nature portfolio

Peer Review File

Annotating the genome at single-nucleotide resolution with DNA foundation models

Corresponding Author: Dr Thomas Pierrot

Version 0:

Decision Letter:

27th Jun 2024

Dear Dr Pierrot,

Your Article, "SegmentNT: annotating the genome at single-nucleotide resolution with DNA foundation models", has now been seen by 4 reviewers (Reviewers 1 and 2 co-reviewed the paper). As you will see from their comments below, although the reviewers find your work of potential interest, they have raised a number of very important concerns. We are interested in the possibility of publishing your paper in Nature Methods, but would like to consider your response to these concerns before we reach a final decision on publication.

We therefore invite you to revise your manuscript, and we are not committed to sending the paper back to our reviewers unless the paper is extensively revised with all their concerns fully addressed, including, among other revisions, making a strong case for the benefits of the model for advancing the state-of-the-art and addressing key challenges.

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 paper:

* include a point-by-point response to the reviewers and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

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 paper within 6 months. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS

When revising your manuscript, please update your reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

DATA AVAILABILITY
We strongly encourage you to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific and community-recognized repositories; a list of repositories is provided here: http://www.nature.com/sdata/policies/repositories

All novel DNA and RNA sequencing data, protein sequences, genetic polymorphisms, linked genotype and phenotype data, gene expression data, macromolecular structures, and proteomics data must be deposited in a publicly accessible database, and accession codes and associated hyperlinks must be provided in the "Data Availability" section.

Refer to our data policies here: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-data

To further increase transparency, we encourage you to provide, in tabular form, the data underlying the graphical representations used in your figures. This is in addition to our data-deposition policy for specific types of experiments and large datasets. For readers, the source data will be made accessible directly from the figure legend. Spreadsheets can be submitted in .xls, .xlsx or .csv formats. Only one (1) file per figure is permitted: thus if there is a multi-paneled figure the source data for each panel should be clearly labeled in the csv/Excel file; alternately the data for a figure can be included in multiple, clearly labeled sheets in an Excel file. File sizes of up to 30 MB are permitted. When submitting source data files with your manuscript please select the Source Data file type and use the Title field in the File Description tab to indicate which figure the source data pertains to.

Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY
Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see:
https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

MATERIALS AVAILABILITY
As a condition of publication in Nature Methods, authors are required to make unique materials promptly available to others without undue qualifications.

Authors reporting new chemical compounds must provide chemical structure, synthesis and characterization details. Authors reporting mutant strains and cell lines are strongly encouraged to use established public repositories.

More details about our materials availability policy can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials

ORCID

Nature Methods 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. This applies to primary research papers only. 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.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to consider your work.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewers' Comments:

Editor: Reviewers 1 and 2 co-reviewed the paper.

Reviewer #1 (Remarks to the Author):

This manuscript introduces SegmentNT, a deep learning model that combines a pre-trained DNA foundation model (Nucleotide Transformer) with a segmentation architecture (U-Net) to predict various types of genomic elements at single-nucleotide resolution in DNA sequences. SegmentNT can predict the base-resolution positions of 14 elements, including genes, regulatory regions, and splice sites, and explores extensions to expand the input sequence length up to 50kb. The authors demonstrate the value of using pre-trained DNA foundation models and techniques to extend the model's context length. They also developed a multispecies version of SegmentNT that can generalize to annotate genomic elements in other animal and plant species beyond humans. While the innovation and effectiveness of SegmentNT are nicely demonstrated, there are several major and minor concerns:

Major concerns:

• There is a lack of discussion about traditional genome annotation tools and how SegmentNT advances the field. For instance, traditional genome annotation approaches can be de novo, alignment-based (CACTUS), or combinations of the two (eg. Comparative Annotation Toolkit, CAT). They can also depend on functional assays, such as chromhmm. Placing this work's contribution within the broader field is important for people to understand how Segment NT is advancing upon previous methods.

• SegmentNT’s performance for splice donor and acceptor in Fig 1e appears the highest. However, the ablations show that UNet performs poorly when using one-hot sequences. UNet was never established as a meaningful model for gene annotations (even though it makes sense); UNet is clearly beneficial when using NT representations as input. On the other hand, SpliceAI, which is not a UNet, but takes one-hot sequences performs much better on predicting splice annotations. Perhaps a better supervised baseline would be to take spliceAI and predict all 14 tasks.

• SpliceAI-10k's performance in the original study was 0.98 PR-AUC. In this paper, it is 0.96 for acceptor and 0.94 for donor. Why is there a discrepancy?

• I would not draw the same conclusion as "These results show that the segmentation capabilities of SegmentNT can be used to predict complex gene rearrangements directly from the sequence, which should be a useful tool for the interpretation of sequence and structural variants that can affect gene regulation and disease." from a PCC of 0.24. I think further evidence is needed to support this claim.

• For the MST1R dataset testing isoform change prediction, SpliceAI should be included in the comparison as a relevant baseline.

• For the Enformer evaluation, 7 ATAC-seq profiles were selected, but it’s unclear which ones. Would it make more sense to use Enformer's representations and place a UNet over that as a direct head-to-head comparison with NT embeddings, especially since Enformer's bin size is not base resolution and the promoter prediction task demands it? On that note, how are edge effects handled due to binned predictions versus base-res labels?

• SegmentNT’s predictions are given at base resolution. I suppose the annotations with hard boundaries make sense for gene features, such as exons, but setting hard boundaries for regulatory elements is a bit more arbitrary. The predictions, at least for regulatory elements, seem to resemble broad distributions. So, are the performance gaps coming from edge effects of annotations or spurious predictions in random locations? How would a Gaussian filter applied to the CRE annotations

perform? Perhaps if the variance of the filter were systematically varied, this could help calibrate how much the metrics are influenced by poorly predicted edge effects.

