.

Reviewer #1:

Remarks to the Author:

Linder et al. describe a novel deep learning model, called “Borzoi”, that can predict RNA-seq coverage in a variety of human tissues at a resolution of 32 base pairs. The model is trained on ENCODE and GTEx and presents a substantial methodological advance over the previous state-of-the-art model of human transcriptional control Enformer. By predicting RNA-seq coverage directly, the model overcomes some major limitations of previous approaches and allows interrogation of transcriptional and post-transcriptional regulation.

Overall, we believe that this paper will be highly valuable to the field and congratulate the authors on their hard work. However, currently several aspects of the manuscript are unclear and may give a wrong impression to the reader. Moreover, the usefulness of the model for the field is presently limited by the fact that the authors present many ways to employ it, without ever making a clear statement as to which method is to be preferred.

We thank the reviewer for their detailed feedback and suggestions on how to improve the manuscript. We have addressed each of the reviewer’s comments below.

Major points

1. The Authors claim (line 188, a section title) that Borzoi “predicts the impact of genetic variation on gene expression”. This is not shown in the manuscript. What the authors show is (1) that a random forest trained on the outputs of the model can be used to classify likely eQTLs vs. matched negatives and (2) that there is a (limited) spearman correlation between the model prediction and the normalized eQTL effect size. Since the former is a classification task and the latter only measures the concordance in terms of rank, neither of these really correspond to “predicting the impact of genetic variation on gene expression”. To do this, the authors would have to demonstrate that the model can quantitatively predict fold changes due to variants. Accordingly, the section should be reworded to reflect that currently it is only shown that the method can rank or prioritize variants.

We agree with the reviewer that this section title could be improved to better reflect the content. We have reworded it to: “Borzoi prioritizes causal genetic variants that impact gene expression” (Line 282).

2. Related to the first point, two analyses would be valuable to better gauge how the model fares in predicting quantitative effects of variation. First, the Enformer paper benchmarked their method on saturation mutagenesis data. This data also is imperfect, of course, but nevertheless a decidedly more quantitative measure of the ability of the model to predict the effects of local variants on expression while sidestepping the problem of linkage. The authors should include this benchmark or provide a compelling reason why this should not be done.

The saturation mutagenesis MPRA data from Kircher et al. (2019) is indeed imperfect for evaluating Borzoi's ability to infer the quantitative effect of variants on gene expression, perhaps even more so than it was for Enformer. The MPRA data measures the effect of variants outside of their endogenous context by sequencing mRNA expression from a mini-gene reporter. Consequently, these measurements ignore the spatial arrangement of the promoters/enhancers to their regulated genes, as well as the gene's transcript structure. These transcript structures are remarkably diverse, and it is largely unknown whether or not promoters/enhancers measured on mini-gene reporters generally behave similarly as when connected to their native genes. In contrast, Borzoi estimates effects on gene expression by aggregating the changes in predicted coverage across these transcript structures. It is thus speculative whether or not Borzoi's predicted fold changes will be well-calibrated with respect to the MPRA measurements and it is hard to draw any strong conclusions.

Nevertheless, we performed this benchmark comparison against Enformer and include it as a supplementary analysis, presented in brief in the revised main text starting on Line 326:

"Finally, in Supp Figure S5G-J we compare Enformer and Borzoi on a subset of the Massively Parallel Reporter Assay (MPRA) data from Kircher et al. (2019). In brief, we found that Borzoi surpassed Enformer's concordance with measured variant fold changes on the majority of promoters tested, and matched the performance on enhancer loci. Borzoi's ensembling strategy contributes significantly to the improvement, as some motifs with large measured effects are not detected by the predictions of the first model replicate but discernible with the full ensemble (Supp Figure S5J)."

Figure panels related to the benchmark are found in Supp Figure S5G-J. Details of the benchmark are presented in a new Methods section titled: "Saturation mutagenesis MPRA benchmark" (Line 1019).

To mitigate the aforementioned problems with comparing fold changes, we use Spearman correlation coefficients as the primary metric for assessing each model's ability to rank the effects of the MPRA variants. We demonstrate competitive performance compared to Enformer, and Borzoi outperforms Enformer for the majority of promoters tested. While Borzoi outperforms Enformer on some promoters using a single model replicate, we found that our model ensembling strategy drives a substantial part of the performance improvement.

3. Second, the benchmark of Enformer cited by the authors (citation [56]) claims that the predicted effect of (validated) enhancers decays strongly as a function of distance to the TSS. To what extent is this also the case for the RNA-seq predictions of Borzoi? Additionally, is the effect of enhancers uniform across the RNA-seq track (indicating that the model uses distal elements to modulate its overall prediction of the transcription of a gene) or are the genic bins closer to an enhancer/variant affected more heavily than those further away? Answering these questions will help practitioners gauge how to compare predicted effects at different distances.

We agree that these are two important questions to quantify for this particular benchmark. To this end, we performed a follow-up analysis similar to the one done by Karollus et al. (2023), where we dinucleotide-shuffle the putative enhancer (using a 2kb window) of every measured enhancer-gene pair in both the FlowFISH and Gasperini data and record the total % change in exon-aggregated K562 RNA-seq coverage due to the shuffle. Scatter plots of predicted % change as a function of distance from the TSS are shown in Supp Figure S4C.

As is annotated in the figure, the average predicted % change of positive examples indeed declines for binned distance ranges that are farther away from the TSS. However, as is also annotated, the fold change between the average % change of positive examples and the average % change of negative examples does not decline monotonically and is at least as large for the most distant category of enhancers as it is for the least distant bin. Hence, the same (high) specificity is largely preserved at distant enhancer loci. In Supp Figure S4H-I, we show that Borzoi's AUPRC surpasses that of Enformer on the FlowFISH data in all distance categories when using the dinucleotide-shuffle approach instead of input gradients, however the performance increase is much more pronounced at farther distances. For the Gasperini data, performance even gets slightly worse than Enformer in the distance category closest to the TSS when only comparing the dinucleotide-shuffle methods. We speculate that this discrepancy could arise from enhancer loci that are within 2kb of exons of the target gene, which would generate very large perturbation scores due to breaking the exon junction(s) by shuffling (Borzoi is more sensitive to exonic features than Enformer; even when breaking exons of other genes it seems). Input-gradients with a more narrow gaussian smoothing might be less prone to generating artificially high scores this way. We visualize false positive predictions in Supp Figure S4D which exemplifies this phenomenon.

The new analysis is described starting on Line 245 of the main text:

"We compared our gradient-based analysis to an in-silico perturbation approach where % change in RNA coverage across the exons of the target gene is recorded upon dinucleotide-shuffling the candidate enhancer. This allows us to observe the total predicted change in expression due to enhancer disruption (which correlates with, but is not necessarily identical to, local gradient saliency). In line with recent work [73], we find a general decreasing trend in % change as a function of distance from the TSS. However, we also find that the relative fold change in average perturbation magnitude between positive and negative examples remains largely consistent across distance categories, suggesting specificity is maintained even at far distances (Supp Figure S4C). Manual inspection of false positive predictions suggest that loci near exons of the target gene, or other genes, may be a source of discrepancy between the model and measurements (Supp Figure S4D)."

More details are also given in the Methods section titled "Gene enhancer prioritization task" (Line 993).

To answer the second question posed by the reviewer ("are the genic bins closer to an enhancer/variant affected more heavily than those further away?"), we performed an additional

analysis that is presented in Supp Figure S4E-G. We again dinucleotide-shuffle the putative enhancers using a 2kb window, but this time we recorded the relative drop (% change) in coverage at bin-level resolution for those bins that overlap the exons of the target gene. We found that the pattern of drop in coverage in some cases remains constant across all exonic bins (Spearman R ~= 0), while in other cases the (negative) drop in coverage diminishes as a function of distance to the enhancer (Spearman R > 0), and yet in other cases the drop increases as a function of distance (Spearman R < 0). We observe all three of these cases regardless of whether or not the enhancer occurs upstream of-, downstream of- or within the gene body. The distribution of Spearman coefficients across all positive Gasperini examples skews to positive numbers. However, on average the relative fold change between the magnitude in drop in coverage at the 20% most distant bins vs the drop in coverage at the 20% least distant bins is 0.95x, meaning the bias is low on average across genes.

The new analysis is described starting on Line 253 of the main text:

“Furthermore, the in-silico perturbation method lets us examine the % change in coverage at individual exon-overlapping bins of a target gene. We find that the bin-level drop in coverage often exhibits a non-uniform pattern that is gene-specific (Supp Figure S4E). While we observed a general trend where coverage drops less in exonic bins furthest away from the enhancer, there are also many genes where this trend is reversed (Supp Figure S4F-G). Overall, any global distance bias is minimal; averaged across all positive examples from Gasperini et al., the drop in coverage at the 20% most distal exonic bins is ~0.95x of the drop in coverage at the 20% proximal-most bins. Finally, the choice of shuffle window size only marginally affected overall accuracy (Supp Figure S4H).”

4. Furthermore, as the authors themselves note, one of the main uses of their model likely will be in interpreting variants. However, many of the design choices by the authors make this nontrivial. Firstly, because of the ensembling strategy, there is not just one Borzoi, but four models. Secondly, sometimes attributions or predictions are averaged over forward and reverse complements. Thirdly, the authors use a variety of different attribution methods (gradients, smoothed gradients, ism, “ism-shuffle”) in a very task-specific manner, without really explaining why a specific method was chosen for a particular task. As a result, after reading the manuscript, it remains unclear to us what the recommended way to interrogate variants with Borzoi is. Especially because running this model is very slow and expensive, it would severely limit the useability if users had to try every combination of possibilities explored in the paper themselves. Accordingly, the authors should evaluate the advantages of these different approaches and determine best practices for employing their method.

We agree that the original version of the manuscript did not give a clear recommendation of best practices for using the model to interpret variants, and that the choice of presenting a particular attribution method for a given task in the main figure was somewhat unmotivated (even though all three attribution methods were presented in each corresponding supplementary figure).

First, regarding the use of ensembling, we showed in the original manuscript that the eQTL prediction task improved upon ensembling (see Line 304 of the revised manuscript), however using a single replicate already improved upon Enformer. In new follow-up experiments, we now also show that: (1) the enhancer prioritization benchmark benefits from ensembling, but a single replicate still is superior to Enformer within all distance categories of the FlowFISH data and most categories of the Gasperini data (presented on Line 262; see also Supp Figure S4I), and (2) the saturation mutagenesis MPRA benchmark also benefits from ensembling, but a single replicate is still competitive with Enformer (Line 326; see also Supp Figure S5J). Thus, we feel that in the revised manuscript we present plenty of evidence demonstrating that while a single model replicate is already competitive with other state-of-the-art models, there is a clear advantage to using all 4 replicates. Any researcher wanting to use Borzoi can now decide, based on these metrics, whether or not they want to incur the additional computational cost of using the full ensemble and gain a moderate but significant boost in performance.

To summarize these results, we have added the following text to the Discussion starting on Line 422: “By averaging predictions across an ensemble of model replicates, Borzoi's performance improved further, although a single replicate already outperformed other models on several benchmarks.”

Second, regarding the use of different attribution methods in a task-specific manner, our reasoning is as follows: ISM is our preferred/ideal method for sequence attribution over gradients, for a number of reasons: (1) input gradients have previously been shown to occasionally be miscalibrated or otherwise a less accurate quantification of total importance than ISM (see for example Figure 4 in the ICLR paper from Ancona et al., 2018; https://openreview.net/forum?id=Sy21R9JAW; here ISM is similar to “Occlusion-1” except that Occlusion replaces features with 0’s), and (2) ISM scores are more intuitive (the scores are calculated by making syntactically valid perturbations to the sequence, i.e. point mutations, and there is a natural reference, namely the original unperturbed sequence). However, ISM is computationally intractable to use in conjunction with Borzoi over large stretches of sequence, making it ill-suited when we want to discover salient TF- or RBP motifs anywhere in sequence *de novo*. Hence, we are forced to use input gradients to generate attributions for *de novo* motif discovery (which is why gradients are used in Figure 3A-C, Figure 4A-D, Figure 6B, and Figure 7B). Luckily, as we show in Supp Figure S3C, the saliency scores computed from input-gated gradients are largely consistent with ISM scores (Spearman R >= 0.70 for all 5 tissues tested).

While our preference would have been to rely solely on ISM scores for any kind of variant interpretation in short predefined regions, we found early on that many repeat-like motifs (e.g. poly-T stretches) near polyadenylation signals in the 3’ UTR exhibited strong buffering effects that could effectively “hide” motifs with ISM. This led us to use ISM Shuffle as the primary interpretation method for variation in the 3’ UTR. In contrast, for other regulatory mechanisms, e.g. splicing, there were much less apparent buffering effects. This, in combination with the knowledge that many splice-regulatory motifs are short (e.g. the canonical part of splice donors and acceptors are only 2nt wide) made us comfortable with keeping ISM as the primary method. We have added new follow-up experiments that exemplify these observations, for both

polyadenylation and splice sites. Specifically, in Supp Figure S6B (and main text starting on Line 347), we have added an in-silico experiment where consecutive T nucleotides are sequentially added to a region downstream of a polyA signal. It is clear that, while the polyA site strength (calculated from predictions) monotonically increases, the ISM score across the T-stretch starts decreasing already after 6 consecutive T's and becomes ~0 again after 9 T's. In contrast, repeating the analysis for a splice acceptor site in Supp Figure S7A by incrementally growing a polypyrimidine tract shows much less buffering in the ISM scores (the average ISM score is >> 0 after 9 C's or T's). Hence, our recommendation is to use ISM Shuffle when interpreting variation related to polyA signals.

In order to make these recommendations clear to readers, we have added the following paragraph to the Discussion (starting on Line 454):

"We also emphasize the challenges in choosing appropriate attribution methods to interpret the model. When discovering TF- or RBP-motifs *de novo*, i.e. when these elements can occur anywhere in the input sequence, it is computationally intractable to use an ISM-based approach and we resort to input-gated gradients, even while recognizing the potential miscalibrations of this approximation. For variant interpretation, ISM-based contribution scores are preferred; the scores are calculated by making syntactically valid perturbations to the sequence (point mutations) and there is a natural reference point (the original, unperturbed sequence). However, we noticed that buffering effects near polyadenylation signals can hide important features from ISM. We thus recommend using window-shuffled ISM for interpretations in the 3' UTR, while ISM is the preferred method for splice junctions and local enhancer/promoter regions. We look forward to exploring methods such as DeepLIFT or SHAP as potential unified replacements."

5. Related to this, the many different attribution methods, Gaussian filters, cutoffs and other processing steps employed in the paper do give the authors additional degrees of freedom, which raises the concern that sometimes the analysis method is tailored to make the results look more impressive in a way that is not generalizable. Specifically, in Figure 3 C-D, the authors compare the ability to prioritize distal enhancers to Enformer. In the Enformer paper, this analysis was done using three methods: gradient*input, attention and ISM. In this paper only the gradient method was used, although the authors additionally "smooth the scores… using a Gaussian filter" (which to our knowledge was not done in the Enformer paper). How were the parameters of this Gaussian filter determined? Why was this method used and not another? Also why do the authors evaluate the gradient for Enformer based on "two 128 bp bins centered at the gene's TSS" (line 817-819) whereas the original Enformer paper used three bins: "We note that 'CAGE at TSS' always means summing the absolute gradient values from three adjacent bins, as is also done in gene-focused model evaluation. The three bins include the bin overlapping the TSS and one flanking bin on each side" (Enformer paper, section "Enhancer prioritization"). Moreover, the Enformer paper specifies that "The enhancer–gene score was obtained by summing the absolute gradient × input scores in the 2-kb window centered at the enhancer.", whereas in this paper, we did not find any information as to the size of the window around the enhancer. Given the overlap in authors between these two papers, it is difficult to understand why the details of this benchmark differ. To make the results interpretable, the

authors should either run the benchmark exactly as was done previously for both Enformer and Borzoi (as 128 is a multiple of 32, the difference in bin size does not prevent this). Or the authors should explain why the new parameters they have chosen are superior to those chosen previously.

As we were developing/adapting the enhancer prioritization analysis for Borzoi, we noticed early on that computing a weighted average score using a (relatively wide) Gaussian filter consistently performed better (in terms of higher AUPRC values) in all distance categories of the two CRISPR benchmarks compared to aggregating the absolute-valued nucleotide scores in a 2kb window (unweighted). The Gaussian filter has a standard deviation of 300 (listed on Line 1009 in the Methods section) and is truncated at numpy's default value of 4 (see https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.gaussian_filter1d.html). The exact value of the standard deviation was not specifically optimized, but was chosen to be roughly the size that we believe salient core enhancer syntax normally is (2-300 nt). Hence the total effective window in which we aggregate absolute-valued gradient saliencies is 1.2kb, but the nucleotide scores on the edges contribute much less to the final score than the center nucleotides. Importantly, we saw a performance improvement for Enformer using this method.