• The performance is summarized for each element across all loci. It could also be informative to look at performance averaged across annotation tasks within loci. This way, one can monitor performance within each loci and then further investigate the loci that are poorly predicted to understand SegmentNT's limitations.

• Although it may be out of scope for the paper, it would be interesting to see if the multi-tasking nature helps or hurts performance. Perhaps task ablations?

• The generalization across species should be benchmarked against existing state-of-the-art gene finder tools (if possible).

Minor concerns:

• How is the probability of belonging to each of the 14 genomic elements at the nucleotide level calculated?

• There is a data leakage possibility for gene elements due to orthologs and gene duplications.

• It would be nice to see performance under different metrics, not just MCC, such as AUPR and AUROC.

• The pictures in Figs 5a and 5b are too small.

Reviewer #2 (Remarks to the Author):

This manuscript introduces SegmentNT, a deep learning model that combines a pre-trained DNA foundation model (Nucleotide Transformer) with a segmentation architecture (U-Net) to predict various types of genomic elements at single-nucleotide resolution in DNA sequences. SegmentNT can predict the base-resolution positions of 14 elements, including genes, regulatory regions, and splice sites, and explores extensions to expand the input sequence length up to 50kb. The authors demonstrate the value of using pre-trained DNA foundation models and techniques to extend the model's context length. They also developed a multispecies version of SegmentNT that can generalize to annotate genomic elements in other animal and plant species beyond humans. While the innovation and effectiveness of SegmentNT are nicely demonstrated, there are several major and minor concerns:

Major concerns:

• There is a lack of discussion about traditional genome annotation tools and how SegmentNT advances the field. For instance, traditional genome annotation approaches can be de novo, alignment-based (CACTUS), or combinations of the two (eg. Comparative Annotation Toolkit, CAT). They can also depend on functional assays, such as chromhmm. Placing this work's contribution within the broader field is important for people to understand how Segment NT is advancing upon previous methods.

• SegmentNT's performance for splice donor and acceptor in Fig 1e appears the highest. However, the ablations show that UNet performs poorly when using one-hot sequences. UNet was never established as a meaningful model for gene annotations (even though it makes sense); UNet is clearly beneficial when using NT representations as input. On the other hand, SpliceAI, which is not a UNet, but takes one-hot sequences performs much better on predicting splice annotations. Perhaps a better supervised baseline would be to take spliceAI and predict all 14 tasks.

• SpliceAI-10k's performance in the original study was 0.98 PR-AUC. In this paper, it is 0.96 for acceptor and 0.94 for donor. Why is there a discrepancy?

• I would not draw the same conclusion as "These results show that the segmentation capabilities of SegmentNT can be used to predict complex gene rearrangements directly from the sequence, which should be a useful tool for the interpretation of sequence and structural variants that can affect gene regulation and disease." from a PCC of 0.24. I think further evidence is needed to support this claim.

• For the MST1R dataset testing isoform change prediction, SpliceAI should be included in the comparison as a relevant baseline.

• For the Enformer evaluation, 7 ATAC-seq profiles were selected, but it's unclear which ones. Would it make more sense to use Enformer's representations and place a UNet over that as a direct head-to-head comparison with NT embeddings, especially since Enformer's bin size is not base resolution and the promoter prediction task demands it? On that note, how are edge effects handled due to binned predictions versus base-res labels?

• SegmentNT's predictions are given at base resolution. I suppose the annotations with hard boundaries make sense for gene features, such as exons, but setting hard boundaries for regulatory elements is a bit more arbitrary. The predictions, at least for regulatory elements, seem to resemble broad distributions. So, are the performance gaps coming from edge effects of

annotations or spurious predictions in random locations? How would a Gaussian filter applied to the CRE annotations perform? Perhaps if the variance of the filter were systematically varied, this could help calibrate how much the metrics are influenced by poorly predicted edge effects.

• The performance is summarized for each element across all loci. It could also be informative to look at performance averaged across annotation tasks within loci. This way, one can monitor performance within each loci and then further investigate the loci that are poorly predicted to understand SegmentNT's limitations.

• Although it may be out of scope for the paper, it would be interesting to see if the multi-tasking nature helps or hurts performance. Perhaps task ablations?

• The generalization across species should be benchmarked against existing state-of-the-art gene finder tools (if possible).

Minor concerns:

• How is the probability of belonging to each of the 14 genomic elements at the nucleotide level calculated?

• There is a data leakage possibility for gene elements due to orthologs and gene duplications.

• It would be nice to see performance under different metrics, not just MCC, such as AUPR and AUROC.

• The pictures in Figs 5a and 5b are too small.

Reviewer #2 (Remarks on code availability):

Code is accessible and easy to use. README file provides good tutorial and information to reproduce the models.

Reviewer #3 (Remarks to the Author):

Summary
—----------
This manuscript introduces SegmentNT, a deep learning model that leverages a pre-trained DNA language model (Nucleotide Transformer, NT) and a 1D U-Net architecture to predict the location of various genomic elements within DNA sequences at single-nucleotide resolution. The authors demonstrate SegmentNT's ability to predict the location of both gene elements (exons, introns, splice sites, etc.) and regulatory elements (promoters, enhancers) in the human genome, claiming superior performance compared to baseline models and generalization capabilities to other species. They further highlight the model's efficiency and its potential for analyzing sequence variants and their impact on transcript isoforms.