To make the point clear that the Gaussian-weighted average aggregation helped both Enformer and Borzoi on this benchmark task, we have provided a new extended analysis in Supp Figure S4I. From this it is clear that Gaussian-weighted averaging improves performance for both Borzoi and Enformer on the FlowFISH data while being at least as good as the original unweighted 2kb averaging on the Gasperini data. We have added the following new text to the Results section, starting on Line 260:

"To understand what drives the improved performance in prioritizing enhancers with Borzoi, we performed a number of ablation studies. First, we tested how various parameter choices affected performance when aggregating the gradient- or shuffle-based scores, for both Enformer and Borzoi (Supp Figure S4I). Briefly, we found that a fraction of Borzoi's improved accuracy was due to model ensembling, and that gaussian smoothing often was superior to averaging the gradient score in a discrete window, for both models."

In Supp Figure S4I, we also explicitly compare different choices in how to compute and aggregate the gradients from Enformer's CAGE predictions. Specifically, we compare the effect on AUPRC when aggregating the signal in either 2 bins centered at the TSS or 3 bins (default from Enformer paper), and we also compare the effect of either averaging the final score across multiple TSS's (for genes with more than 1 TSS) vs taking the score of the maximum-predicted peak. Using 2 bins was consistently better than using 3 bins on the FlowFISH data, while 3 bins were superior on the Gasperini data. Since the gap in performance was larger on the FlowFISH benchmark (which is generally believed to be less noisy), we conclude that using 2 bins can be considered marginally better than using 3 bins. Furthermore, averaging across multiple TSS was marginally better than taking the maximum score. In Main Figure 4C-D, we use the best combination of parameters when presenting Enformer's results (computing gradients from 2 CAGE bins, averaging across multiple TSS).

The details of this analysis, including the rationale for choosing these parameters for Enformer, are added to the Methods section titled “Gene-enhancer prioritization task” starting on Line 1003: “For both Enformer and Borzoi, we scored putative enhancers using input gradient analysis. For Enformer, we computed the gradient of the K562 CAGE prediction in the two 128 bp bins centered at the gene’s TSS. Computing the gradient using 3 bins (as in the original paper) resulted in marginally worse performance. The gradient score statistic was averaged for genes with multiple TSSs, which performed better than taking either the max or sum. For Borzoi, we computed the gradient of the K562 RNA-seq prediction for all bins overlapping the gene’s exons in GENCODE v41. For each nucleotide, we took the absolute value of the reference nucleotide gradient. For each regulatory element, we computed a weighted average of the nucleotide scores using Gaussian weights (standard deviation 300), centered at the element’s midpoint. This approach improved performance for both Enformer and Borzoi compared to a simpler strategy of averaging the absolute-valued gradients in a 2kb window centered on the enhancer. To calibrate scores across genes with different expression levels, we normalized the scores by the mean nucleotide score across the entire region.”

6. Also, why are ISM-based, and not gradient-based analyses, then used for the splicing and polyadenylation examples? Did gradient-based methods not work for these cases?

This comment overlaps to a large degree with Major Point 4 above. In the response to point 4, we argue that we (need to) use gradient-based methods for *de novo* motif discovery anywhere in the sequence (which is used in conjunction with MoDISco for expression, polyA and splicing analyses) because of the computational intractability of running ISM on large sequence regions.

In that response, we also argue that ISM is a better, more intuitive attribution method (that is often better calibrated) for interpretation of short local regions, and that is the reason why we focus on ISM attributions in the context of variant interpretation (as is done for eQTLs, paQTLs and sQTLs). For polyadenylation variants, we resort to focusing on ISM Shuffle in the main figure because we detect troublesome buffering effects that effectively hide some features in ISM attributions (see the new follow-up analysis presented in Supp Figure S6B).

Also, it’s worth emphasizing that, while ISM and ISM Shuffle are displayed in the main eQTL, paQTL and sQTL figures, we also provide attributions based on input-gated gradients in every supplementary figure as a way of presenting the differences transparently.

7. Why is the overall ability to predict variation in coverage between exons and introns “evaluated on the top 20% of test set genes with highest variance in coverage across the span of exons and introns” and not, say 10% or 25% or 50%? What happens beyond 20%?

As the variance in bin-level coverage across the span of a given gene decreases, it makes less and less sense to evaluate Borzoi’s (or any model’s) ability to predict this variation as it will be driven more by measurement noise than a deterministic sequence component. However, we agree with the reviewer that the exact choice of 20% was chosen somewhat arbitrarily. To

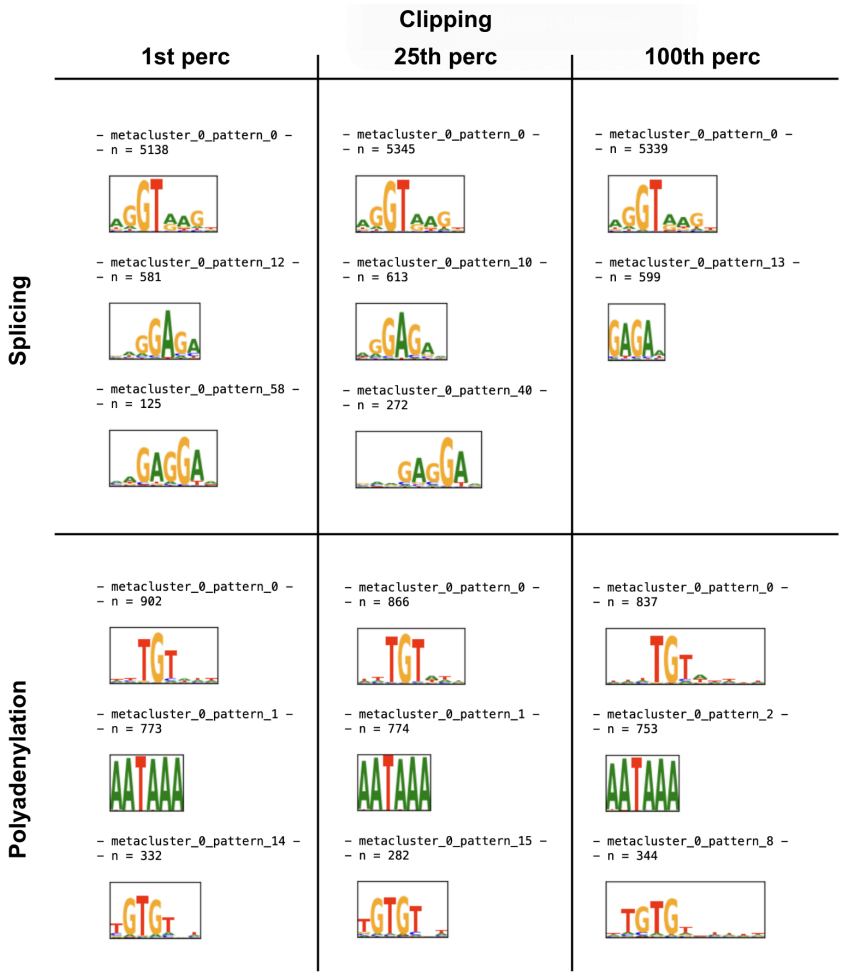
remedy this, we have included a new analysis in Supp Figure S1I, which details the relationship between the choice of variance percentile cutoff (x-axis) and Borzoi's ability to predict the variation within genes passing this threshold (y-axis, measured as Pearson R). We have updated the following text in the Results section, starting on Line 100:

"Finally, we note that Borzoi accurately predicts variation within the transcript structure; evaluated on the top 20% of test set genes with highest variance in coverage across the span of exons and introns, the average Pearson R between predicted and measured RNA coverage (at bin-level) was 0.88 across all genes and samples (Supp Figure S1I)."

8. "We clip the lower end of the 4, 778 maximum deviations at the 25th percentile (to avoid up-weighting gradients with very low magnitudes) and divide each gene's gradient by this number." Similar question: why did the authors choose the 25th percentile here and not the 20th as above?

We wanted to clip the lower end of the calculated maximum deviations at a reasonably large percentile so that gradient magnitudes would be within the same order of magnitude (for splice sites we are less interested in the difference in absolute magnitude of gradients across genes, which on the other hand we are interested in for expression analyses). However, we agree with the reviewer that the exact choice of using the 25th percentile was chosen somewhat arbitrarily.

Ultimately, we want MoDISco to produce reasonably stable results. To understand whether or not the choice of this parameter matters much for the downstream analysis, re-ran MoDISco on the same set of gradients, for both splicing and polyadenylation motif discovery, but we either clipped the maximum deviations at the 1st percentile or the 100th percentile. In the end, we identified most motif clusters across all parameter choices, and the number of supporting seqlets did not vary substantially. See the attached figure below (**Response Figure A**) which compares some well-known splicing- and polyadenylation motifs parameter choices (SRSF1 was not found among the splice clusters for clip = 100th perc.). We can conclude that this parameter did not fundamentally change the results.



Response Figure A: *Splicing and polyadenylation motif clusters (and their PWMs) when running MoDISco on coverage ratio gradients which have had the lower end of their maximum deviations clipped at different percentiles.*

To make readers aware that this parameter ultimately did not have a large impact on the results, we added the following sentence to the Methods section “Tissue-pooled splice motif discovery”, starting on Line 764: “We tried varying the percentile threshold (from 1 to 100) and the results were robust to this parameter (the same motif clusters were identified with roughly the same number of supporting seqlets).”

9. “Finally we selected one negative SNP for each fine-mapped causal paQTL by requiring that they have identical distances to an annotated PAS and that the negative SNP occurs in a gene with expression levels at most 1.625-fold within the expression levels of the paQTL gene” (698-700). Where does 1.625-fold come from?

We thank the reviewer for pointing our attention to this particular step in the procedure for curating negative SNPs. We agree that the level of detail (and a better rationale) could be given. Also, after inspecting the code again, we note that the number is actually less than 1.625.

The following procedure is used to efficiently find candidate genes with matched expression levels to a given gene in a particular tissue and then filter this subset for genes that contain distance-matched non-causal SNPs:

1. Discretize and bin the log2-TPM values of all genes into buckets of size 0.4 (in log2-space).

2. For query gene g (and its associated log2-TPM value), take all candidate genes that hash into the same bucket. Scan this subset of genes for any that contain a distance-matched SNP.

3. If none of the genes in the bucket was a suitable candidate (because none of them had a non-causal SNP that occurred at the target distance), then subtract 0.15 from the query log2-TPM value and take all candidate genes that hash into the new bucket (if subtracting 0.15 did not change the bucket, skip to step 4). Scan this new bucket for suitable genes.

4. If no suitable gene has been found, repeat step 3 but instead add 0.15 rather than subtract 0.15 to the original log2-TPM value. Scan this (potentially) new bucket for suitable genes.

5. If no suitable gene has been found, exit with an error (unmatchable).

This procedure makes searching for expression- AND distance-matched negative SNPs relatively fast, which otherwise would run slow with a naive scan on the entire GTEx VCF.

The value of the bucket size, 0.4, was chosen to be narrow (preferably restricting genes of the same bucket to be within 1.5-fold of each other's expression levels) while retaining at least 100 genes per bucket. The search offset, 0.15, was (somewhat arbitrarily) chosen to be close to, but less than, half the bucket size, so that query genes often have two matching buckets.

With these parameters, the maximum log2-fold change that two genes can be within and still match is 0.4 + 0.15 = 0.55 (i.e. 1.4641-fold). We incorrectly concluded that this value was 0.4 + 2 * 0.15 = 0.70 (1.6245) when we previously analyzed this procedure.

We have updated the details of this procedure in the Methods section titled "Fine-mapped paQTL classification task", starting on Line 871.

10. "For each type of assay (e.g. DNase), we exhaustively search for the window size W that maximizes the spearman correlation between the resulting scalar predictions and the TRIP measurements. Note that when we perform 20-fold cross-validation, we only use the training split of the current fold of the data to search for the optimal window size" (lines 805-807). Do we understand correctly that not only for every promoter tested in trip-seq, but potentially even for each cross-validation fold of every promoter, a different window size was used? If so, the resulting metric, while not formally wrong, is difficult to interpret and not useful for practitioners, as this finetuning cannot be done for promoters which have not been experimentally measured. The "natural" window size to measure DNase at a promoter is the minimal window that covers this promoter. Why not use this? Also, why do the authors not optimize the window-size for

other analyses, e.g. the enhancer prioritization analysis using Enformer (see (a)). Why is this suddenly a concern here but nowhere else?
Overall, we urge the authors to either give reasons for these choices or to benchmark different approaches.

We agree with the reviewer that the search for optimal window sizes through cross-validation, which indeed results in a unique window size per promoter and fold (and even output assay type), are hard to interpret and inconsistent with the much simpler approaches taken for other benchmark tasks. The main reason why we did this is because there are a relatively large number of different output assay types that we use (CAGE, RNA-seq, DNase, H3K4me3, H3K9ac, H3K36me3, H3K79me2, H3K9me3, H3K27me3), and at that time we thought the most unbiased approach (and the approach requiring least thought, admittedly) in setting these window sizes would be through cross-validation.

In the revised manuscript, we have re-done and updated this benchmark in line with the reviewer's suggestions. Specifically, we now only use two different window sizes (globally):

- For CAGE and RNA-seq, we use a 4kb window, which tightly covers the inserted reporter (including promoter). We saw that the performance from the RNA-seq tracks would actually increase if we used a more narrow window size (and this was consistent across all promoter types), e.g. 128bp, but we refrained from this since 4kb is a more intuitive choice. Overall the performance drop was less than 0.02 Spearman R.

- For the DNase- and histone modification tracks, we used a wider window of 8kb. We noticed that the performance (Spearman R) increased for nearly all of these tracks monotonically with wider window sizes (for the majority of promoter types), up to a mean optimal window size of ~10.5kb - 11.5kb across promoters, suggesting that it is beneficial to also aggregate flanking regulatory signal present in the native context. Again, the overall difference in performance between a narrow and wide window is marginal (e.g. only about +0.02 difference in Spearman R for ARHGEF98 when using a 8kb window instead of a 4kb window).

We have added supporting data and plots to Supplementary Figure S4K-M, and have updated Main Figure 4E with the new results.

We have updated Methods section "Predicting TRIP expression" with the new rationale for window sizes, starting on Line 984:

"For CAGE and RNA-seq outputs, we used a 4,096 bp window size which tightly covered the full reporter construct (and tightly covered the average signal profile, as exemplified in Supp Figure S4K for promoter ARHGEF98). While this was technically a sub-optimal choice (a narrow 128 bp window maximized the average Spearman R across promoters for RNA-seq; see Supp Figure S4L), the difference in Spearman R was small (e.g. <0.02 for ARHGEF98) and a 4,096 bp window size was a more intuitive choice. Similarly, the average optimal CAGE window size

was 8,813 bp, but the 4,096 bp window had near-identical performance (<0.01 difference in average Spearman R across promoter types). For DNase and histone ChIP tracks, we used a slightly wider 8,192 bp window size as we noticed that the correlation to measured expression saturated less quickly than CAGE as a function of window size (e.g. ~0.02 difference in average Spearman R across promoter types when comparing a window size of 4,096 bp to 8,192 bp for H3K4me3)."

11. A potential pitfall could be that the model predicts better exon boundaries for mRNAs, which are enriched for coding exons with characteristic sequence distributions, than for non-coding RNAs (See also the SpliceAI paper). The author should comparatively assess the performance on non-coding RNAs.

As suggested by the reviewer, we have updated the splice site detection benchmark and broken the results down by protein-coding and non-coding RNA. The updated benchmark can be found in Supp Figure S2M. The results show that accuracy indeed drops for non-coding RNA compared to coding RNA. However, the drop in performance is only moderate (AUPRC stays above 0.95 for more than half of the tested tissues).

12. In the cited benchmark of Enformer (figure 5A-B of citation [56]), it was shown that Enformer can predict very well the strength of promoters measured in a STARR-seq reporter assay, but struggles to predict the enhancer effect. It was speculated that this is because the enhancer in STARR-Seq is transcribed, leading to post-transcriptional effects that Enformer does not account for, as it was trained primarily on CAGE. It may be interesting to rerun this analysis to see if Borzoi performs notably better, as it could theoretically account for such biases if they are present, although we would understand if the authors believe that the computational cost (c. 600k predictions) is not justified.

While we agree with the reviewer that this sounds like an interesting analysis to perform, we do feel that it is out of scope for this study. As we currently do not find any canonical 3' UTR motifs related to mRNA half-life in the salient features learned by Borzoi, and since it is unclear whether canonical enhancers present in the 3' UTR would be out-of-distribution anyway, coupled with the fact that mini-gene reporters are very different from native transcript structures found in the genome, it is difficult to justify this computational experiment and what might be learned from it. As suggested by the reviewer in Major Point 2, we did compare Borzoi to Enformer on the saturation mutagenesis MPRA data of Kircher et al., which showed comparable performance on enhancers. While this reporter data is different from STARR-seq, it does allude to Borzoi having learned roughly the same local sequence grammar of enhancer sequences.