General Remarks
—----------
This manuscript presents SegmentNT, a deep learning model for genome annotation that leverages a pre-trained DNA language model. While the exploration of this approach is timely given the increasing interest in foundation models for genomics, the manuscript currently falls short of demonstrating its practical value and significance to the broader biological community. The authors primarily focus on showcasing the technical feasibility of using a DNA language model for annotation, but lack a clear articulation of how SegmentNT addresses the limitations of existing methods and advances the state-of-the-art in genome annotation. The benchmarking strategy is currently relying on artificially constrained comparisons, a misuse of existing tools, and could focus more on known tools and relevant specialized methods. We are left with a premature set of investigations, whose documentation has perhaps the merit to attract the attention to a traditional computational biology area that will benefit from LLMs, but with the drawback that it does not advance the state-of-the-art and will confuse further newcomers to computational biology coming from other areas of machine learning by providing a set of flawed benchmarks and tasks. Overall, the study requires a clear focus to establish the practical utility of the approach, to provide more rigorous comparative analyses, and to clearly articulate its contribution to addressing key challenges regarding genomics research.

Major Points
—----------
1. The motivation as currently presented is too limited to be of broad interest and this is reflected in many analyses presented by the author. Currently, the main aim of the paper seems to be to prove that DNA-LMs can „also“ do genome annotation. This is interesting from a methodological perspective. However, to be impactful, the work must be placed into the context of the general literature on genome annotation, clearly identify where current tools are lacking and explain how a DNA-LM based approach would improve upon state-of-the-art problems of interest to biologists. Specifically, the study entirely ignores how genome annotation (see pipelines e.g. used by the Ensembl annotation team) is performed. Those pipelines and the tools they build upon have a single-nucleotide resolution. Ignoring this literature is concerning, starting from the fact that Segment-NT takes the output of these pipelines as ground truth. The limitations of these tools and pipelines are not spelled out.

2. Related to this, the manuscript lacks comprehensive benchmarking against established gene annotation tools. To demonstrate SegmentNT's practical value, the authors should compare its predictions to established gene annotation

pipelines such as for example AUGUSTUS or MAKER, using standard evaluation metrics and datasets for gene prediction accuracy. Additionally, while tools like Deepromoter are mentioned in the introduction, a comparison against these specialized methods is missing.

3. Statistical significance testing is largely absent. The authors should perform appropriate statistical tests to determine whether the reported performance differences between models are statistically significant.
4. Currently, SegmentNT is benchmarked against ablations or competing methods that are artificially constrained. SegmentNT is benchmarked against a version that is randomly initialized. This comparison does not account for the amount of compute spent on pretraining the Nucleotide Transformer: “the version with random initialized NT showed much slower convergence and have not converged yet even after 68M training sequences (34B tokens), a training more than three times longer”. The authors should train the model until convergence if they really want to make a fair point about the effect of pre-trained DNA language models.
5. When comparing the supervised Enformer to SegmentNT, the comparison would only be fair if both had the same U-Net-head on top of their embeddings. Currently, the authors use the prediction over 7 selected ATAC-seq (or DNase?) profiles as promoter-proxy, neglecting that many regions in the genome can be accessible, including enhancers, transposable elements and potentially parts of coding sequences, and thus accessibility is not an exclusive mark for promoters. Encode, for example, categorizes promoters as „Elements (cCREs) putatively labeled as promoters defined by a signature of high DNase and H3K4me3 signal and that lie within 200 bp of an annotated GENCODE transcription start site (TSS).“ which indicates that accessibility alone is insufficient.

6. Related to the previous point, for regulatory elements such as promoters/enhancers, it is somewhat unclear whether there is an added value to predicting a label such as „enhancer“ compared to predicting biochemical marks such as accessibility, histone acetylation or transcription factor binding, as is done by Enformer. In the end, the biochemical marks represent the biological mechanism whereas annotations such as „enhancer“ reflect a human attempt to systematize these mechanisms. The existence of enhancer RNAs, for example, implies that the distinction between enhancer and promoter is not always clearly definable. As such, it is not obvious whether a model that predicts human labeling of regulatory elements really can provide a benefit over a model such as Enformer that directly predicts biochemical tracks, particularly for applications in human genetics, where ENCODE already is near exhaustive. Potentially the authors could address this by showing that their labeling of regulatory elements provides some benefit in the cross-species setting, e.g. by using it to predict when the orthologues of enhancers cease to be functional [Kaplow, et al. Science 2023].

7. The application of Enfomer is flawed. The authors use a reimplementation of it. There is no data showing that the reimplantation is correct (how much do the predictions differ). It is referred to a non-reviewed preprint which does not show this either. More critically, Enformer takes 200kb as input. The authors run Enformer with only 10kb padding the rest of the sequence. It is therefore very misleading to label such predictions as “Enformer” predictions.

8. Prometre and enhancer results are cherry-picked. The authors apply their approach to “the two best classes of regulatory elements”. Promoter tissue-invariant and enhancer tissue-specific (Fig 2a,b). Comprehensive results should be shown.

9. The comparison between SegmentNT and SpliceAI is conceptually flawed. SpliceAI is developed to predict splice sites within precursor RNAs as clearly stated in the original publication. Genome annotation is not done by running SpliceAI on the plus strand of a genome. Nonetheless, the authors apply SpliceAI outside transcribed regions and on the reverse strand of genes. The authors then claim the superiority of their tool specifically on genes encoded in the minus strand (Figure 4b). This is extremely misleading.

10. For splice site prediction within precursor RNAs, Pangolin should be compared as well as it is multi-species and claims to outperform SpliceAI.

11. For splice site prediction, the performance of SegmentNT on non-coding RNA splice sites should be shown as some of the performance could be driven by some correlative signal from coding sequence.

12. Any single metrics summarizing confusion tables has issues and MCC is not immune to this. Moreover, confusion tables depend on the cutoffs applied. Precision-recall curves in case of highly imbalanced classes, or ROC curves otherwise, should be shown to assess whether methods are uniformly better than others or only at some cutoffs. The 0.5 cutoff should be indicated on those curves for SegmentNT and the recommended cutoffs of other tools (SpliceAI has three recommended levels). For very imbalanced tasks, we recommend reporting the top-k accuracy, where k is the actual number of positives, see the SpliceAI publication and reference therein.