Minor points

13. The authors write that "Nevertheless, the predicted exon-to-intron coverage ratio surrounding a potential splice site discriminates between annotated sites and negatives with accuracies on par with the state of the art model Pangolin (Supp Figure S6B)" (lines 286-288). Either the figure S6B is wrong or the sentence is wrong, as the figure strongly indicates that the

model is not on par with Pangolin, showing significantly worse performance for both kinds of splice donors analyzed.

We agree and have re-written this paragraph to emphasize exactly for which type of splice junction performance is matched between Borzoi and Pangolin (acceptors), and explicitly state that performance is worse than Pangolin for splice donors. See the updated sentence starting on Line 178: “Nevertheless, the predicted exon-to-intron coverage ratio surrounding a potential splice site discriminates between annotated splice acceptors and negatives with accuracies on par with the state-of-the-art model Pangolin [16], while performance is slightly lower for canonical splice donors (GT; mean dAUPRC = -0.01) and worse for non-canonical donors (GC; mean dAUPRC = -0.04) (Supp Figure S2L-M).”

14. The authors note that Borzoi outperforms Pangolin on variants close to the splice site but underperforms on far-away splice variants (lines 305-307). The authors note that “Most of these far-away SNPs are de novo splice-gain mutations”, but it is not clear to us what this implies. Intuition would suggest that, if anything, Borzoi should perform better at long-range splicing (due to the attention mechanism and its large receptive field) and worse at close-by events (due to the binning at 32 bp resolution). The fact that this relationship is reversed may warrant further comment or investigation.

We believe there are two different effects at play that make the results appear the way they do. First, Borzoi’s predicted effect sizes for *de novo* splice-gain mutations in intronic regions are quite low and not as well-separated from the background as splice-enhancing or repressing variants near canonical junctions. We hypothesize that we could improve this with a different score statistic (absolute max-difference in bin-level coverage, which is currently used, is likely not optimal for splice-gains, as we would rather want to integrate the total change in coverage over the intronic region between the splice-gain and nearest junction). However, improving this benchmark further is out of scope for this paper and we aim to pursue this in future research.

Second, and in contrast to Borzoi, Pangolin can seemingly very easily detect splice-gain mutations simply by looking at the local predicted score right at the variant position (i.e. even though the SNP occurs far away from an annotated splice junction, Pangolin makes use of a local predicted score right at the variant allele for splice-gains). As it seems, these predictions are marginally more concordant with the sQTL labels than predictions of SNPs closer to exons.

We have updated the relevant paragraph to reflect this hypothesis, starting on Line 390: “Most far-away SNPs are de novo splice-gain mutations and are relatively easy for Pangolin to classify based on the local predicted effect at the variant allele, while Borzoi's splice-gain predictions in intronic regions appear less well-calibrated. In contrast, Borzoi is better at distances closer to the junction (Figure 7D-E; dAUPRC = 0.02 when only considering the subset of variants that occur at distances <= 200 bp).”

15. The introduction states that “From a single RNA-seq experiment, Borzoi derives the primary cell type/state-specific TF motifs and a genome-wide map of nucleotide influence on gene

structure and expression" (line 55). It is unclear to us what the authors are trying to say here, but as written, it does not correspond to the manuscript. The model is never trained on a "single RNA-seq experiment", it is trained on hundreds of RNA-seq and other experiments simultaneously in a multi-task fashion, and can therefore pool information across experiments. It is never investigated what the model learns if it is trained on only a single experiment. Accordingly, this sentence should be clarified or removed.

What we meant was that we can apply interpretation methods (e.g. input-gated gradients) to each of the trained-on RNA-seq coverage tracks, as predicted by Borzoi in a multi-task fashion, in order to derive sequence features that drive expression specific to this particular RNA-seq experiment (e.g. to identify tissue-specific regulation as is done in Figure 3).

We have reworded the sentence to clarify this (Line 57 in the revised manuscript): "By applying attribution methods to predicted coverage patterns of individual RNA-seq experiments present in the training data, Borzoi derives the primary cell type/state-specific TF motifs and a genome-wide map of nucleotide influence on gene structure and expression."

16. The introduction cites a previous work by the last author which demonstrates a slight advantage to multi-species training (lines 46-47). But this was done only using a particular model design which did not use the attention mechanism. Since models have become significantly larger and more powerful in the meantime, it is not entirely clear to what extent this insight still applies. It may not be within the scope of this paper to investigate this, but in that case this should be clarified. As it stands, it is not shown that Borzoi benefits from the mouse data, nor is it even really investigated whether the mouse predictions are particularly meaningful, since almost all the benchmarks presented are in human.

We performed ablation studies where we re-trained (slightly smaller, but arguably still comparable) Borzoi instances and removed specific data modalities or other features, including the mouse data. As demonstrated on GTEx gene-level test predictions or eQTL coefficient concordance (Supp Figure S5D), performance improves consistently when including mouse data during training. This benefit is largest when only training on RNA-seq data, with a smaller boost in performance when also including human epigenetic data such as DNase/ATAC-seq. We also saw benefits, albeit smaller (and not entirely consistent ones), on the CRISPR benchmark, with increased AUPRCs for multi-species models on both FlowFISH and Gasperini data when training on DNase-, ATAC- and RNA-seq but seemingly no apparent benefit when also including human ChIP-seq in the training data (Supp Figure S4J).

17. The authors make a number of highly un-intuitive choices when preprocessing the data, such as exponentiating the bin values by 3/4 and taking a square root if the resulting value is above 384, choosing which GTEx samples to include by k-means clustering metatissues with k = 3, and processing GTEx data in an unstranded fashion. Firstly, for reasons of readability, these transformations should be presented as formulas or pseudocode, rather than as text, with all hyperparameters clearly stated. Secondly, the authors should either clearly explain why these

choices of transformations, preprocessing steps and cutoffs are important, or note that they were arbitrary choices.

We apply a squashing transformation to the RNA-seq tracks due to the rather exceptional dynamic range in coverage observed across genes for some experiments, and we do not want a few extremely highly expressed genes to contribute many orders of magnitude more to the Poisson/MNLL-loss than other genes. We arbitrarily chose the exact values of these parameters (e.g. 384), but our general aim was to limit the maximum bin values to numbers around 1,000 (a reasonably large number for maintaining a rich range of value, while still working with numbers small enough that tensorflow can handle them without special data types).

We have added clarifying sentences to reflect this, including a formula of the transforms applied to the coverage track(s), in the Methods section starting on Line 496:

“These operations effectively limit the contribution that very highly expressed genes can impose on the model training loss. The below formula summarizes the transform applied to the *j*:th bin for tissue *t* of target tensor **y**:

<Formula>

We refer to this set of transformations as 'Squashed scale' in the main text. The parameters were chosen such that most genes had bin values less than 1,000 (a reasonably large maximum value that is handled well by standard tensorflow data types).”

The GTEx project has hundreds of “replicate” experiments performed on many individuals for each tissue. We believed that training on a few representative examples would be sufficient for this analysis. Rather than choosing these examples randomly, we performed k-means as the simplest clustering algorithm that we could think of. Unfortunately, the GTEx datasets did not use a stranded RNA-seq protocol.

18. Relatedly, while we appreciate that the training data is available at all, it can only be downloaded against a substantial payment. For reproducibility and a wide accessibility, the authors should provide their preprocessing scripts that generate the training data from publicly available resources, particularly in light of the many complex steps described above.

We have uploaded the training data processing scripts and QTL generation scripts, along with documentation, here: https://github.com/calico/borzoi/tree/main/src/scripts/data. Additionally, we have made available the scripts and python notebooks relevant for replicating the evaluations (and results) presented in the paper, located at: https://github.com/calico/borzoi-paper. This has been updated in the “Availability of data and software” statement.

Ps., Downloading the full TFRecord dataset would require a negligible payment for users relative to other research costs in our opinion. We previously made such datasets freely

available, but many users set up training pipelines in which they were constantly accessing the data, which caused us to incur a substantial payment.

19. Many figures showing quantitative information, e.g. 5A,E,F, 6A,C,F, do not display the scale of the y-axis. Including scales will make these plots more interpretable and will allow the reader to judge to what extent predictions and observations are quantitatively comparable.

We have updated all the main and supplementary figures to include y-axes (both for coverage prediction plots and attribution sequence logos).

20. Figure 3C-D: the Enformer paper helpfully noted the number of enhancers in each distance bin, as well as the number of positive examples in each bin. As the AUPRC is quite sensitive to the class balance and since these metrics generally become more shaky with smaller sample sizes, it would make the figure more interpretable if this information was included. Additionally error bars should be indicated, or some other information should be provided with which the reader can gauge whether the performance improvements vis-a-vis previous methods are statistically significant.

We have now included the number of positive- and total examples in each distance bin, and the bar plots have been updated to show the 95% confidence interval as determined from 1000-fold bootstrapping (the updated figures are listed as Figure 4C-D in the revised manuscript).

21. The paper should state the hyperparameters used during training of the model

While we understand the reviewer's point of view in explicitly documenting all of the parameters and details of training the model, we feel that this would quickly make the manuscript unwieldy and hard for the general audience of Nature Genetics to read. Instead, all of the parameters used during training are available on this link (which is linked from our Github): https://storage.googleapis.com/seqnn-share/borzoi/params.json

22. In the gradient based attribution methods, the authors always subtract the mean across nucleotides. Why is this done? If it is because of the analysis presented in the paper "Correcting gradient-based interpretations of deep neural networks for genomics", then this should be cited.

We thank the reviewer for pointing this out. We have added the reference.

23. The conventional abbreviation for the spearman correlation is the greek letter rho (ρ) and not R, which generally refers to the pearson correlation.

While we agree with the reviewer, both us and other peers have seen the greek letter rho used for spearman correlation as well. To remove any possibility of confusion, we now explicitly annotate correlation metrics with "Pearson R" or "Spearman R" in every figure panel.

24. The plots showing RNA-seq coverage of Ref and Alt (ex.6C,F) could benefit from including a version with a measure of the difference between them in the y axis, such as the ratio between Ref and Alt coverage. It is sometimes hard to see the changes in coverage claimed by the authors.

We agree with the reviewer, though we found that it was hard to fit additional "difference" coverage plots into the figure(s) without making them appear overly complicated. As a compromise, we have added annotations to all the coverage prediction/measurement plots depicting variant effects that explicitly states the estimated variant effects size (e.g. log fold change in exon-to-intron coverage ratio) derived from the reference- and alternate tracks.

25. In the example depicted on 6C the coverage prediction by Borzoi has substantial differences from the ground truth. It generally predicts much more coverage for the intronic regions, the second to last exon and the longest version of the last exon. Such differences are even higher than what can be observed due to the effect of a SNP. This should be commented.

While it is true that the predicted coverage track in (old) Figure 6C (using either the reference or alternate allele) has substantial differences to the measured tracks, it is worth noting two things:

- The measured coverage displayed in this figure, or in any variant example, are not necessarily the same GTEx coverage measurements that Borzoi was trained on (Borzoi was trained on 2 or 3 GTEx subjects' RNA-seq data per tissue, while what we are showing here for measurements are samples of individuals selected because they have either the reference or alternate allele), which means that the model may have learned subtle differences (or biases) in coverage patterns for some genes that are not present in other subjects' RNA-seq data. That said, even if these were matching predictions and measurements from the same RNA-seq experiment, one must remember that this is only one example coverage prediction for a 4kb window of a single gene. While Borzoi is far from perfectly recapitulating the exact coverage shapes of held-out genes (which we already state in the Discussion), we do feel that we have presented substantial (unbiased) genome-wide benchmarks on held-out test data in Figure 1-2 to convince readers that the predictions overall are well-correlated with the measurements.
- We want to emphasize that, even though there may be differences in the coverage shape compared to the measurements, only the component of the shape relevant to the variant of interest is actually altered in the prediction, such that once we subtract the alternate prediction from the reference predicted coverage, the residual shape matches the measured residual quite well.

In light of this, we leave this comment without any change.

26. For both panel 6C and 6F it is not mentioned why the coverage examples are in Testis and Nerve.

We focused on a tissue-agnostic paQTL analysis (and benchmark) in Figure 5 (which is now Figure 6 in the revised manuscript), which is why we also want to show examples with pooled

predictions. In contrast, the larger set of sQTLs allowed for a tissue-specific analysis of variants (In old Figure 6, which is Figure 7 in the revised manuscript), and this is the reason we want to also show example predictions that are from individual tissues. We show the predictions and measurements in these tissues because it is there we find a significant effect of the sQTL(s).

To clarify this, we added the following sentence to the paQTL section, starting on Line 356: “We focused on tissue-pooled predictions since the limited number of QTLs prohibited a tissue-specific analysis.”. We also added the following sentence to the sQTL section, starting on Line 382: “This relatively large set of variants allowed for a tissue-specific analysis (there are approximately 4x more sQTLs than paQTLs).”. Finally, we updated the caption for Figure 7C, stating that the sQTL is significant in Testis.

27. In panel 6F there is no scale labels on the attribution scores which makes it hard to understand the score differences between reference and alternate alleles

Y-axis scales have been added to all attribution scores in all figures. Also, whenever we display sequence logos for reference- and alternate alleles, the logos are equally scaled (which we now clarify in the figure captions).

28. In 4E the legend refers to a precision recall curve when it is a ROC curve

We thank the reviewer for pointing this out. We have corrected the figure caption.

29. In line 624, the upper bound of the summation over the layers should be 7.

We thank the reviewer for finding this mistake. We have updated the bound to 7.

30. In the caption for Figure 1, the upsampling operation followed by a convolution is referred to as a deconvolutional layer. This term, though often used, is technically confusing and should be avoided. See:
https://medium.com/@marsxiang/convolutions-transposed-and-deconvolution-6430c358a5b6

We have revised this sentence in the caption for Figure 1, which now reads: “The output is then repeatedly upsampled and put through a number of convolution layers with matched U-net connections to predict at 32 bp resolution.”.

31. In Supp Figure S8, the hidden dimension (C) of the Transformer block should be 1536.

We thank the reviewer for finding this error. We have corrected the hidden dimension to 1536.

Reviewer #2:

Remarks to the Author:

This study presents a novel deep learning model, named Borzoi, which can learn to predict cell- and tissue-specific RNA-seq coverage from DNA sequence. Importantly, the trained Borzoi models are effective for understanding different layers of gene regulation, including transcription, splicing, and polyadenylation via model interpretation and variant effect scoring. It is also shown that Borzoi is competitive with, and often outperforms, state-of-the-art models trained on task-specific regulatory data. Model performance is assessed using highly appropriate measures, and caveats in performance (and hence areas for improvement, which are useful for the field to know about) are honestly presented throughout the study.

There are multiple core improvements for this Borzoi model over other models:
1. It extends previous work in the field for modeling sequencing read count features by moving from local, small windows to much-expended regions capable of incorporating enhancers and RNA elements located at distal regions of a given gene.
2. On the technical side, the model has an improved receptive field size as well as an improved transformer/attention layer context window size. These are very important for modeling the long-range dependencies. Further, its coupling with the U-net structure provides additional support for using distal sequences as well as better resolved output.
3. The showcases presented in this study combine the visualization of attention maps alongside other classic genomic data model-interpretating methods. As demonstrated, this is particularly useful for learning the mechanisms underlying multi-layered gene regulation.
4. The paper demonstrates how a deep learning model trained with RNA-seq coverage data can be used for several important, biologically-motivated, downstream tasks.

Overall, this is a very well-done study with few flaws. That said, there are several opportunities for further improvement. I have grouped them into “Moderate issues”, which I would like to see addressed, and “Minor issues”, which largely comprise suggestions for improving performance or interpretation, but are not viewed as necessary for publishing this particular study - they are possible opportunities for the future that could be either discussed or quickly tested here if feasible.

We thank the reviewer for their overall positive feedback. We have addressed each of the reviewer’s comments and suggestions below.

Moderate Issues

1. Understanding the driving forces behind the performance improvement. While the Borzoi model did present better utilization of distal regulatory sequence features, e.g., in Fig. 3 A-D, it’s unclear where these improvements came from. Is it because of the U-net structure (which provides support for marginally distant sequences)? Can an ablation analysis be performed to show whether this is or is not the case? On top of this, it’s hard to understand Borzoi’s

extraordinary performance in Fig. 3C, where the trend of decreasing AUPR w.r.t distance is completely reversed for only the Borzoi model only within the 130k-262k bin.

We agree that an ablation study where various data modalities and/or architecture components are removed would aid readers in understanding what's behind the performance improvement for this benchmark. We have added a new ablation analysis in Supp Figure S4I-J that does precisely this. In Supp Figure S4I, we show that part of Borzoi's improvement on this task comes from the ensemble (though the model outperforms Enformer using a single replicate). In Supp Figure S4J, we re-train (somewhat smaller instances of) Borzoi on subsets of training data, or on the full data but with the U-net removed (and modeling coverage at 128 bp resolution). We found that on this particular benchmark, the increased resolution does not change performance much. (In contrast, however, the U-net does aid interpretability within the transcript structure when exon/intron boundaries are very short, see an example of this in Supp Figure S7G.) We further found that the addition of mouse data, as well as the addition of auxiliary data modalities such as DNase-, ATAC-seq and ChIP-seq in the training set, to various extents are crucial for utilization of distal features (see the analysis below for more details). Please note: When comparing these different model instances, we always use the K562 RNA-seq output tracks to compute predictions and gradients, meaning that any observed improvement when including e.g. chromatin accessibility data is due to indirect learned interactions between the data modalities during training (i.e. the model makes better use of enhancers to predict RNA-seq).

We have added the following paragraph in the revised manuscript, starting on Line 260: "To understand what drives the improved performance in prioritizing enhancers with Borzoi, we performed a number of ablation studies. First, we tested how various parameter choices affected performance when aggregating the gradient- or shuffle-based scores, for both Enformer and Borzoi (Supp Figure S4I). Briefly, we found that a fraction of Borzoi's improved accuracy was due to model ensembling, and that gaussian smoothing often was superior to averaging the gradient score in a discrete window, for both models. Next, we re-trained additional instances of Borzoi where various subsets of the original training data were removed, or where the architecture was changed (Supp Figure S4J). Consistently, we observed diminished performance if only the RNA-seq data was used for training. We generally observed improved performance on the data from Gasperini et al. when training on mouse in addition to human assays, although human-only and K562-only models were performant on the data from Nasser et al. as well. Including DNase-/ATAC-seq training data improved performance in both the human-only and multi-species settings. However, including ChIP-seq only boosted performance of the human- or K562-specific models. Finally, we note that removing the U-net component and modeling coverage at 128 bp resolution does not degrade performance on this benchmark."

Furthermore, the performance for the 130k-262k bin in Figure 3C of the original analysis appeared extraordinary (and in reverse direction of the general trend of the other bins) solely due to stochasticity (the FlowFISH data has only a handful of positive gene-enhancer examples at this distance). In response to a comment we got from another peer we actually changed the distance categories to different bin ranges (previously, there was a mismatch in the number of

examples scored by Borzoi and Enformer in the 60-130K bin since Enformer only sees variants up to 98kb; in the updated manuscript we have changed the bin ranges to start a new, Borzoi-exclusive bin exactly at 98kb - 262kb), and now the results look less extraordinary. Finally, in the updated Figure 4C-D (previously Figure 3C-D), we now include 95% confidence intervals that clearly shows the last FlowFISH distance category has a lot of uncertainty due to the small sample size, which is not the case for the larger Gasperini data.

2. Training data. Were all data types used by Borzoi necessary for performance? For example, would Borzoi have performed equally well with just chromatin accessibility and RNA-seq? This is of great interest to the field because ATAC-seq and RNA-seq are routine and feasible for most labs as both bulk and single-cell assays. Is it possible to present the weight (relative importance) of each data type or dataset?

In the response the reviewer's previous comment, we presented an ablation study where we re-trained Borzoi on various data modalities and showed that for the CRISPR benchmark task, chromatin accessibility data was crucial to include in the training data for utilization of distal regulatory features. Furthermore, in a new analysis presented in Supp Figure S5D, we compare the same model instances on (1) GTEx gene-level test set predictions, and (2) eQTL coefficient predictions, and again find that multi-species training and the inclusion of chromatin accessibility data substantially improves performance. On this benchmark, however, including ChIP-seq and CAGE training data does not improve performance.

Note that Borzoi is trained on relatively few ATAC-seq samples, and these were mainly from brain tissue, which is why we did not separate DNase and ATAC data in the ablation study.

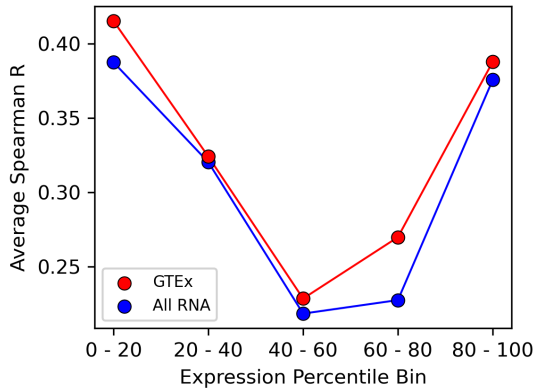
We have added the following sentence to briefly summarize this new analysis, starting on Line 305: "Comparing Borzoi model instances that had been re-trained on different subsets of data modalities indicates that both gene-level test predictions and eQTL coefficient predictions improve substantially when including DNase-/ATAC-seq data in addition to RNA-seq, or when training on mouse RNA-seq in addition to human RNA-seq (Supp Figure S5D)."

3. Use cases. How does Borzoi gene expression prediction compare for highly versus moderately versus lowly expressed transcripts? I'd like to see Spearman correlation coefficients broken down in this way. Similarly, how does Borzoi gene expression prediction break down based on number of exons or gene length? If there are indeed differences, please provide some interpretation.

We agree that it would be useful and interesting for readers to see a breakdown of gene expression prediction performance by various axes.

However, we want to note the difficulty in taking an analysis comparing predicted to measured expression across genes and breaking it down by measured expression in a meaningful way, especially an analysis based on Spearman correlation. In particular, depending on how granular we bin genes by measured expression, the Spearman correlation coefficient between predicted

and measured expression within each bin will change drastically (in general, the smaller we make each bin, the lower the correlation within each bin, as the variance in measurements will be driven more and more by measurement noise). While we include a figure in the response below depicting this breakdown (**Response Figure B**), we do not think this analysis is easy to interpret and refrain from including it in the revised manuscript.



Response Figure B: *Mean Spearman R when comparing predicted vs measured exon-aggregated coverage values of held-out test genes (average prediction of 4 replicates), as a function of measured expression quantile (5 bins).*

As a compromise, we have included a similar breakdown of held-out gene-level prediction performance in Supp Figure S1H, which depicts the average absolute prediction error (in log2-space) between predicted and measured exon-aggregated coverage values, as a function of measured expression. We note that the error increases the most in the lowest measured expression bin, though we believe this is largely driven by noise in the sequencing depth of lowly expressed genes. To highlight this, we include scatter plots of predicted vs measured gene-level coverage for two example assays in Supp Figure S1E. It is clear from these plots that low read depth in the lower left quadrants cause horizontal “streaks”, which are obviously measurement noise.

As requested by the reviewer, we have also included a breakdown of performance based on either spliced transcript length or the number of exons in Supp Figure S1F-G. While performance degrades with more exons, the performance is still reasonably high for even very long genes (average Spearman R >= 0.77).

Minor Issues

4. Suggestion 1 for possible technical improvement. The Borzoi model and its applications extended the frontiers of modeling and understanding multi-layered gene regulation in our field. This improvement is largely based on building its core modeling capability off of its predecessor, the Enformer model. Both models use transformer-based architectures with large receptive fields, plus convolutional kernels as sequence encoders to accommodate large genomic region of interest. Both also stick with 128bp resolved bins in their core transformer/attention layers as a compromise for balancing between feasible computational complexity/cost and better

modeling capacity. In this study, the context window of the transformer/attention layers has been extended to around 4k from 1.5k in the Enformer model, enabling much needed attending to gene positions spanned in large region. I am a little concerned that the important internal attention solution is left at 128 bp resolution, thus making it harder for precisely modeling certain interactions in the genome. This was also raised by the authors in the Model section on page 14, claiming that “Conversely, mammalian exons regularly cover fewer than 128bp, and modeling the coverage patterns around these exons at such a coarse resolution can obstruct splice site learning.”

There are several tools and tricks that can be used to help extend the context window beyond 4k (and therefore better resolved information for attending) - these include the use of flash attention and multi-query attention. And compared to the approximation-based approach “Scatterbrain” mentioned in the Discussion section, these directly optimized and alternative implementations might offer better performance. It would be fantastic if the authors could train a better resolved model with a larger or better resolved context window in their attention layers to show if such improvement would indeed be needed to help generate better insights and/or emerging knowledge and understanding for any of the downstream analysis.

We thank the reviewer for their ideas and suggestions of model improvement, and we fully agree that it would be very exciting to extend the effective size of the self-attention beyond 4k, either so that we can perform attention at a finer resolution (e.g. 32 or 64 bp) while keeping the total receptive field fixed at 524kb, or so that we may widen the total input window beyond 1 Megabase (or a mix of these). However, pursuing this will be a substantial undertaking, one that we feel is out of scope for this study and is better suited for a follow-up paper (either by us or by peers in the field, who might be inspired by the results of our modeling at 524kb in this paper). For example, flash attention is unavailable in Tensorflow, so trying it will require porting our whole codebase to PyTorch. Furthermore, it does not natively allow for our relative positional encoding scheme, so we would need to evaluate alternatives. We hope the reviewer understands and agrees with this.

5. Suggestion 2 for possible technical improvement. Although deep neural network models can and do learn complex relationships from data, there might be a trade-off between (1) multi-task learning of all cell types and data types together (e.g., where the model must learn that HepG2 H34me3 is valuable for HepG2 RNA-seq prediction and less so to K562 RNA-seq, etc.) versus (2) single-task learning of RNA-seq using only relevant data types for each cell type. With the addition of CLIP-seq, could a single-task learning model (e.g., in K562, HepG2, or some other cell type with abundant CLIP-seq) outperform the multi-task? Relatedly, are “singleton data tracks” (RNA-seq from a cell type with no other epigenome data, etc.) helpful to model performance?

The supplementary analyses in Supp Figure S4I-J (ablation study benchmarked on the CRISPR data) and Supp Figure S5D (ablation study benchmarked on the eQTL data) address most of the questions raised in this comment. See our responses to reviewer comments 1 & 2 (under “Moderate Issues”) above for details. The analysis in Supp Figure S4J includes a set of 3 model

ablations trained on data for a single cell type (K562), including either RNA-seq, DNase-/ATAC- and RNA-seq, or DNase-/ATAC-/CAGE-/ChIP- and RNA-seq assays respectively. Overall, the K562 models perform well on both the FlowFISH and Gasperini data, but are generally worse than the model versions that include all human data, or for that matter human and mouse data. There are a few exceptions, e.g. one of the K562 models outperforms the other models on the two most distance bins furthest away from the TSS for the FlowFISH data, but we suspect this may be due to the low sample size, and this trend generally does not hold on the Gasperini data. Importantly, the K562 models indicate that additional epigenetic training data beyond RNA-seq greatly boosts performance.

Regarding CLIP-seq: We are eager to make use of CLIP-seq data and include it in the training set, but we hope the reviewer understands the large amount of work that would go into gathering, uniformly processing, and analyzing this modality, and we leave it for future work.

6. Additional application/discussion topic. It would be nice for the authors to discuss their views of the opportunity for applying this kind of approach to multiome data (scATAC-seq+scRNA-seq), which in principle would enable the direct tying of enhancers to function. It would be even better to see how the model worked using transfer learning with application to scRNA-seq + scATAC-seq data.

We have added a sentence to the Discussion regarding the possibility of training on multiome data, starting on Line 433: “This observation suggests that recent multiome datasets, which measure both accessibility and expression in single cells, would be quite valuable as joint training data.”.

Regarding transfer learning approaches: We agree that it would be interesting to develop and evaluate transfer learning approaches for leveraging Borzoi’s large pre-training data while fine-tuning for a newly collected (possibly multi-omic) experiment. There are many options for how one could implement this and many baselines to compare to. Thus, we view this as out of scope for this study and leave it for future dedicated work.

7. Missing details. It would be helpful to include # of positive and negative examples for the prediction tasks in Fig. 3C, D in main text or figure legend, in addition to Methods.

Per the reviewer’s suggestion, we have added these numbers below each distance category in Figure 4C-D of the revised manuscript (this was Figure 3C-D in the original manuscript).

8. Model interpretation. In Fig. 3A, 3B, true positive Borzoi examples are interpreted. It would be interesting to know what features contributed to false positives. (Do the four model folds attend different sequence regions or disagree in some other way…? Are they attending signal in the wrong cell type? Is there any intuition to be gained?) What lessons can be learned from these mistakes?

From the new supplementary analysis provided in Supp Figure S4I, it can be concluded that the use of the four-replicate ensemble boosts performance quite considerably on both the FlowFISH and Gasperini data. By inspecting the per-replicate gradient patterns plotted for the two example genes in Figure 4A-B in the revised manuscript (formerly Figure 3A-B), which shows both true positive and false positive enhancer-gene links, it becomes apparent that the replicates generally don't attend to wildly different loci. Instead, there exists variance in the exact gradient magnitudes, which averaging across the replicates usefully smooths over to increase performance. Similarly, in the new supplementary saturation mutagenesis MPRA benchmark provided in Supp Figure S5G-J, the use of ensemble predictions clearly reduces variance. When inspecting an example loci (Supp Figure S5J), it can be seen that replicate 0 of the model fails to utilize some specific motifs properly, while the full ensemble's motif use agrees well with the measurements.

Additionally, in response to other reviewers' comments, we compared Borzoi and Enformer on the CRISPR benchmark using a 2kb dinucleotide-shuffle approach instead of input-gated gradients (also shown in Supp Figure S4I). Intrigued as to why Borzoi's performance diminished quite dramatically at short TSS distances with the shuffle approach, we visualized a few false positive predictions per the reviewer's suggestion in Supp Figure S4D, and found a number of examples (annotated as negatives in the measured data) where the candidate enhancer was located so close to one of the target gene's exons that the 2kb window shuffle destroyed the exon junctions. Interestingly, this drop in performance is not visible in the benchmark in Supp Figure S4I for Enformer at short distances, suggesting that Borzoi is much more sensitive to perturbations affecting the gene structure (which we already knew, or guessed at). We added the following sentence to the revised manuscript to highlight this observation (Line 251): "Manual inspection of false positive predictions suggest that loci near exons of the target gene, or other genes, may be a source of discrepancy between the model and measurements (Supp Figure S4D).".

As to whether the model attends to enhancers related to the incorrect cell type (not K562) as a failure mode in the CRISPR benchmark: Since the strongest peaks in the gradients of the two visualized examples in Figure 4A-B highlight TAL- and GATA motifs, it is unlikely that the model is attending to incorrect (non-K562) enhancers. Anecdotally, we have never detected false positive promiscuity of enhancer usage across tissues. Assessing this carefully at a large scale seems quite complicated and difficult. To alleviate the reviewer's concern, we refer to our analysis of cell- and tissue-specific motifs found in gradient saliencies in Figure 3 (Figure 2 in the original manuscript), from which we concluded that they agreed well with known biology.

Reviewer #3:

Remarks to the Author:

SUMMARY OF KEY RESULTS

A very technically dense manuscript. While I hold expertise in disease-relevant RNA splicing and ML, I struggled greatly with the complexity of information in this manuscript. Training and Test data sets for ML and deep learning are an embodiment of critical thinking. If explained well and clearly, any scientist can appraise the critical thinking.

Borzoi is a deep learning neural network trained on the raw DNA sequence, methylation data from different tissues, as well as adjunct genomic data from a variety of assays – that each inform inference of binding sites of transcription factors, promotor machinery and transcriptional activators, splicing regulators, polyadenylation machinery.

The authors claim Borzoi can: 1) predict RNA isoform expression in different cells and tissues based on the DNA sequence alone; 2) identify key transcription factor binding motifs with purported likelihood to influence tissue-specific isoform expression of genes, though rigorous evaluation of precision and recall using a high confidence dataset is not conducted; to 3) interpret the effects of intragenic SNVs upon RNA isoform expression, with purported application to help prioritise/identify of SNVs associated with human disorders (monogenic disorders, polygenic risk alleles etc).

Whether Borzoi is an incremental advancement or transformative was difficult to discern, due largely to limited explanation of the critical thinking behind, and the specific content of, the training and test datasets for each Figure. An algorithm is only as good as its Training and Test Data. Most Test Datasets used are big, genome-wide datasets. Assignment of True Positives versus True Negatives within the Test datasets was not defined. Further, I was unable to discern if the authors re-trained the Neural Network from scratch without any data from these genes, or whether Borzoi was fed all data other than the empirical RNA-Seq data - or otherwise how sources of predictive bias was mitigated in the Training Data for each Test Set.

We appreciate the reviewer’s feedback and opinion.

The allocation of data into training and test sets follows all work previously performed in this subfield from DeepSea and Basset to BPnet to Basenji and Enformer. Specifically, a portion of (randomly selected) genomic sequence is held out as a test set. The remainder of genomic sequence serves as the training set. For each independent sequencing experiment collected, there is an associated measurement with each genomic nucleotide via aligned read counts. The model is trained to predict all of these values simultaneously given the underlying sequence as input. All datasets are included during training and evaluation. We do not aim to generalize to new data; we aim to generalize to new sequences. In this work and many others, the labels aren’t binary—there are no true positives, true negatives, false positives, or false negatives. Every 32 bp genomic bin has a numeric value derived from the aligned read coverage in each of the independent sequencing experiments. The loss function of the model makes this a regression problem.