13. Regarding the mutagenesis screen for splicing, the authors should compare to other models including Pangolin, SpliceAI and MMSplice - the latter having been developed for this.

14. In the multi-species scenario: since genes can be shared across species by homology and species have different numbers of chromosomes, the strategy of helding out chromosomes is subject to data leakage. The authors should assess this issue discuss and implement a strategy to circumvent it. This problem can particularly affect models like the one proposed which has a very high number of parameters in striking contrast to state-of-the-art genome annotation tools (generalized HMMs).
15. The manuscript claims that SegmentNT species generalization improves when trained on multiple species. However, the base model, NTv2, is in both cases already a multispecies (self-supervised) model and should have captured the semantics of different species. DNA language models have previously been shown to capture regulatory motifs across evolution and not

only on one target species [Nguyen et al,bioRxiv (2024); Karollus et al Genom Biol 2024]. It should be explored how much the SegmentNT U-Net is overfit to single species, if NTv2 has not learned these species well enough to contribute to genomic annotation tasks, or if worse performance is mostly an artifact of worse genome annotations for distant species.

Minor Points
—----------
16. The name of the tools is a misnomer. Segment-NT is not a segmentation algorithm. It is a single-base multitask classification. Proposing a segmentation algorithm could have been methodologically innovative and may be a safer direction at this stage that the authors could investigate.

17. In Fig. 1a, the pictogram incorrectly displays 3-mers as inputs instead of nonoverlapping 6mers.

18. What is the purpose of the U-Net architecture? As described in Fig. 1b, one could just go from (L, 1024) to (L, 6*14) directly. In contrast to Borzoi, which maps from single base-pair input to 32bp bins, the base-model NTv2 should have already learned to mix along the sequence dimension during pretraining, so that pooling in the U-Net might no longer be necessary. An explanation would be interesting for readers. Furthermore, is the NTv2 variant for the human SegmentNT the best option, or would that be the human-only NTv1?

19. The emphasis on the speedup (Fig. 2C) is not motivated in the text. It is indeed unclear if inference speed is a concern for genomic annotation models that mostly will be run once per genome.

20. The manuscript could benefit from the inclusion of negative controls to assess the robustness of the model's predictions. Several recent studies analyzed how foreign DNA is interpreted by human cells. If SegmentNT correctly annotated this foreign DNA, it would provide compelling evidence that the model is using causal principles of genome architecture to make its predictions.

21. p. 10 and in Discussion, it is argued that it is encouraging that the multispecies model still retained predictive performance for species whose genome underwent large-scale rearrangements. The authors should clarify their argumentation and explain why would ploidy and large-scale rearrangements affect the genomic code (eg, have to do with predicting splice sites, etc)?

22. In Discussion, it is stated ““Here we focused on a more challenging task of segmenting genomic elements in DNA sequences at nucleotide resolution.” The authors should explain segmenting genomic elements more challenging than other (not named) tasks? In fact, computational biology have developped tools to annotate genomes way before gene expression models could be developed.

23. The Discussion should refer to concurrent approaches to transcript prediction based on experimental data (RNA-seq coverage): borzoi, bigRNA. If these methods get peer-reviewed and published they should be part of the benchmark.

Typos
—-------
24. Everywhere: finetuning -> fine-tuning.

25. Figure 1b: U-Net in all caps

26. Figure S4b,c do these constructs have indeed multiple mutations? Was the model applied to a set of mutations jointly?

27. Figure S4c the authors should denote whether the mutations correspond to a branch point mutation and donor and acceptor site mutation. These are fine examples but already well-modeled since decades.

28. Figure 5a, “zero-shot generalization” is misused. The classes are the same than during training (exons, splice sites, etc.). Zero-shot would mean to predict an entire new class (say enhancer by only having learnt from promoters or exons.). The crux of zero-shot learning is to use some auxiliary information about the class [https://arxiv.org/abs/1707.00600] and this is not the approach undertaken here. Generalization to unseen species should be used instead.

29. Discussion: “... of segmenting genom*i*c elements”

Reviewer #3 (Remarks on code availability):

I did not review the code at this stage given the major concerns overall. I do appreciate that code was available.

Reviewer #4 (Remarks to the Author):

Overall, it is very nice to see a new, important application of pre-trained language models in genomics (and not just more benchmarks). Comparisons with spliceAI and generalization to other species are very interesting. Commend the authors for including randomly initialized baselines, but several of the comparisons need to be done more carefully (e.g., fine-tune the head only).

Although not related to the work here, thanks to some of the authors for the mRNA vaccine. It's an honor to review a paper from some of the people who helped to protect us from covid-19.

Thanks very much for your patience with the slow review turnaround. This was a very interesting paper with many areas for improvement so it took some time to think about it.

Essential

1. The authors need to explain more clearly how they are evaluating segmentations. Naively, MCC is not a great measure for segmentation because it's not comparable between segments of different lengths and treats each nucleotide as a separate prediction. (E.g., if half of the sequence is gene and the other half is non-coding, it's much easier to get a high MCC if there is only one changepoint than if there are 10 change points. Even more complicated issues if it's not half and half) If the authors have computed MCC using each nucleotide as a separate prediction (not explained in the text, but that's what I guess), they should present at least one other metric that is at the "region" level. Examples of this kind of metric are intersection over union (as in the original Unet paper), Dice coefficients, or (probably nicer to stick with bioinformatics) the classical Sov score that was developed for protein secondary structure segmentation (PMID: 10022357)

2. Where are the error bars? Not a single performance measure is presented with confidence intervals/estimates of uncertainty. Given the variation in sample sizes for the different types of elements, estimates of performance for some elements must be more confident than others. As in any scientific study, all reported measurements should include estimates of uncertainty. I realize that it is commonplace in the computer science conferences and online machine learning competitions to achieve "state-of-the-art" by simply reporting a better number than the other methods, but in a serious peer-reviewed study, I believe we need to see estimates of uncertainty. (And connected to point 1, the error bars/uncertainty estimates cannot be based on the assumption that each nucleotide is independent. They must be based on the number of regions of each type. That is the real "sample size".)