To make these points clear, we have added details to the Methods section “Training data”. For example, starting on Line 493 we now state that we train on normalized coverage values: “We train Borzoi to directly predict these continuous coverage values in 32 bp genomic bins.”.

We describe how the genome-wide dataset is split into training and test sets in the Methods section, starting on Line 512. In this research, relative to others, we introduced an additional layer of complexity, in that we trained four model replicates (with randomly initialized weights) on the same train/test split. These four replicates enabled us to assess performance metric variance and together form an ensemble that is more accurate than its individual members for appropriate tasks. In all analyses for which we refer to held out test sequences, we specifically refer to the regions of hg38 that these model replicates were not trained on. We reviewed the manuscript and explicitly state when we are using a single replicate model or the ensemble.

We also want to point the reviewer to the following places in the Methods section of the revised manuscript for descriptions of the curation of positive and negative examples for all of the downstream benchmark tasks (responding to the point raised by the reviewer that “Assignment of True Positives versus True Negatives within the Test datasets was not defined.”):

- Line 802 of Methods section “Fine-mapped eQTL classification and regression tasks”: Paragraph describing the definition of positive fine-mapped causal eQTL SNPs.
  Line 809: Definition of negative SNPs (TSS-distance controlled).

- Line 865 of Methods section “Fine-mapped paQTL classification task”: Definition of positive and negative paQTL SNPs.

- Line 923 of Methods section “Fine-mapped sQTL classification task”: Definition of positive and negative sQTL SNPs.

- Line 949 of Methods section “Splice site identification task”: Definition of positive and negative examples for the splice site detection problem. Note: This paragraph also describes how loci are selected to be part of both Borzoi’s and Pangolin’s test sets.

- Line 956 of Methods section “Classifying rare and common variation from gnomAD”: Definition of positive and negative examples.

- Line 998 of Methods section “Gene-enhancer prioritization task”: Paragraph describing the definition of positive or negative gene-enhancer pairs based on CRISPRi measurements.

Finally, we want to emphasize that, while we take care in splitting the original dataset which Borzoi is trained on into distinct training and test sets, we do not make any such distinction for downstream tasks where we are trying to predict the effect of sequence perturbation (e.g. e-/s-/paQTLs, enhancer perturbation, TRIP, etc.). We, and all of the previous work in this space, consider these all to be held-out test examples, regardless of whether or not Borzoi saw the

genomic fragment of the hg38 reference genome during training because the model never saw measurements for the perturbed version. Correctly predicting such data requires proper causal attribution within the model, which an overfit model would not achieve. Importantly, this is assessed using orthogonal, not trained-on, high quality datasets, such as CRISPRi measurements.

Consequently, I am unable to weight Borzoi's predictive accuracy and innovation over other tools. Some Pearson coefficients were low @ 0.53 and some higher @ 0.83. In many instances I cannot grasp exactly what metrics are being compared by Pearson correlation. Discrepancies with other DL tools was glossed over.

In Supplementary Figure 2 of Kelley et al. 2018, we showed that Pearson correlations on test set sequences vary due to the quality of the experiment; they track closely with replicate-replicate concordance of the experiment. We do not believe the correlation values meaningfully reflect the absolute utility of the model, but they can be useful for relative comparison between models. We report them here for comparisons to the Enformer model.

* Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res 28, 739 750 (2018).

REVIEWER RESPONSES TO NG EVALUATION CRITERIA

ORIGINALITY: Borzoi is based on the Enformer algorithm. While some differences are clearly articulated (e.g. training set window set at 32 nt rather than 128 nt), other points of difference/novelty are hard to discern. Figure 4 shows an incremental improvement in Borzoi (AUC 0.794) over Enformer (AUC 0.747).

The originality of this work is the use of RNA-seq read coverage data for training, not the neural network architecture itself. Enformer was not trained on RNA-seq, and any previous sequence-based model has only ever trained on RNA-seq data after aggregating counts into scalar gene TPM values. A description of previous work on modeling (aggregated) RNA-seq is given in the paragraph starting on Line 31 of the Introduction, where we also propose benefits of modeling RNA-seq coverage directly (our approach). Changes to the neural network architecture (compared to Enformer) only serve to better model RNA-seq data (such as the increased resolution, or the greatly increased input window).

SIGNIFICANCE: Demonstrating a capability of deep-learning to perceive patterns in raw DNA sequence that can RELIABLY inform tissue-specific isoform expression is highly significant to transformative. However, a lack of clarity around training and test datasets, critical thinking behind the methods, made it very difficult to weight the strength of evidence supporting conclusions made.

Please see our response above (under “SUMMARY OF KEY RESULTS”), where we clarified the train/test splits used when training Borzoi. This is also detailed in the Methods section, starting on Line 512.

DATA AND METHODOLOGY were opaque. Extremely difficult to understand and beyond the comprehension of most Nature Genetics readers. Every approach is technically complex. Nevertheless, it is possible to present a high-level explanation of the critical thinking behind the training and test datasets.

Of particular importance is lack of clarity around Test and Training datasets and what, exactly, is compared or presented in figures. Figure 1: Why EGFR? Without being 100% sure that Borzoi was not trained on any EGRF RNA-Seq data, and without being plainly informed of a significant variation in tissue-specific EGFR isoform expression that you were stress-testing Borzoi’s ability to spot, one cannot weight Figure 1B or most of Figure 1. If the algorithm was trained on any EGRF RNA-Seq data, and there is only one main EGFR isoform expressed similarly in all tissues, getting the isoform prediction right is itself predictable, right?

We have updated the caption for Figure 1B, clarifying that the genomic sequence containing INSR (the currently plotted example in the revised manuscript) is from the held-out test set. We chose to visualize this gene because it is a recognizable gene to many researchers that has a large span with many exons, making it a representative complex example. In Figure 1A-D, we have not yet begun to study tissue-specific expression, and this example was not chosen to evaluate tissue-specific activity. Figure 1E is the first panel that evaluates tissue-specific gene expression, which is done by considering all genes from the test set, that were never seen during training, and evaluating how well Borzoi’s predicted gene-level expression values correlate with measured values after subtracting the average predicted (or measured) expression of each gene.

Finally, the reviewer repeatedly emphasizes “isoform”. Alternative isoforms of a gene come in many forms—alternative transcription start sites, splicing, and polyadenylation can all result in different isoforms. In the original version of the manuscript, we only evaluated tissue-specific total gene expression and tissue-specific alternative splicing. Polyadenylation was only analyzed in a tissue-pooled setting. In response to suggestion #5 below, we have restructured and updated the manuscript with a new section and figure (Figure 2), which we believe clearly assesses Borzoi’s capabilities in predicting tissue-specific regulation, including expression, alternative start sites, alternative polyadenylation sites, and (as detailed in Supp Figure 2) alternative splicing. In particular, we clearly state that tissue-specific alternative splicing is not learned well by the model. See our response to suggestion #5 below for more details on the new analysis and figure.

CONCLUSIONS: Though transparently declaring I struggled with this review, I hold reservations about most test datasets. I could not understand what was being shown or compared in many figures. Therefore, I am circumspect and uncertain about most conclusions. I feel Borzoi may well identify key transcription factor binding motifs likely to influence tissue-specific isoform

expression of genes. I am not convinced Borzoi can reliably predict RNA isoform expression based on DNA sequence alone. I feel somewhat convinced Borzoi can assist with prioritising SNVs that may affect RNA isoform expression and annotated versus unannotated (erroneous) polyadenylation. I am unsure if Borzoi is superior to other tools.

SUGGESTED IMPROVEMENTS

1. Define acronyms and explain in plain language what each assay does that Borzoi is trained on. For each technique explain simply what you hope the algorithm might 'learn' from this dataset. Clarify the specific nature of RNA-Seq data used from ENCODE i.e. whether PolyA enriched, length of reads etc. These details should be provided for all training data in Methods.

We have reviewed the manuscript to ensure all acronyms have been defined and modified the language to be more plain and clearly state the learning objectives. We have expanded the subsection "Training data" in Methods (Line 481) by describing what we hope to learn from each experimental assay trained on:

"The training data for this analysis consisted of a large set of human and mouse RNA-seq experiments. To help the model utilize important sequence features for making its RNA coverage predictions, we also included the experimental assays studied by the Enformer and Basenji models in the training data [8, 9, 13]. This includes a curated set of human and mouse CAGE assays from the FANTOM5 consortium, which we hypothesized would help the model relate TSS usage and strength to RNA-seq coverage between multiple (alternative) TSS [42, 105], as well as DNase and ChIP-seq from ENCODE / Epigenomics Roadmap [31, 38] and pseudo-bulk single-cell ATAC-seq data from CATlas [106, 107], which focuses the model towards distal regulatory elements."

A shorter summary of this is provided in the Results section (Line 81 of the revised manuscript).

The RNA-seq data from ENCODE that was used for training was quite diverse in experimental configuration; many experiments are polyA-enriched, some are not enriched. Similarly, some experiments are limited to the nuclear fraction of RNA, some are cytosolic RNA and while others are total RNA-seq experiments. All details of these experiments are available via the ENCODE portal (https://www.encodeproject.org/), which even supports aggregating various statistics from a set of RNA-seq experiments. Hence, we feel that it is better to refer the reader to this portal for more detailed information on the experiments. We provide a file which lists all of the trained-on RNA-seq experiments from ENCODE on our Github: https://raw.githubusercontent.com/calico/borzoi/main/examples/targets_human.txt. We can also upload this file as an extended data item if and when the manuscript is published.

We have updated the "Availability of data and software" statement to highlight the file enumerating trained-on RNA-seq experiments, as well as refer the reader to the ENCODE portal for further details and statistics on experiments.

2. Provide high-level explanations of the critical thinking and content of each TRAINING and TESTING dataset. Explain and justify how you mitigated sources of bias in your TRAINING set for each TESTING set.

The human reference sequence was divided into train and test sets randomly, which is a standard and widely accepted way of generating genome-wide training data for sequence-based ML models. During this phase, there are no train versus test datasets; all datasets are included. We aim to generalize across sequences, not datasets. Most downstream tasks apply the model without fitting any new parameters, e.g. scoring distal enhancers via model gradients. These examples perturb the sequence away from the reference and correct prediction requires correct causal attribution by the model. For the e-/s-/paQTL analyses, we do choose negative examples to match basic properties of the fine-mapped causal sets. We reviewed the manuscript to carefully describe these procedures.

3. Limit field-specific jargon. Explain terms/techniques like: “saliency scores”, “attention matrix”. I remain unsure exactly what you are correlating in a “Gene Level Pearson correlation” – relative splice-junction usage across the predicted isoform, akin to a sort of Anova?

We have elaborated on the definitions of these terms in the revised manuscript. Specifically, we have updated the paragraphs in the Results sections where these terms first appear as follows:

Attention map:

- Line 106 : “In order to predict RNA coverage, Borzoi must have internally learned a detailed representation of gene structure. We hypothesized that this learned structure is embedded in the higher-level attention layers (in their *attention maps*, which are dynamic matrices that specify from which input positions an output position will receive information; Methods).”

Gradient saliency:

- Line 115: “We can visualize changes to global interactions in the attention map through in-silico sequence perturbation. Furthermore, while the attention map has 128 bp resolution, we can elucidate base-resolution determinants by estimating the contribution that each nucleotide in the input sequence has on the total attention in a region of the map, e.g. based on *gradient saliency* ('Gradient x Input'; Methods).”

More details on saliency scores are given in subsection “Input sequence attribution” in Methods (Line 637). More details on the attention visualization are provided in subsection “Attention matrix visualization” (see Line 791 for a mathematical definition of the attention scores).

Finally, we address the reviewer’s question about gene-level analysis in comment #4 below.

4. Explain key results plainly. I am unable to decipher: “To study predictions at the gene-level, we sum coverage in bins overlapping exon annotations and compute log2. Transforming the experimental and predicted RNA-seq coverage similarly, we observe a mean 0.89 Pearson R

across genes (Figure 1D). Finally, by quantile-normalizing the gene-level predictions and subtracting the average expression value taken over coverage tracks, we observe a mean 0.62 Pearson R (Figure 1E), indicating that the model explains a significant amount of the variation observed between tracks (such as tissue- and cell type-specific differences)."

The model is gene-agnostic in the sense that it has zero knowledge of where the genes are—it merely outputs aligned read coverage predictions in 32 bp bins. If we want to study genes with the model, we need to aggregate those bin predictions that overlap gene exons. In the analysis referred to by the reviewer, we summed coverage predictions in bins overlapping gene exons. As is standard in RNA-seq analysis due to the large dynamic range across many orders of magnitude, we subsequently computed a log transform. As is also standard in RNA-seq analysis, we carefully normalized the gene expression vectors across experiments before comparing them; we chose quantile normalization as the most conservative method. Finally, we subtracted each gene's mean expression across experiments so that the value represents the residual expression in that experiment beyond the mean. If a gene is liver-specific, it now has small negative values in non-liver experiments and large positive values in liver experiments. Only after that transformation can we properly evaluate the model's ability to predict tissue-specific expression. This analysis has no relationship with ANOVA.

We have reviewed the presentation of this analysis to clarify these aspects. Specifically, on Line 91 in the Results section of the revised manuscript, the sentence has been updated as follows:

"To study predictions at the gene-level, we sum and $\log_2$-normalize coverage in bins overlapping the exon annotation of each gene (using the average prediction of the model ensemble). When comparing predicted to measured gene-level coverage values, we observe a mean 0.87 Pearson R across held-out genes (0.86 per model replicate) (Figure 1D; Supp Figure S1E)."

We have also updated the following sentence, which starts on Line 96:

"After quantile-normalizing the gene-level predictions across experiments and subtracting each gene's mean expression (so that the value represents the residual expression in that experiment beyond the mean expression of the gene), we observe a mean 0.58 Pearson R (0.55 per replicate) (Figure 1E), indicating that the model explains a significant amount of variation observed between tracks (such as tissue- and cell type-specific differences)."

5. Consider detailed evaluation of higher confidence, albeit smaller, TESTING sets to enable a non-expert to more readily critically evaluate Borzoi's predictive abilities and accuracy. For example, the ability to understand Figure 1 would be transformed if it focussed upon a 'Hold-out TESTING set' of ~30 genes from among the ~1000 genes that show the most tissue-specific differences in expression level and isoform expression. Strategically include genes that stress test different aspects of Borzoi's predictive capabilities. Any scientist can understand critical thinking. For example:
A) ~5 'Test Genes' with expression tightly restricted to a specific tissue. Most organs (and particularly skeletal muscle, kidney, eye, heart, brain) have genes expressed predominantly or

exclusively in only those tissues. Show Borzoi predicts expression of rod and cone genes in the eye, and nowhere else. This is easy to understand.
B) 5-10 ‘Test Genes’ expressed in a couple of tissues, not in others, and further, select example genes with empirical evidence in RNA-Seq data that confirms tissues-specific isoform expression in the tissues they are expressed in. Show Borzoi can pick this up (or not).
C) 10-15 Test Genes that may be expressed ubiquitously though use alternative TSS in different tissues, some with different patterns of alternative splicing of exons within the gene body, some with annotated use of different PolyA/3’UTRs.
D) Go deep into these genes. Articulate there is no data relevant to these 20-30 genes in the TRAINING set, mitigating possible sources of predictive bias. Expand your analyses to hundreds or thousands of genes if you must – but if so, explain how you have controlled for this by not training your algorithm on any of the data you are asking to offer a prediction for. I can’t figure out how you removed all of the relevant data from the Training Set using such large Testing sets.

We agree with the reviewer that the manuscript would benefit from a clearer, more structured evaluation of tissue-specific regulation across the different layers of regulation presented in one place, as well as including a number of test set predictions for genes that exhibit each form of tissue-specific regulation (overall expression, TSS isoforms, 3’ UTR APA isoforms, alternative splicing isoforms). To this end, we have written a new subsection in the Results titled “Inference of tissue-specific expression and TSS usage” (Line 133). We have also moved up and deepened the sections on alternative polyadenylation (Line 152) and alternative splicing (Line 169). These analyses are now jointly presented in a new figure (Figure 2 of the revised manuscript) and evaluate, step-by-step, Borzoi’s capabilities in predicting: (1) differential expression (Figure 2A-B), (2) differential TSS usage (Figure 2C-D), (3) differential APA usage (Figure 2E-F), and finally (4) differential splicing (Supp Figure S2K-O). We focused on the same 5 GTEx tissues that we analyze for TF motifs in Figure 3 (Figure 2 in the original version of the manuscript). For all analyses, we tested Borzoi’s ability to predict the fold change of either expression or isoform coverage ratios in a given tissue with respect to the mean expression (or isoform ratios) of the 4 other tissues. These evaluations were done on all held-out test genes, and we used either one or all 4 of the model replicates when making predictions. These numbers are presented in Figure 2B, 2D and 2F. We argue that this is a large-scale, genome-wide, unbiased analysis.