3. The test of the segmentation method on species that are distant from the training data is fantastic, and in my view the most interesting part of the paper. However, the authors should clarify where the "ground truth" is coming from for these genome annotations. For well-studied model systems like arabidopsis, c elegans, e coli or drosophila will have high quality genome annotations that are based on numerous data sources and expert curations. If these are not included in the training data and are evolutionarily distant from the species in the trainin data, this is a true test of out of sample generalization. On the other hand, most other species are studied by fewer molecular biologists, and will only have annotations that are mostly based on bioinformatics predictions. These should not be treated as "ground truth." In these situations we are only comparing the nt-transformer segmentation to classical bioinformatics aprpoaches.

4. The comparison to the Unet alone vs. the embedding + Unet is not a fair comparison because of the difference in tokenization. As far as we understand, the Unet alone is on a 4-word vocab, while the embedding + Unet is on a 4096 word vocab. Therefore, as far as we understand it, the number of parameters and the dimensionality of the models in not comparable. The authors need to redo this key control experiment where the UNet is trained on a 4096 word vocab.

5. The random initialization is a great negative control, but it's not a fair baseline if they didn't let it converge. In addition, they need to 1) use "frozen" embeddings from nt-transformer and only train the segmentation head and 2) use a randomly initialized, but frozen embedder (untrained) and train the segmentation head. The performance differences on these experiments will show the importance of the pre-training task.

6. many overstated claims:
-The authors claims about 0-shot and few-shot learning are misleading. They use 0-shot and few-shot to mean generalization out of species or out of context length, which are both impressive, but are not 0- or few-shot in the sense that the term is used in the field. We believe that 0- and few- shot should only be used to describe performance on a task the model was never trained to do. Since they have fine-tuned the parameters of the embedder on the segmentation task, there is no sense in which this is 0-shot learning (as far as we can understand).

-“We show improved performance on the detection of splice sites throughout the genome and demonstrate strong nucleotide-level precision”

The generality of this claim is not supported by the data presented. The performance of the model on detecting splice sites is similar to SpliceAI for protein-coding genes

-“We show that SegmentNT achieves high nucleotide accuracy for all elements” Not convincingly true, some elements like lncRNA are still very poorly segmented.

-“SegmentNT accurately detects splice donor and acceptor sites in both strands in any given input DNA sequence” Is not supported by the results. The performance of SegmentNT is high but not that high.

- “a new paradigm on how we analyze and interpret DNA” seems overstated. The authors should restrict themselves to claims that can support with evidence.

7. The attention to detail of the manuscript is at times poor, with many errors which inhibited understanding. These are too numerous to fully account for here, however the below list details some:
• In the abstract the authors claim SegmentNT can “generalize to elements of different human and plant species” which doesn't

make sense because there is only one human species.
• Supplementary Fig 3 contains the exact same data as Fig 4b and 4c, despite the text indicating it should contain different values.
• The text refers to Supplementary Fig. 4d-g, however figures 4e, 4f, and 4g do not exist
• Figure 5a caption indicates the held-out test set contains 12 species, however the figure indicates there are actually 17 species. Also misspells “species.”
• Citations 28, 44, 53, 54, and 57 are missing key information. Some citations in the text not correctly ordered when within the same brackets.
• Section A.5 has details which conflict with those elsewhere (different model context length, does not indicate human annotations are included in the finetuning)
• First sentence of Section D.2 refers to the methods section despite being in the methods section
• The link to the Illumina Basespace platform cannot be accessed
• Y-axis of Supplementary figure 1b is incorrect
• Question marks in negative region of axes in Supplementary figure 4b) and 4c)
• Various instances of minor inconsistencies in wording across the manuscript. Approach sometimes referred to “SegmentNT” and other times as “Segment-NT”, x-axis and y-axis titles between adjacent plots changed despite showing the same metrics. Models given names not found elsewhere in first sentence of “SegmentNT generalizes to sequences up to 50kb” section.
• Many minor grammatical and formatting errors (e.x. Tetraodon spelt incorrectly in methods section, Time tree of life links not working, etc.)
A thorough re-reading of the manuscript must be performed to eliminate these errors.

Suggestions.
1. In my view the comparison with enformer is problematic. As the authors show, using binary models is not nearly as good as training the segmentation head (Unet) on top. Since Enformer was not trained on a segmentation task, to make a fair comparison the authors should use the bp-level predictions of all the tracks from enformer as a pre-trained representation and then train a unet segmentation head on top. This could also be evaluated on the out-of-sample "out-of-species" test. Otherwise, the authors should drop this analysis or at least acknowledge that it is not really a fair comparison. My sense is that the spliceAI results are strong enough to make the case for the fine-tuned segmentation model. I don't feel that the enformer results (especially as presented) are necessary.

2. If I am reading the paper correctly, the authors did not use the "largest" of the nt transformer models (in terms of parameters and training data). Seems like they only used a 500M multispecies (this means the model has seen some of the other species genomes? so it is not really held out data?). This should be discussed/explained, especially in the context of the current thinking around "scaling" of "foundation models." If using larger models does not improve performance, this is an important (negative) result about the scaling of current MLM pre-training. If the larger models are just too "big" to use in practice, that's also an important result on the practical side for the computational genomics community.

Minor essential issues.

1. The have provided a Method that can segment a genome. However, code they provided for segmenting the genome in their README.md does not run because of a syntax error (we found the bug and were able to correct, but this is not really our job).

2. We cannot replicate the paper's findings because the methods are not described clearly enough (and no codes are provided).