We visualize specific examples in Figure 2A, 2C and 2E (more examples are displayed in Supp Figure S2B, S2C and S2G). For the tissue-specific expression examples, we show predictions and measurements in all 5 tissues. For the TSS- and APA examples, we only show predictions and measurements in the two most differential tissues due to space constraints (these plots take up a lot of space due to isoform annotations). The visualized genes were all part of the held-out test set. The specific criteria we used for selecting these examples are detailed in the new subsection “Tissue-specific expression, TSS and APA predictions” in Methods (Line 619). Briefly, we visualized genes with highly differential expression by picking the examples with largest measured fold changes of GTEx TPM values. We picked genes with differential TSS isoforms from the set of genes with large measured fold change in TSS usage according to

FANTOM5 CAGE data. We picked genes with differential APA isoforms from genes with large measured fold change in 3' coverage ratio between the target tissues (based on the measured GTEx RNA-seq coverage), but we cross-referenced these genes against PolyASite 2.0 and only visualized examples with upregulated differential distal usage in reasonably matched cell types (to mitigate selecting genes where differential 3' usage is largely driven by RNA-seq bias).

Finally, we evaluate Borzoi's performance on tissue-agnostic splice site detection in Supp Figure K-M, and highlight some very clear examples of tissue-specific alternatively spliced genes which Borzoi fails to predict in Supp Figure S2N-O. (These results were presented in the original version of the manuscript, in what used to be Supp Figure S6C-D.)

6. Define exactly what is being measured or evaluated in each Figure. As another example, for Figure 6, what qQTLs were selected in the Test set. Why? How did you control for bias in your Training Set. How exactly did you compare Borzoi with Pangolin? How did you assign each qQTL variant in your Test Set as a True Positive or True Negative. Just like polyadenylation can occur erroneously at motifs that mimic genuine adenylation signals, cryptic splice-site usage also occurs erroneously. I am an expert in splice-site recognition and usage, hold deep expertise in SpliceAI and Pangolin, and genuinely haven't any idea what you have done or shown in Figure 6.

We have reviewed all figures to exactly define what is being evaluated. The reviewer repeatedly wrote "qQTL", which is not a term that we used in our manuscript or have encountered before. We'll guess that the reviewer is referring to splicing QTLs since they reference Figure 6, SpliceAI and Pangolin.

As described in the Methods section "Fine-mapped sQTL classification task" (Line 919 in the revised manuscript) and in the Results (Line 378), we collected splicing QTLs from the eQTL Catalog (Kerimov et al., 2021, Nature genetics) which have originally been discovered through statistical fine-mapping of GTEx RNA-seq data. We selected the set of 4,105 variants that were determined by fine-mapping to have >0.9 probability of being the causal variant for the association. We evaluated Borzoi and Pangolin's ability to classify these variants from a negative set of variants that we selected to control for splice site distance and gene expression. There is no additional training in this analysis, so there are no train/test splits. Neither Borzoi or Pangolin train on genetic variation data, so all variants compose the test set for metric calculation. To make this point clear (for all variant analyses), we have added the following sentences to the Methods section "Training" starting on Line 564:

"In subsequent analyses (e.g. variant effect prediction in Figure 5), we make no distinction between train- or test splits of hg38. This technically means that the ensemble is applied to genomic loci seen during training. We argue that these are still unbiased analyses, as the evaluations are done on out-of-domain measurements not trained on (e.g. the alternative alleles of fine-mapped QTLs, and their estimated effects, were not part of the training data)."

We derive simple but distinct statistics from the Borzoi versus Pangolin model output because the models are very different. For Pangolin, we take the absolute-valued maximum difference between splice site classification probability across all positions (described in the Methods section “Comparison to Pangolin” on Line 935). For Borzoi, we take the absolute-valued maximum difference in normalized coverage across the gene span (Methods section “Splicing variant effect prediction” on Line 927). In Figure 7D-E of the revised manuscript (Figure 6D-E of the original version) we compare Borzoi’s and Pangolin’s ability to classify the positive sQTL SNPs from the negatives, using the metrics described above. Figure 7E displays the classification performance of each model when evaluated on the subset of SNPs that are within a specific distance to an annotated splice site (the distance threshold is varied along the x-axis). We have rewritten the paragraph starting on Line 388 to clarify this:

“When comparing Borzoi to Pangolin [16] at the task of classifying the causal sQTLs from matched negatives, Pangolin has a slight advantage (Figure 7D-E; dAUPRC = 0.01, evaluated on all SNPs within a distances <= 10,000 bp from an annotated splice site). Most far-away SNPs are de novo splice-gain mutations and are relatively easy for Pangolin to classify based on the local predicted effect at the variant allele, while Borzoi's splice-gain predictions in intronic regions appear less well-calibrated. In contrast, Borzoi is better at distances closer to the junction (Figure 7D-E; dAUPRC = 0.02 when only considering the subset of variants that occur at distances <= 200 bp). Importantly, the average rank prediction of both models is superior to either model alone (dAUPRC > 0.02). More examples are shown in Supp Figure S7C-D. Supp Figure S7E includes a tissue-pooled benchmark.”

We do not specifically consider whether the variants represent “cryptic” splicing. Every positive variant was defined by the genetic analysis from the eQTL Catalog (Kerimov et al., 2021) as generically influencing alternative internal exon usage. Every variant from the negative set lacked such evidence in every GTEx tissue studies.

7. If the purpose of developing the Borzoi tool was to assist identification/interpretation of human genetic variation associated with genetic conditions or susceptibility risk factors, I cannot understand rationale for including murine RNA-Seq data. The vast body of scientific literature indicates that variation in gene/isoform expression is what underpins most variation over mammalian evolution. Please justify inclusion of murine RNA-Seq data in the training set.

The mouse data was included because it doubles the number of genomic sequences available for training and leads to better models for both human and mouse data. We studied this in great depth in a previous publication (Kelley, PLoS Comp Bio). In the revised manuscript, we have added a supplementary analysis as Supp Figure S5D where we train instances of the Borzoi model on both human- and mouse data or on the human data only (details are provided in Methods, starting on Line 569). The results are summarized on Line 305. In brief, we demonstrate greater gene-level prediction performance on held-out human test sequences and greater eQTL coefficient concordance for models trained on data from both human and mouse.

OTHER COMMENTS:

BOLD STATEMENTS THAT DON'T MAKE SENSE:

"Borzoi accurately predicts RNA-seq and other sequencing assays"
"Borzoi - A neural network for predicting RNA-seq coverage from sequence"

Do you mean Borzoi accurately predicts the pattern of alternative splicing of known alternatively spliced genes from the DNA sequence alone?

To support these statements, one must show more robust case-level data. It is not clear to me why EGFR was selected (known alternative splicing in different tissues?). It is not explicitly clear that all EGRF data was removed from the TRAINING set.

These sentences merely describe our work—we trained a machine learning model to predict RNA-seq coverage data. We specifically stated in the original version of the manuscript that the model does not learn alternative splicing well (see subsection "Inference of splicing events from RNA coverage" on page 10-11, and Supp Figure S6C-D, of the original manuscript). Nevertheless, alternative splicing represents one case of regulatory biology, and the lack of performance on that specific task doesn't invalidate the sentences describing the work, nor the performance on other tasks which are generally higher than state-of-the-art models.

We hope and believe that we have provided plenty of new evidence (including numerous case-level test data) in new Figure 2 of the revised manuscript, along with the new or updated sections related to the figure (Line 133 - 185 in the Results section), to convince the reviewer of the following claims: (1) Borzoi predicts differential expression of tissue-specific genes, (2) Borzoi predicts differential alternative TSS usage across tissues, (3) Borzoi predicts differential 3' UTR APA across tissues, and (4) Borzoi does not predict alternative splicing well. These regulatory mechanisms are quantified based on the predicted RNA-seq coverage patterns, which indeed are predicted solely from DNA sequence, on held-out test sequence only.

Finally, the INSR gene (which is the current example visualized in Figure 1B) was part of the test set. We chose to visualize this gene because it is a relevant gene for many genetic researchers, and it is also a fairly complex gene with many exons. This is clarified in the corresponding section in the revised manuscript.

OVERSTATEMENTS:

Borzoi outperformed CADD (mean AUROC 0.54 vs 0.53). Do the authors consider this significant? What variants comprised the Test Dataset? Why? How did you mitigate sources of bias from your Training Set? How did you assign each Test Variant as a True Positive or True negative to event derive a AUROC?

We agree with the reviewer that the difference in performance between Borzoi and CADD is only marginal, and so we have updated the corresponding sentence (Line 322 in the revised manuscript) and removed the claim that we outperformed CADD.

The details of which variants were curated as positives and negatives for this task are presented in the Methods section “Classifying rare and common variation from gnomAD” on Line 956 of the revised manuscript, which reads as follows:

“We sampled a set of 14,198 singletons and 14,198 matched common variants (allele frequency >5%) from the GnomAD v3.1 database (https://gnomad.broadinstitute.org), with sampling restricted to regions overlapping ENCODE cCREs. To control for sequence mutability, we excluded variants within CpG islands and low-complexity regions. For each singleton sampled, we sampled a negative example as a matched common variant with the same reference and alternate allele as the singleton. We also matched the variants’ background DNA contexts, sampling common variants that lie within the same tri-nucleotide as the singleton. Finally, we removed variants overlapping gene exons in coding sequences (GENCODE v41), focusing only on regulatory variants for our evaluation. For all sampled variants, we used their CADD raw score and CADD phred scores (v1.6) from the GnomAD v3.1 dataset. We trained ridge regression models to discriminate common variants from singletons and used 10-fold cross-validation to evaluate the models. The CADD-based model uses the CADD scores as features, whereas the Borzoi-based model uses the L2 scores across all RNA-seq tracks as features, averaged across the four model replicates. We derived a third (combined) model by averaging predicted variant ranks for the Borzoi-based and CADD-based models. For a second genome-wide benchmark, we sampled uniformly from across the genome instead of restricting the variant sampling to ENCODE cCRES. This resulted in a variant set containing 17,360 singletons and 17,360 matched common variants.”

As described in more detail above, we, and others training models in this framework, consider all variants to be unseen data, as Borzoi was trained on measurements from the human (and mouse) reference genome only.

CAVEATS WITH TEST DATASETS.

The eQTL and common versus rare gnomAD variants is encouraging, but there caveats with these Test Datasets (and again is not clear how data was held-out from Training Set). The zygosity of rare variants can be important re inferred likely impact on RNA expression levels or isoform. Similarly, common variants can exert an effect upon RNA expression levels or isoform that is functionally benign. I appreciate the value in a genome-wide approach. However, surely it is more robust and rigorous to evaluate (a smaller) TESTING set of variants empirically confirmed to alter isoform expression - and prove conclusively via precision-recall analyses that Borzoi picks these easy and better than any other tool.

For example, re the ability of Borzoi to interpret SNVs that may affect isoform expression. One could readily curate 50-100 True Positive variants from literature in UTRs or non-coding regions

that are classified Pathogenic in ClinVar, which don't affect an annotated splice-site, for which empirical data confirms an affect upon isoform expression levels or processing. Recurrently identified Start loss or Stop loss pathogenic variants from Clinvar might be OK to also include in this mix. Combine these True positives with an equal number of similarly positioned (with respect to location in exons, introns, promoter regions, UTRs) common variants in gnomAD (for the reasons you describe). Sure it's a smaller test dataset, but MUCH higher confidence. This would give compelling evidence for sensitivity and specificity of Borzoi versus other tools.

As explained in our response to the comments above, we consider all genetic variants (and their measured/annotated impact or quantitative effect on any regulatory mechanisms relative to the reference nucleotide) to be held-out test data, as Borzoi (or any other benchmarked model) was trained on measurements mapped to the reference genome only.

Also, while we appreciate the reviewer's opinion, we disagree that ClinVar would be better, higher quality, or less biased than fine-mapped QTLs. We believe that unbiased genome-wide test sets such as e-/s-/paQTLs (estimated from standard RNA-seq protocols, which supports quality controls, significance tests, and even visualization of raw sequencing coverage of individuals harboring the variant allele) is better than a small set of manually resolved variants (vanishingly small, if we were to look only at non-coding CRE or 5' / 3' UTR variants).

CLARITY

Numerous abbreviations neither defined nor explained. TF, TF CHiP, ATAC, CAGE assays are not understandable terms. Many clinician geneticists will not be aware what these assays are or show.

There is an explanation of ChIP-seq, ATAC-seq, and DNase-seq in the Introduction (Line 22 of the revised manuscript): "For example, TF ChIP-seq or DNase/ATAC-seq reads indicate a transcription factor (TF) binding event or accessible DNA at the site where the reads align.".

We have added a brief explanation of CAGE-seq in the Introduction (Line 34): "Other models have been trained to predict CAGE-seq data, which measures expression at the 5' end of capped RNA and does not capture coverage at individual exons.".

The clear explanation in Methods why 32 nt resolution is better than the 128 nt of the Enformer algorithm was welcomed. However, rationale why 32 nt was selected and not another window length is not provided.

We explain in the Results section, in the paragraph starting on Line 65 of the revised manuscript, that we believe single nucleotide resolution would be optimal for modeling RNA-seq, but due to computational limitations we cannot go to a resolution finer than 32 nt. If we want to model a large input sequence (524kb), we cannot have that model predict at finer resolution than 32 bp without running into issues with GPU memory or infeasible training time. Our technique for increasing the resolution beyond the 128 bp bins used in the attention blocks

(which again are maxed out at 128 bp resolution due to GPU memory constraints) is to use U-net blocks with 2x upsampling. Hence, 32bp is better than 64bp resolution (and it fits on the GPUs we have access to), while 16bp causes problems.

Line 376: The authors note “We processed the data slightly differently relative to prior analyses” – but do not explain to which ‘other studies’ they refer.

We have updated this sentence and added references to the Basenji2 and Enformer papers (Kelley et al., 2020, PLoS Comp Bio, and Avsec et al., 2021, Nature Methods, respectively).

Similarly Line 394: The authors note “In contrast to previous efforts in which a single train/validation/test split is chosen”, but do not define to which ‘previous efforts’ they refer.

We have updated this sentence and expanded on the details of train/test set construction.

How is one supposed to understand what the authors have done and how it is different or better to other studies?

We train a model that predicts RNA-seq coverage in 32 bp bins given 524 kb sequence as input, and this model is shown to predict coverage on held-out test loci (and the held-out test genes located within these loci) with high concordance to measured RNA coverage.

We compare this model to a number of other models (Enformer, APARENT2, Pangolin, CADD) on various tasks, for example to predict functional variant effects, and demonstrate by standard metrics (Spearman correlations, AUPRCs or AUROCs) that Borzoi performs competitively, or better than, all of these models. Importantly, when we evaluate these models on perturbation-based benchmark tasks (e.g. variant effect prediction or distal enhancer prioritization), we assume that all measured effects from these data constitute an unbiased, large-scale test set, seeing as none of the models were explicitly trained on measurements of both the reference and variant sequences.

We believe this is how one is supposed to understand the key results of this paper. We hope that with the revisions and clarifications made in response to the reviewer’s comment above, the manuscript has become easier to understand.

A lot of important information and critical thinking is buried in high technical detail or is omitted.

By updating the manuscript in response to the reviewer’s suggestions above, e.g. by defining acronyms and abbreviations, by explaining the rationale of train/test splits, and by stating our assumption on what constitutes unbiased test data for each respective benchmark task, we believe that the important missing information requested by the reviewer has been provided. We hope that these actions have clarified the contents of the manuscript; if not, we kindly request the reviewer to provide examples of any outstanding items that need clarification.

Response to Referees - Borzoi Manuscript, Nature Genetics (NG-A63358R)

Reviewers' Comments:

(Author responses = blue)

Reviewer #1:

Remarks to the Author:

The authors have addressed most of our concerns. Moreover, they have spotted and addressed a train-test leak we had not noticed. We are now raising two major points related to train vs. test performance which we think are important to be addressed.

We thank the reviewer for their comments.

Major points

---

1. We found a substantial discrepancy between train set performance (with profile Pearson R > 0.85 across tracks, not mentioned in the manuscript) and the test set performance (approximately 0.75 Pearson R, Fig 1C), consistent with the drop in performance between the original manuscript and the revision. Is this gap indicative of overfitting, and could it negatively impact variant analyses or other downstream evaluations? Perhaps, the choice of the smallest test and validation folds (3,4) likely exacerbates this problem. The authors should include a discussion on this topic to highlight potential issues in downstream analyses. For comparison, train vs. test performances should be provided for Enformer.

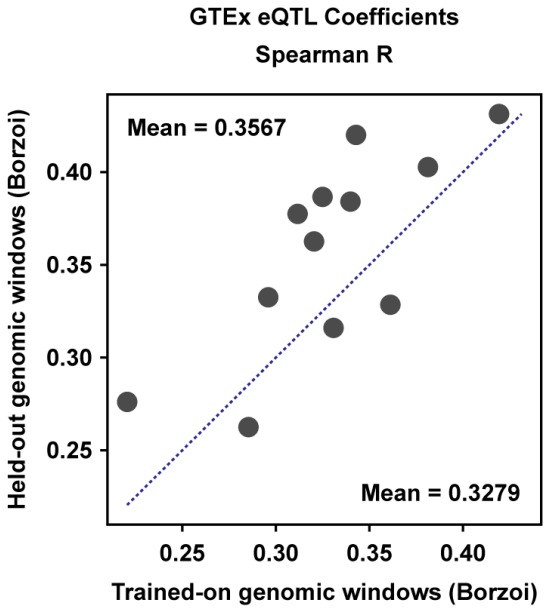
We understand and appreciate the reviewer's concerns regarding the difference in bin-level Pearson R correlation between trained-on and held-out genomic loci.