-the output shape in the figure (L x 6 x 14) contradicts what they write in the methods section A.1 (N x K x 2) which we believe is N = 6 x L and K = 14 but there is an extra factor of 2.

-For human the test and validation set chromosomes are indicated, but not for the other 5 species used to train the multispecies model. This is necessary information for replicating the results.

-we cannot understand the experiments where they compared to their binary promoter classifier or to the enformer enhancer predictor. There are so many issues that we can't follow, we will not be able to list them all here.

e.g., "We note that for Enformer we had to pad the
10kb sequences for inference, since the original model predicts scores for 114,688bp."
To our knowledge Enformer takes as input 196,608. We've been padding our sequences to that size. Can you please clarify what you are actually padding your sequences to.

e.g., choice of 10 nt window is arbitrary and unjustified.

-there are also problems with the comparison to spliceAI

E.g., the numbers in the text do not match the numbers in supplementary figure 3.

e.g., It's hard to understand why they did the experiment of limiting to 10Kb test sequences.

-Figure 1a has a labeled detection threshold, however such thresholds are not referred to elsewhere in the manuscript. The threshold needs to be defined in the manuscript or it cannot be reproduced.

-inconsistent use of MCC and PR-AUC in different analyses. We wonder if they are hiding results? OR the paper was done in stages and not well organized. The authors should be consistent.

Version 1:

Decision Letter:

4th Mar 2025

Dear Dr Pierrot,

Your Article, "Annotating the genome at single-nucleotide resolution with DNA foundation models", has now been seen again by 3 reviewers. As you will see from their comments below, our reviewers still raise a number of concerns. We are interested in the possibility of publishing your paper in Nature Methods, but would like to consider your response to these concerns before we reach a final decision on publication.

We therefore invite you to revise your manuscript to full address all these remaining concerns.

After we shared with them your update on code availability, Reviewer 4 has updated their comments on code, which can be found in their remarks on code availability (please see below).

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 paper:

* include a point-by-point response to the reviewers and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

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 paper within 2 months. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please update your reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html

or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

EXTENDED DATA FIGURES
When re-submitting your manuscript, please ensure that any supplementary figures and tables that are crucial to the manuscript's conclusions are converted into Extended Data figures and tables to increase visibility of these data. Extended Data figures and tables are online-only (present in the online PDF and full-text HTML versions of the paper), peer-reviewed display items that provide essential background to the article but are not included in the main article due to space constraints. A maximum of ten Extended Data display items (figures and tables) is permitted.

DATA AVAILABILITY
We strongly encourage you to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific and community-recognized repositories; a list of repositories is provided here: http://www.nature.com/sdata/policies/repositories

All novel DNA and RNA sequencing data, protein sequences, genetic polymorphisms, linked genotype and phenotype data, gene expression data, macromolecular structures, and proteomics data must be deposited in a publicly accessible database, and accession codes and associated hyperlinks must be provided in the "Data Availability" section.

Refer to our data policies here: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-data

To further increase transparency, we encourage you to provide, in tabular form, the data underlying the graphical representations used in your figures. This is in addition to our data-deposition policy for specific types of experiments and large datasets. For readers, the source data will be made accessible directly from the figure legend. Spreadsheets can be submitted in .xls, .xlsx or .csv formats. Only one (1) file per figure is permitted: thus if there is a multi-paneled figure the source data for each panel should be clearly labeled in the csv/Excel file; alternately the data for a figure can be included in multiple, clearly labeled sheets in an Excel file. File sizes of up to 30 MB are permitted. When submitting source data files with your manuscript please select the Source Data file type and use the Title field in the File Description tab to indicate which figure the source data pertains to.

Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY
Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

MATERIALS AVAILABILITY
As a condition of publication in Nature Methods, authors are required to make unique materials promptly available to others without undue qualifications.

Authors reporting new chemical compounds must provide chemical structure, synthesis and characterization details. Authors reporting mutant strains and cell lines are strongly encouraged to use established public repositories.

More details about our materials availability policy can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials

ORCID

Nature Methods 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. This applies to primary research papers only. 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.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to consider your work.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

The authors have addressed all my concerns thoroughly. The revisions, including expanded benchmarking across relevant tasks, stronger baselines incorporating state-of-the-art tools, additional supervised embedding methods, improved data splits to prevent data leakage, and better contextualization within the field, have significantly enhanced the rigor and clarity of the manuscript. I believe it now meets a very high standard and will undoubtedly be impactful for the broader field.

Reviewer #1 (Remarks on code availability):

I checked to see the organization of the repo and found it to appear quite complete with thorough documentation, though I did not attempt to install and run it.

Reviewer #3 (Remarks to the Author):

General comments
The revision is strong and has addressed most of the concerns. The following points shall still be addressed. We are reviewer 3. We first pick up on a point by reviewers 1 and 2.

Reviewers 1 and 2:
SpliceAI-10k's performance in the original study was 0.98 PR-AUC. In this paper, it is 0.96 for acceptor and 0.94 for donor. Why is there a discrepancy?

Response: This discrepancy is due to small adaptations in the dataset to allow a direct comparison with SegmentNT settings. Namely, we consider sequence windows of 30kb (instead of 10kb in the original publication) and compute the predictions on the whole window (instead of not predicting for the flanking +/- 5kb sequence). We have also removed sequences with Ns due to the constraint of the SegmentNT architecture. We have added these details in the Methods section (see page 26).