First, to clarify, we stopped training each replicate once the average Pearson R on the validation set had not increased for 15 epochs, and used the weights from the best-performing epoch according to the validation metric as the final model. So the model did not overfit in the classical sense of training to the point of reduced performance on held-out examples. Better performance on the training set versus held out sets is a ubiquitous phenomenon in machine learning, especially neural network models with larger parameter counts, and does not indicate a problem. Reducing the gap between train/test performance would require constraining the model and lead to reduced test set performance, which is ultimately the most important metric. Per the reviewer's request, we have added Enformer's (and Borzoi's) train- and test set metrics as a new panel in the revised manuscript (Supplementary Figure S1E). The difference between these sets is very similar for Enformer.

The reviewer raises interesting questions: Does an increased performance gap on the training set relative to the validation/test set translate to worse performance on variant effect prediction

tasks (e.g. due to the possibility of diminished quality in the feature attributions)? Would a model that had stopped training earlier (or had more regularization) result in better variant effect predictions, even if validation performance on wildtype sequences was lower?

While we cannot answer these questions perfectly without laborious ablation experiments that we feel are out of scope (namely re-training & evaluating Borzoi models with different levels of regularization), we can compare the eQTL prediction performance of the existing model ensemble on trained-on vs held-out genomic windows. In **Response Figure 1** below, we do this using a *gene-agnostic* eQTL coefficient analysis. As a reminder, we calculate a gene-agnostic variant effect score as the difference between the sum of log(coverage + 1) for the reference and variant allele, and compare this predicted effect to the beta coefficient estimated for that SNP averaged across all significant eGenes (within the output window). This is identical to the analysis presented in Figure 5D of the paper. We observe a small overall difference in eQTL coefficient Spearman R between train and valid+test set (< 0.03 on average across tissues), i.e. there is a marginal decline for trained-on windows (although there is admittedly a relatively small sample size to evaluate performance on held-out loci).

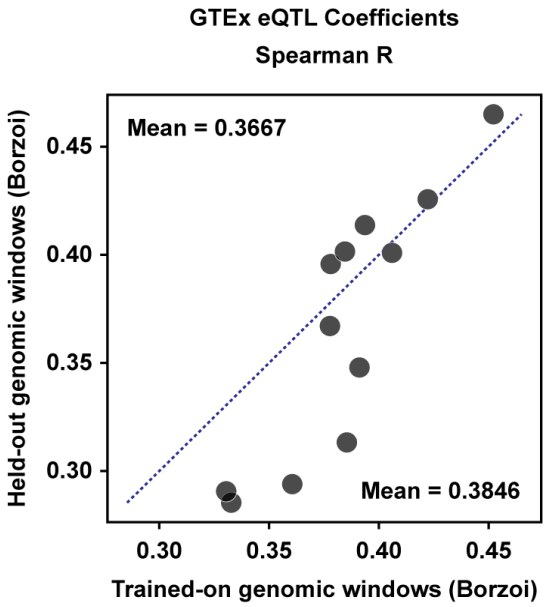


Response Figure 1: Variant effect prediction performance (eQTL coefficient Spearman R) when separating SNPs based on whether they occurred in trained-on or held-out hg38 sequence windows during Borzoi training. Each dot represents a tissue. Only tissues for which there were at least 200 SNPs in held-out sequence windows were included.

When we re-run the evaluation above with a *gene-specific* score, calculated as the difference in log(sum coverage) for a specific target gene (and compared to the specific beta coefficient of said gene), we obtain the results shown in **Response Figure 2** below. The difference in performance on trained-on vs held-out hg38 windows remains small (< 0.02 Spearman R), but now the difference points in the opposite direction (better performance on trained-on windows). Note: The gene-specific eQTL coefficient analysis was also the method used for the ablation

experiments in Supplementary Figure S5E of the revised manuscript. We only use the gene-agnostic method for less biased comparisons against Enformer; the gene-specific method is our recommended way of predicting effect sizes (which we now state in the Discussion).

Thus, we believe that there likely is no real difference in performance, and what we're observing is stochastic. Remember that many SNPs are shared across tissues, so even though multiple tissue dots appear to be shifted off-diagonal, they are driven by partially overlapping variant sets, lowering significance. Anecdotally, we are aware of several additional research labs working in this space that have similarly observed that variant effect prediction quality is consistent between training and held out sequences. Hopefully this convinces the reviewer that there is likely **not** a meaningful difference in performance.



Response Figure 2: Variant effect prediction performance (eQTL coefficient Spearman R) when separating SNPs based on whether they occurred in trained-on or held-out hg38 sequence windows during Borzoi training. The variant effects were computed as gene-specific scores and compared to each SNP-gene pair's beta coefficient.

We have added these plots as a new supplementary figure (Supplementary Figure S5D), and we have also revised the figure captions & manuscript text to comment on these observations.

Specifically, in the Figure caption of Supplementary Figure S1E we added the following text:
"... Note that, for both Borzoi and Enformer, the train set Pearson R is higher than test set Pearson R, even though model training stopped once validation performance saturated.".

Second, in the Figure caption of Supplementary Figure S5D, we added the following:
"Overall, only a small difference in average Spearman R is noted between trained-on and held-out genomic sequence windows (< 0.03)."

Finally, we added the following to the Discussion, starting on Line 445:
"We also investigated whether variant predictions were better or worse in genomic sequence windows seen during training. We found only small differences in eQTL Spearman R for trained-on windows, suggesting a minor impact."

2. The authors mention (lines 88-90) that test set prediction performance improved with the use of more than one model replicate. Given the quadruple runtime, the benefit of using multiple replicates for a minute improvement (0.74 to 0.75 Pearson R across tracks) should be critically evaluated. Which replicate showed the strongest performance? Single replicate performance should be presented with standard deviation to facilitate a proper study of performance variance, as claimed by the authors (line 83).

Per the reviewer's request, we have added a new supplemental panel related to Figure 1 (Supplementary Figure S1F), evaluating single replicate performance across held-out loci for bin-level, gene-level, and quantile-normalized+mean-subtracted predictions. Each plot compares the mean Pearson R as well as the 5th & 95th percentile of Pearson R values across RNA-seq tracks (which we felt was more appropriate than standard deviation) for each replicate or for the full ensemble. The results show that the replicates perform very similar and no replicate is the best across all three types of evaluation, while the ensemble consistently has the highest mean, 5th and 95th percentile values. These observations agree with the performances reported in the bar chart of Figure 2B (e.g. replicate 2 has highest blood-specific log-FC Spearman R among replicates, while replicate 0 has the highest muscle-specific Spearman R). The Results section has been expanded, starting on Line 95, to state these findings:

"The concordance with measured bin-level and gene-level coverage increased consistently when averaging predictions across the 4 model replicates (Supp Figure S1F-G). Conversely, no individual replicate was best among all other replicates by both metrics simultaneously."

Also note that while the average bin-level performance indeed increases only modestly when using the ensemble, the use of the full ensemble results in larger performance increases by other (perhaps more interesting) evaluation metrics on downstream tasks. For example, the ensemble performances reported in Figure 2B for tissue-specific gene expression prediction display increases in Spearman R ranging from 0.025 to 0.045 compared to the average replicate performance. As another example, on Line 307 in the manuscript we report that the eQTL coefficient performance drops from 0.334 Spearman R to 0.292 across tissues when using only a single replicate. In general, fold change predictions (in the form of variant effect predictions, or in the form of comparisons between tissues) benefit the most from ensembling. The Discussion section has been expanded, starting on Line 433, to clarify this point:

"While the increase in performance due to replicate ensembling was relatively small for bin-level test set predictions, the gain was more noticeable when predicting gene-level fold changes across experiments or when predicting eQTL effect sizes."

Minor points

---

3. In the “Model ablation experiments” section (lines 578-579), the authors refer to a “baseline model” trained on 4 folds. Is this model still available, or is it based on a single fold with 4 replicates? Have the ablation experiments been conducted against the correct common fold? This should be clarified.

We thank the reviewer for verifying this with us. Indeed, the model ablation experiments used a correct cross-validation strategy (the training bug only occurred for the original Borzoi models which we trained on a cloud service; there was a bug in our cloud-based train script). Hence, the current description of ablation cross-fold training on Line 598 remains accurate.

4. The authors state that model performance is challenging to compare to Enformer due to differences in data processing. However, since the data processing for ChIP, DNase, ATAC, and CAGE is similar (except for different binning as noted in line 82), it is unclear why Borzoi underperforms on every common assay compared to Enformer, except for CAGE, even when adjusting for 128bp resolution. This statement should be considerably weakened, and a discussion should be included on which modalities Borzoi is suitable for, and where Enformer still outperforms. The authors may discuss why a more locality-enforcing architecture like the U-Net used in Borzoi performs worse than the Enformer on more local assays.

There are several potential reasons for why the reported bin-level Pearson R metric evaluated on CAGE tracks is different compared to Enformer. First, it is possible that Borzoi improved its CAGE predictions compared to Enformer due to positive interactions with the RNA-seq tracks (since the TSS regions are marked by the start of RNA-seq coverage and CAGE coverage). Second, we processed the CAGE experiments into separate stranded tracks this time, whereas we studied unstranded merged tracks in previous work, including Enformer. Stranded processing fundamentally alters the distribution of peak densities per track compared to the original Enformer CAGE profiles, and it is unclear how these changes may inflate or deflate average per-track bin-level Pearson correlations. Training on tracks separated by strand may also further improve Borzoi’s ability to predict CAGE.

The only assay where Borzoi is substantially underperforming Enformer is for DNase experiments. We believe there are two main reasons for this. First, the published Enformer model was actually fine-tuned on the human data only (without mouse data) as a final step. This improved Enformer’s test set performance on human experiments (although it did not improve variant effect predictions, which is why we did not do this for Borzoi). In Supplementary Figure S1E, we include the test set performance of Enformer before this final fine-tuning step (with permission from Ziga Avsec, the first author of Enformer), and as can be seen the average DNase Pearson R is now near-identical to Borzoi. Second, Enformer is a larger neural network in terms of the number of weights. It applies residual convolutions in the convolution tower, softmax-pooling layers, and 11 attention layers, resulting in a total of ~245M parameters; compare to Borzoi’s ~185M per replicate, which includes the U-net weights. Furthermore, Enformer is tasked with learning fewer experimental assays (namely, no RNA-seq and fewer

DNase & ATAC assays). Thus, Enformer does not need to dedicate any parameters to learning post-transcriptional regulation, such as polyadenylation. It is reasonable to speculate that increasing the size of the model to match (or exceed) Enformer might further improve Borzoi's held-out performance, seeing as the architectures are basically identical besides the U-net.

We have expanded the discussion of the differences in performance between Borzoi and Enformer, starting on Line 89 in the Results section and on Line 447 in the Discussion, to clarify the above points.

Line 89: "Performance is difficult to compare directly to Enformer due to differences in data processing and modeling choices (Methods). Nevertheless, test accuracies on overlapping datasets are broadly similar (Supp Figure S1A-E), with two exceptions (the average Pearson R is lower for Borzoi compared to Enformer for DNase tracks, while it is higher for CAGE tracks). We return to this observation in the discussion."

Line 447: "Finally, as we compared Borzoi to Enformer on held-out bin-level predictions, we noticed that Enformer had marginally higher average Pearson R for DNase tracks, while conversely Borzoi's CAGE predictions had a higher correlation. However, due to differences in data processing (data splits, bin resolution, stranded tracks), it is difficult to quantify exact differences. Additionally, Enformer was fine-tuned on the human data without mouse data as a final step, while Borzoi was not. The performance on DNase tracks was comparable to Enformer before the fine-tuning step."

Reviewer #2:

Remarks to the Author:

The authors have done a fine job of addressing my concerns, no further issues on my end.

Reviewer #3:

Remarks to the Author:

I acknowledge the author's investment of time and energy to carefully and evenly address the concerns and suggestions raised by Reviewers. I also appreciate the author's patient and detailed rebuttal, which is a little easier for a non-expert to understand. For the authors, I declared transparently to the editors that I held no technical expertise in deep-learning. However, it was felt a generalised opinion from a possible future user at the research-clinical interface may be helpful.

Your rebuttal and revisions helps clarify what Borzoi is and does, at least I hope I have it right this time. Trained on small windows of gDNA sequence, uniformly processed RNA-seq from ENCODE and GTEx, paired with training datasets from the Enformer model, the authors trained four Borzoi deep-learning models. I am convinced that Borzoi has a general capacity to identify:

gene loci within the genome; and
promoter and enhancer elements regulating gene expression; and
exons demarked by flanking splice-sites within gene loci; and
transcription start sites, termination sites and polyadenylation signals; to predict
relative levels of transcript expression arising from all gene loci in different tissues; which is a transformative advance for the deep-learning field.

This explains why the Borzoi model works better when also trained on murine data – because core motifs/elements regulating gene expression and transcript content are evolutionarily conserved in mammals. It is the finer tuning of gene expression that diverges between humans and rodents.

In summary, I feel reviewer comments and author responses and emendations collectively support Borzoi as a major advancement in deep-learning. However, I am not yet persuaded Borzoi can “unravel the full transcriptomic consequences of a variant”. Trying to interpret one clinical variant as a possible cause of a patient’s condition is different to assessing large QTL datasets, which have caveats, as every dataset does. Being able to identify variants affecting gene regulatory motifs proximal and distal to gene loci has potential transformative application for clinical genomics, though the authors don’t yet have hard evidence it can do this. A Pearson co-efficient of 0.70-0.85 shows Brozoi is doing a pretty good job at spotting expression QTLs and splicing QTLs but at the single variant level there remains significant uncertainty. I think where I land is that Borzoi may provide a good variant prioritization tool for cases unsolved after genome sequencing to identify variants that may alter transcription of the target gene. A comment below relates to exactly how to use Borzoi for this purpose, which score?

I am grateful for the opportunity to review this impressive piece of work. I have learned a great deal from the authors and Reviewers 1 and 2, who offered important and careful critique of technical aspects of the deep learning. I hope using Borzoi maybe within the grasp of keen users like myself who are not technical deep-learning experts – such that we can put its abilities to good use for clinical genetics, quickly. If the authors would like stress test Borzoi on a large panel of challenging clinical variants with rigorous RNA assay data showing the exact consequences upon splicing and/or expression (incl. UTR, miRNA binding sites, adenylation, TSS and promoter variants) levels then please reach out.

We thank the reviewer for their comments and feedback. We are glad that we managed to better convey the model, its strengths & weaknesses, and its use cases in the revised manuscript (and in the response). The summary written above matches how we ourselves would describe Borzoi and the study as a whole.

We also appreciate the remaining comments raised by the reviewer, which in general relate to doing an even better job at clarifying the uses of Borzoi, in particular for geneticists and clinical researchers (such as the reviewer) wanting to take advantage of the model for variant scoring. This group of scientists represent an important target audience for this work and so we try to address the remaining concerns below as best we can, while being mindful of the space- and

word constraints of Nature Genetics. To reduce some of the complexity and length of the paper, we are also proposing to the editor that we move one of the overly technical analyses (namely the Attention matrix analysis on Line 108) to the supplement. The core results stand on their own without this analysis in the main text, and we can then focus on adding clarifications where needed to address the concerns below. We hope the reviewer agrees with this change.

Ps. The clinical variant data mentioned by the reviewer sounds like an excellent benchmark that we would be keen on applying the model to in the future. Hopefully we can get the contact details from the editor once the work is published.

I have a few additional questions and comments:

Q1: Training Set: Can you clarify if the ENCODE datasets included developmental/pediatric specimens from humans and mice. Arguably some of the most significant changes to transcript expression levels and composition is developmentally regulated. I checked methods, Lines 491-511, and could not find details.

The ENCODE datasets indeed include RNA-seq samples (and other experimental assays) across the developmental spectrum for both humans and mice. For reference, we point the reviewer to the list of trained-on experiments:

Human experiments: https://storage.googleapis.com/seqnn-share/borzoi/hg38/targets.txt
Mouse experiments: https://storage.googleapis.com/seqnn-share/borzoi/mm10/targets.txt

These are also linked from the Data availability statement and the Github repository.

To mention a few examples, Mouse Track 2216/2217 correspond to whole-brain tissue from 14.5-day mouse embryo (RNA-seq, sense/antisense), while Mouse Track 2478/2479 are whole-brain tissue from 10-week adult mice. Similarly, Human Track 6140/6141 correspond to 20-week embryo frontal cortex, and Human Track 7539 corresponds to adult brain. There are also some pediatric data available, e.g. RNA-seq in pancreas tissue in a 16-year old child (Track 6132/6133).