Reviewer 3: The authors may wish to modify SpliceAI or its usage to better match characteristics of their model (input and output size). However, reporting the performance of a tool when, instead, a modified version of the tool was applied is very misleading. The manuscript may include this modified SpliceAI but should also report the original SpliceAI performance. Relatedly, the new Fig. 5a shows Pangolin and SpliceAI predictions outside of transcript boundaries. This is misleading as this is outside of the application domain of these tools. As mentioned in another point, the authors want to show that predicting splice sites genome-wide is hard because best-in-class tools have been developed for predicting splice sites within transcript boundaries. Therefore, these tools are suboptimal for de novo annotation of genomes for which transcribed regions are not experimentally determined or accurately predicted. This point is valid but confusing because Pangolin and SpliceAI false positives outside their application domain are conceptually different than those found within. To avoid confusion, the authors could highlight in Fig 5a the gene boundary (e.g. with a grey background outside the transcribed region) and add to the legend that SpliceAI and Pangolin have not been trained for out-of-transcript-boundary applications and make further false positives when applied there.

Reviewer 3 (this reviewer):
3. Statistical significance testing is largely absent. The authors should perform appropriate statistical tests to determine whether the reported performance differences between models are statistically significant.

Response: We agree with this reviewer and have now added error bars and confidence intervals to all models and metric comparisons [....]. These results provide stronger support to our conclusions, many thanks!

Reviewer 3: Statistical assessment was more often performed in this revision. However, the authors shall i) take reviewer 4's point 2 into account about bases within regions (e.g. in one annotated exon, one annotated intron, etc.) problematically not being i.i.d. and ii) perform statistical assessments systematically. There are missing statistical assessments in new Fig. 4, new Fig. 5, and new Fig. 7i. Also, the Statistics section of the reporting summary shall be revised. Statistical testing, sample size reporting, etc. are not "not applicable".

13. Regarding the mutagenesis screen for splicing, the authors should compare to other models including Pangolin, SpliceAI and MMSplice - the latter having been developed for this.

Response: Following the different reviewers' comments, we have revised the manuscript to focus more on the segmentation aspects, adding extensive analyses and baselines on gene annotation, splicing prediction and species generalization. As such, we have removed the analyses of this isoform mutation dataset and will leave the extension to mutations for follow-up studies.

Reviewer 3: Variant effect prediction can indeed be considered out of the scope of a genome annotation method, whereby exploiting correlative signals may be very effective. Consequently, the following statement of the discussion should be replaced by a statement about future work to evaluate the capacity of this approach for variant effect prediction:
" Third, the accuracy of SegmentNT predictions can be leveraged to evaluate the impact of sequence variants on the different types of genomic elements, as we showed for splicing isoforms."
Related to this, the last word of this sentence is wrong:
"In addition, the SegmentNT-30kb-multispecies model showed improved performance over its human counterpart for the prediction of splicing variants (Supplementary Fig. 6d)."
It should be:
"In addition, the SegmentNT-30kb-multispecies model showed improved performance over its human counterpart for the prediction of splice sites (Supplementary Fig. 6d)."

Reviewer 3: Minor Points
16. The name of the tools is a misnomer. Segment-NT is not a segmentation algorithm. It is a single-base multitask classification. Proposing a segmentation algorithm could have been methodologically innovative and may be a safer direction at this stage that the authors could investigate.

Response: We would like to respectfully disagree with this reviewer's opinion. SegmentNT is indeed aligned with segmentation principles - with segmentation being simply the task of assigning a class label to every single pixel of an input image for computer vision, or assigning a class label to every nucleotide of a DNA sequence in this genomics context (see for instance book "Digital Image Processing", Rafael C. Gonzalez & Richard E. Woods, 2018). More precisely, SegmentNT is doing instance segmentation - computing a binary mask over entities for each "instance" (here instance refers to an element type and entities to nucleotides; see for example https://www.ibm.com/topics/instance-segmentation). While this is a form of multitask classification, it effectively segments DNA sequences by assigning functional labels to each nucleotide, predicting the boundaries and labels of genomic elements at single-nucleotide resolution. We will clarify these points in the paper. If the reviewer is asking for us to perform semantic segmentation instead (attributing one label to each entity), this would limit us as we would have to remove some element types; for instance we couldn't have both protein coding genes and introns/exons.

Reviewer 3: Thanks. Other fields, other definitions. A segment, geometrically, is a part of a straight line bounded by two distinct endpoints. It is commonly understood that segmenting a genome amounts to partitioning it into intervals (other example: https://academic.oup.com/bioinformatics/article/22/16/1963/208107). Some classes of algorithms more naturally favour such outputs, where adjacent bases tend to be of the same class. One example is Tiberius, the deep learning follower of AUGUSTUS, which employs an HMM to this end.
This point is not as academic as it seems. It relates to another miscommunication regarding statistical assessment with Reviewer 4. Reviewer 4 underscores in their points #1 and #2 that adjacent bases are not independent and, therefore, that statistical assessment shall not be done at the base level. To address this, the authors should use ground truth regions (e.g. a ground truth whole exon) as one instance instead of considering all bases of a region as independent observations.

New minor points:
2) P.9 the main ioform -> the main isoform
3) The discussion should now mention concurrent work published since initial submission, notably Tiberius (Gabriel et al. Bioinformatics 2024)
4) P. 10 "Pangolin compared with SpliceAI on non-coding RNA splices" -> splice sites.

Code:
We tested the HuggingFace code which worked. It should somewhere be mentioned that the HuggingFace uses FlashZoi, a different implementation of Borzoi and slightly different model, than the Jax implementation upon which the results are based.

Reviewer #4 (Remarks to the Author):

We appreciate that the authors have done a lot of work and responded to nearly all of our comments. We especially appreciate the convincing demonstration of the value of pre-training and the exploration of model mis-predictions. Overall, the manuscript is significantly improved. However, we believe there are still some key issues that need to be addressed before the manuscript can be published.

-The code for segment-enformer and segment-borzoi is not available on their huggingface space or github.