To emphasize this point, we have added the following sentence in Methods, on Line 509: “We collected ... RNA-seq coverage tracks from ENCODE. This set includes samples for a diverse set of tissues and cell types, with measurements spanning the developmental spectrum for both human and mouse.”

Q2: Thank you for providing more clarity around Test and Training Sets. Reading your methods, I understand only one model fold is used for the test set (rather than all four). However, I can’t glean the exact nature of hold-out Test Set for Figs 1 and 2. How many genes, why these genes were selected. Specifically with respect to predicting exon coverage for EGFR, can you clarify if all sources of data concerning EGFR were depleted from the Training Set (RNA-Seq + CAGE, DNase, ATAC, ChIP-seq) or was only the RNA-Seq data “held out”.

First, whenever a 524 kb genomic sequence window was held-out from training and used for testing, that means both the sequence as well as the associated coverage measurements across all assays (RNA, DNase, CAGE, ATAC and ChIP) were not seen during training. So yes, for the examples shown in Figure 1 and 2 (although we switched the specific EGFR example to INSR in the previous revision), all measurements for a gene were held out. To clarify this, we have added the following sentence to Methods, Line 533: “Note that all coverage measurements of all experimental assays (RNA, DNase, CAGE, ATAC, ChIP) are held out (and not seen by the model) whenever a particular 524 kb sequence window is not in the training set.”

Second, the reviewer asks why specific genes were held out from training. The held-out sequence windows (and implicitly – the genes within those windows) were randomly chosen (see Line 531). That means that roughly 1/8th of genes were not seen during training. The purpose of holding out parts of the genome during training is to determine when to stop training to avoid overfitting, by identifying when the validation performance plateaus, and then to evaluate performance. Ultimately, we believe the most interesting and important use of the model is for variant scoring (or for other kinds of sequence perturbations), for which we are confident that the model is equally accurate on trained-on vs held-out windows (because the model never sees human variant alleles, only hg38 and mm10. Thus, we did not specifically consider which exact genes were held out from training.

Q4: Lines 512-514. “We fragmented the human (hg38) and mouse (mm10) chromosomes and randomly divided these fragments into eight roughly evenly sized partitions, pairing orthologous regions into the same partition. One partition was held out for validation and testing each, and the remainder of the data (~ 75%) was used for training. This does not clarify which chromosomal regions are in the Training Data and which are held-out in the Testing Data. SpliceAI (understandably) has blind spots for genes/splice-sites on chromosomes that it was not trained on. I need to know if a deep-learning model is trained on data from the clinically relevant gene/variant I am scrutinizing clinically. Please can you clarify in the methods exactly which human chromosomal regions the model is or is not trained on (if the latter is easier/shorter).

First, we want to emphasize that we do not believe the model performs substantially worse with regard to variant effect prediction on sequences/genes that were held out during training. Please see our response to Reviewer #1 Major Point #1 where we argue for this by performing additional eQTL variant effect prediction analyses, showing that variant prediction performance is roughly equal between trained-on and held-out windows. I.e. it is marginally worse on trained-on windows by one type of variant score, and marginally worse on held-out windows by another variant score.

That said, we fully agree that a reader should be able to see for themselves which sequence window was trained on or not. The .csv file enumerating all chromosomal sequence windows, including a description of how to search for the ones held-out from training, is given in the github repository readme file (https://github.com/calico/borzoi).

The github page refers to this .csv file:
https://github.com/calico/borzoi/blob/main/data/sequences_human.bed.gz

The published model ensemble (of 4 replicates) all had data folds 3 and 4 held out (this is also detailed on the github page for future readers). This means that by searching for rows where column 4 is “fold3” or “fold4” will return sequences not trained on.

To make this more clear for future readers, we have added a description of this list to the Data availability statement.

Q5: Lines 125-132. For Fig 1 H and I - what is the real coverage for the variants affecting the splice donor and PolyA signal?
a) Are these real or synthetic variants? If synthetic, why use a synthetic variant?
b) Please provide the HGVS annotation of the donor and adenylation site variants being modelled/assessed? I can’t check if exon-skipping is likely because I don’t know exactly which variant is being modelled.
c) The real insight is whether Borzoi can accurately predict exactly HOW the mRNA will be altered from loss of the GT. Exon-skipping is only one possible outcome: many donor variants cause activation of one or more cryptic donors and/or intron retention. 50% of all donor variants create more than one aberrant transcript. Do you have the corresponding RNA-Seq data to show that Borzoi got the prediction right – or didn’t?
d) Similarly, not all cryptic AATAA/AC PolyA cleavage sites are usable and some PolyA variants just cause loss of transcripts because either adenylation does not occur properly and/or negative feedback to the RNA Pol and abnormal transcription termination. Did Borzoi get it right? Was the alternate PolyA site used as predicted?

We thank the reviewer for pointing out this possible source of confusion. To clarify, the variant examples shown in Figure 1H-I are indeed completely synthetic/hypothetical. We do not have any RNA-seq measurements that would confirm the predicted effects of the model for these particular variants. The reason why Figure panels 1F, 1G, 1H and 1I exist is that we wanted to visualize and interpret how the Borzoi model is parsing, considering, and connecting certain gene-regulatory features in the sequence when making its predictions, across the span of regulatory mechanisms (transcription, splicing, polyadenylation), and we wanted to exemplify this for a single gene in the same figure (to convey to readers that the model considers multiple layers of regulation). A common visualization technique from the Natural Language domain is to look at the attention matrices within the neural network and observe how they change as the input sequence changes. Since we could not find a gene that jointly has high-confidence promoter variants, splice-altering variants, and polyadenylation-altering variants with associated RNA-seq measurements, we picked an arbitrary gene and induced hypothetical mutations in order to highlight that Borzoi makes reasonable changes to the underlying attention matrices.

However, we agree with the reviewer that these predicted changes may in fact be less “believable” or “reasonable” than we initially thought. As the reviewer points out, a number of possible alternative scenarios could play out when these splice- and polyadenylation altering

variants are induced. We thus suggest (to the editor and to the reviewer) that we cut Figure 1F-1I from the main text and figure and move these analyses to the supplement. These examples are not required to convey that Borzoi can score transcriptional, splice-altering, and polyA-site-altering variants with reasonable accuracy, but may still be of great interest to readers interested in model interpretation and explainable AI. While moving these panels to the supplement, we will also follow the reviewer's request and annotate the coordinates of the splice-altering and polyadenylation-altering variants shown. For now, we have added the missing coordinates to panels 1H and 1I but kept them in the main figure.

Later in the manuscript, we thoroughly evaluate the model's ability to predict various modifications to the mature mRNA caused by genetic variants via genome-wide QTL evaluations and detailed example interpretations coupled with measured RNA-seq profiles, namely in Figure 5A and S5A-B for transcriptional variants, Figure 6C-D and S6E for polyadenylation-altering variants, and Figure 7C, F and S7C-D for splice-altering variants. In future work, we would like to validate variant interpretations through more elaborate measurements (beyond average RNA-seq profiles of fine-mapped QTLs), but we feel that this is out of scope for this study. We hope the reviewer agrees that the above mentioned examples do enough to bring credibility and vividness to the prediction accuracy of variant effects reported in the paper.

Comment 6: Lines 386 – 387 is an overstatement. Moreover, the authors state Pangolin offers the same prediction and therefore the word unique is not appropriate.
a) The RNA-Seq coverage plot shown in Fig 7C shows high relative levels of intron retention – is this reflective of the natural expression pattern in the relevant specimen? Leaky expression of genes in testis often doesn't mirror the clinically relevant specimen. It may be more prudent for the authors to choose a sQTL to highlight that holds clinical relevance to a particular tissue and show RNA-Seq from this tissue, rather than leaky expression in testis.
b) The alternate acceptor activated in testis by the sQTL is annotated. Can the authors confirm Borzoi was not trained on any RNA-Seq or CAGE data for this gene?
c) It also may be helpful for readers trying to evaluate the variants being assess to provide the full HGVS annotated that details the gene, variant and chromosomal position.

We agree that Line 386 was an overstatement, and we have reworded this sentence (which starts on Line 389 in the revised text): "This overall change was correctly predicted by Borzoi.".

a) We felt comfortable visualizing the sQTL example shown in Figure 7C because it is classified as a fine-mapped causal splice-altering variant specifically in Testis, and the particular gene in question has reasonably high TPM in Testis (https://gtexportal.org/home/gene/OBP2A). Finding an sQTL to highlight that holds "clinical relevance" (i.e. that is associated with some phenotype or pathogenicity) is probably very hard since this is not what sQTLs from a healthy cohort such as GTEx is enriched for. We agree that this would be interesting of course, but we feel this is out of scope for this study. We visualize two other, non-Testis, sQTLs in Figure 7F and S7C.

b) First, we again want to emphasize that the model did not see the variant allele in the input sequence during training; only the hg38 reference was used. So in this regard, the variant allele was “held-out” (as are all non-hg38 alleles). To answer the reviewer’s question: the gene shown in Figure 7C (OBP2A) happened by random chance to be part of the held-out test set, so Borzoi was indeed *not* trained on any measurements relating to OBP2A (RNA, CAGE or any other assays).

c) We appreciate the reviewer’s suggestion and will do this if the editor also agrees that we should, however we already provide the hg38 chromosome, position and alleles in the figure panels, as well as the RS identifier of each variant in the captions (which allows users to look up variants in dbSNP and see, among other details, the HGVS annotation. For example: https://www.ncbi.nlm.nih.gov/snp/rs55695858#variant_details). In our experience reading Nature Genetics articles, we often see variants annotated with RS identifiers, but we will change this to HGVS if that is indeed the journal recommendation.

Q7: It is not clear from reading the manuscript how Borzoi can be deployed to screen whole genome sequencing data to prioritise variants with increased likelihood of altering gene expression. It is possible for the authors to provide a plain explanation regarding how to use Borzoi for this purpose? For example, does Borzoi return a predictive score of some kind and there is a threshold score the author’s recommend as a ‘high confidence prediction’? It would also be helpful to have a rough idea how many candidate variants Borzoi identifies from genome sequencing for a heterogenous condition with ~100 associated genes (e.g. epilepsy), whether it identifies 10 or 10,000?

This question is quite complex to answer with a one-fits-all type of analysis, since the particular context or question being studied would reshape the analysis. However, in general, there are a number of different ways we foresee (and recommend) researchers to use this model to score and prioritize genetic variants:

First, given a set of variants of interest and a specific target gene, we recommend prioritizing the variants by scoring their impact on gene expression, polyadenylation and/or splicing based on the three statistics defined in the methods and used in the eQTL, sQTL and paQTL benchmarks (see Line 819, Line 916, and Line 951 of the Methods section). To summarize what’s in the Methods section, the statistics are:

For expression: predicted expression log fold change obtained by aggregating coverage across the exons of the target gene (i.e. $\log(\text{sum of gene coverage})_{alt} - \log(\text{sum of gene coverage})_{ref}$)

For polyadenylation: Maximum difference in predicted coverage ratio between any annotated polyadenylation junction within the gene body. See Methods for the formula.

For splicing: Maximum difference in any of the 32bp coverage bins overlapping the gene body.

One would then rank the variants in descending order of effect size (absolute-valued or signed, depending on analysis). In order to say anything about effect size significance, one would have to construct a matched set of negative/background SNPs for the same target gene that match all the features one would want to control for (alleles, distance to gene, etc.). Again, this would usually depend on the specific study (e.g. is the user studying promoter variants, or 3' UTR variants?). We do not have a general framework or script developed for doing this, but following the workflow of the three QTL analyses presented in the paper would be appropriate in many circumstances. Also note that the analysis can be performed through any of the available RNA-seq tracks; this also depends on the particular context and study, but a good start is to rely on the GTEx tissue RNA-seq tracks.

Second, if a target gene is not known apriori, we recommend computing the gene-agnostic 'L2' variant effect score (which is defined in the Methods section on Line 839). This score computes the L2 norm between the predicted coverage profile (alt vs ref) across the entire 524 kb window. Variants can then be ranked by this statistic.

Finally, for a maximally generic single score, we believe the output of the classifier that we trained to distinguish common from rare variants in the Gnomad database shows promise. We demonstrated that this score is competitive with CADD at classifying held-out variants and will release the classifiers for other researchers to try. In follow-up work, we see that classification performance continues to improve by training nonlinear classifiers like gradient-boosted decision trees on larger numbers of common variants (again using Borzoi outputs as input features to this auxiliary classifier). We plan to explore the application of this score for rare variant analysis in future work.

The scripts and workflows for scoring variants by the various statistics are available in the Github repository, including example notebooks for predicting and interpreting specific variants (for expression, splicing and polyadenylation). To emphasize this point, we have updated the initial sentence of the Data- and Software availability statement (Line 1067), which now reads: "The code repository for training RNA-seq deep learning models, including example code to use the model as well as scripts for variant scoring, is available under the Apache 2.0 open source license at: 'https://github.com/calico/borzoi'."

We have updated the Discussion to summarize the recommended variant scoring use cases detailed above. Specifically, on Line 474, we added the following text:

"For researchers intending to use Borzoi in their genetic variant analyses, we recommend using the *gene-centric* variant effect scores derived in this paper to prioritize variants with respect to a particular target gene. These scores include (1) predicted exon-aggregated coverage log fold-change of the target gene (for abundance differences), (2) predicted maximum difference in coverage log ratio between any 3' cleavage site (for polyadenylation differences), and (3) predicted maximum normalized difference in any coverage bin within the gene body (for splicing differences). If a target gene is not known apriori, we recommend using a *gene-agnostic*

statistic, such as the one based on total L2-norm, to quantify potential changes in coverage patterns across the entire output window."

As for the final question posed by the reviewer, we do not have any results yet on applying Borzoi to large clinical WGS data or GWAS studies where particular disease phenotypes are investigated, and we therefore cannot answer this question. In many clinical scenarios we can imagine, the researcher will have to decide on a tuneable threshold of top scoring variants/genes to follow up on. Importantly, Borzoi has zero knowledge of gene function and will produce similar scores for similar gene expression effects on genes that code for essential proteins or irrelevant ones. This creates an intriguing opportunity to combine Borzoi with protein annotations in creative ways. We are actively working on building statistical fine-mapping approaches based on Borzoi variant effect predictions (and for rare variant analysis), but this is out of scope for the present study and constitutes future research.

Comment 8: As a final remark, the title is not intuitively understandable for me and therefore may not be for other Nature Genetics readers. Though the decision re title wording rests with the authors. The phrase 'RNA-Seq coverage' is inherently related to RNA-Seq read-depth which can vary from 20M reads to 1 billion transcriptomic reads and is not entirely the same thing as relative transcript expression levels.

We thank the reviewer for pointing this out. We will work with the Nature Genetics editorial team to figure out the most appropriate title, but a possible alternative would be:
"Predicting RNA-seq profiles from DNA sequence as a unifying model of gene regulation"

Response to Referees - Borzoi Manuscript, Nature Genetics (NG-A63358R1)

Reviewers' Comments:

(Author responses = blue)

Reviewer #3:

The authors have addressed my remaining comments.

We thank the reviewer for their valuable feedback.

I had surmised all data sets must have been held out from the Training Set - but could not be certain. Equally important is specifying the donor and PolyA variants modelled in Fig1 are synthetic. This was not obvious. Being synthetic variants, and with no way of proving if the Borzoi prediction is accurate, I agree these examples hold little utility for presentation in Figure 1. A better option would be to find a published article describing RNA assay data for pathogenic splicing variant causing a Mendelian disorder in a gene within your Testing Set.

We agree with the reviewer, it was previously not obvious that the examples in the attention matrix analysis were synthetic. We have updated the text describing this analysis, which now resides as Section 1.2 in the Supplementary Note, emphasizing that the variants visualized are hypothetical and not measured.

Furthermore, while it would be interesting to visualize pathogenic variants known to cause Mendelian disorders, we believe that the current examples of fine-mapped QTLs shown in Figure 5-7 and Extended Data Figure 6-9 provide sufficient credibility to the unbiased QTL benchmark results reported in the paper. We also want to emphasize that these examples are visualized along with actual RNA-seq measurements. We hope that the reviewer agrees with this, and we look forward to pursuing variant interpretations of other data in future work.

I leave the decisions about which figures to go where to the authors and editorial team.

Inclusion of overview statements related to which scores to assess for different classes of variant is also welcomed. The best outcome is if deep learning tools can be harnessed by both fundamental and clinical researchers alike - with the latter needing user-friendly interfaces and simplified guidelines how to use it, including author recommendations regarding what constitutes a high confidence score from a low confidence score.

We agree with the reviewer. In addition to the overview statements of variant effect scores that we previously added to the Discussion, we have included computational notebooks of example variant interpretations and tutorials on how to run the variant scoring scripts on the Github page.

These examples and tutorials are documented at: https://github.com/calico/borzoi