-We are grateful that the authors decided to address the incorrect use of “the term 0-shot” by removing it, but we still find the term in the figure legend of Figure 2, where the authors report zero-shot generalization of segmentNT. We emphasize that including these “buzzwords” will only serve to further sow confusion in our field. Please avoid. Similarly, use of the term “foundation model” has been recently criticized by Yun Song’s group when referring to genomics models (see “Genomic Language Models: Opportunities and Challenges” https://arxiv.org/html/2407.11435v1). We also suggest that the authors avoid this term in their work and focus on their actual findings related to genome annotation.

-The authors still have not presented a region-based segmentation evaluation statistic appropriate for segmentation. Several were suggested in the first review, but the authors write:

“Regarding the metrics at “region” level, we do not understand the definition of regions used by the reviewer. As mentioned above, we have used metrics (and error bars/uncertainty estimates) appropriate for instance segmentation, accounting for intersection over union and average precision. If the reviewer provides a mathematical definition of “region” and metric we would be happy to compute it.”

First, authors appear to misunderstand the concept of “instance segmentation.” If we take the analogy with images, their classification of individual nucleotides into 14 categories can be considered a multi-label “semantic segmentation.” Instance segmentation would distinguish each specific CDS or UTR. The authors need to correct this confusion in the final manuscript. https://en.wikipedia.org/wiki/Image_segmentation

Second, the authors’ confusion about region-level metrics is analogous to confusing pixel-level metrics and object-level metrics in image segmentation, which is shown as a pitfall in Figure 1a of this review: “Metrics reloaded: recommendations for image analysis validation, Nature Methods volume 21, pages 195–212 (2024)” The authors claim to be segmenting multiple objects of different sizes (e.g., exons, UTRs, splice sites, etc., etc.,). Therefore, an object-level metric for the different regions would give an actual estimate of segmentation performance. As it is now, the authors have only given nucleotide level classification performance. A region-level metric would estimate the agreement between the nucleotides within the predicted segment and the known boundaries. For example, if CDSs go from X:200-300 and X:500-505, and the predictions are from X:190-310 and X:490-515, the intersection over union at the region level is (0.83 + 0.20) / 2 = 0.52. However, if evaluated at the nucleotide level as the authors are doing, we get 100 + 5 / 120 + 35 = 0.68. This shows how pixel/nt-level metrics are strongly biased towards the performance on larger objects/regions. Also note that the true sample size is n=2 exons, while at the nt-level we have n=155. The nucleotide-level metric will be associated with greatly over-estimated confidence (biased variance estimates).

-We thank the authors for doing additional baselines to assess the value of pretraining. However, the authors claim to use a 1024-vocabulary, but this is not true. In fact, they are using a 1024 embedding dimension of the 1-hot encoded DNA. The authors should correct this in the manuscript. (To make an actually fair comparison, they could tokenize the DNA into a 6-mer vocabulary because that is the input to the nucleotide transformer. This would also address the difference in length of tokenized input (L vs. L/6) for this experiment.)

We still noticed many cases where detail is lacking and mistakes have not been corrected:
1. We thank the authors for removing the chunks with homologous genes from their test data. The authors should clarify whether this also removes homologous distal regulatory elements, or whether homology could still be inflating performance on these regions.
2. Similarly, the authors “sample” the test set 10 times to obtain estimates of confidence, but the sampling strategy is not explained. Ideally these are non-overlapping partitions.
3. There are also minor writing issues like English spelling errors (inclused vs. included, ioform vs. isoform, etc), surnames in references missing diacritics, e.g., Tomáš Brůna), and the authors refer to “matched U-Net connections” but this is not a term we are familiar with.
4. The authors write “Overall, SegmentNT detects splice donor and acceptor sites in any given input DNA sequence with high accuracy” but the accuracy on non-coding RNAs is still very low (supplementary figure 6, panel c shows auPRC is approximately 0.02). This needs to be corrected.
5. Figure 1 panel b has not been corrected (length L should be N).

Reviewer #4 (Remarks on code availability):

The codes for all the models were available, although run inference_segment_borzoi.ipynb and inference_segment_enformer.ipynb did not run as expected on colab. Codes should be checked more carefully.

Version 2:

Decision Letter:

Our ref: NMETH-A55872B

2nd Jun 2025

Dear Dr. Pierrot,

Thank you for submitting your revised manuscript "Annotating the genome at single-nucleotide resolution with DNA foundation models" (NMETH-A55872B). It has now been seen by the original referees 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 Methods, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We think the remaining points raised by the two reviewers are very important. Please well address them and supply a point-by-point response letter when resubmitting the paper.

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

TRANSPARENT PEER REVIEW
Nature Methods offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.

Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Thank you again for your interest in Nature Methods. Please do not hesitate to contact me if you have any questions. We will be in touch again soon.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewer #3 (Remarks to the Author):

The authors have addressed most points satisfactorily, with the exception of the statistical assessments in figs 4a,c,f,g and 5b.

I remain puzzled as to why it is not possible to provide at least a basic statistical evaluation of the observed trends—if nothing else, the variability in their own data should be quantifiable.

The explanation that the analysis follows "published benchmarks" is not convincing. If previous studies did not include statistical assessments, that should not preclude this manuscript from meeting basic scientific standards.

I trust that the editors will be more patient and can enforce this fundamental requirement. I do not need to review the revised manuscript again.

Reviewer #4 (Remarks to the Author):

The authors have addressed all of the issues, and have finally added the SOV score as a region-level metric. Unfortunately, there are still two issues. Thanks to the authors for their patience through so many rounds of reviews.

1. They have implemented the SOV in a perplexing way. They wrote:

"The SOV score is sensitive to the predicted threshold and focuses primarily on true positives. To offer a more balanced

evaluation, we computed the SOV score for both positive and negative regions, and averaged these two SOV scores to obtain a robust measure of performance."

Can the authors give a citation/precedent for computing an evaluation on the negative regions? For highly imbalanced data, such as genome annotations, the negative regions are most of the genome for most of the classes. Finding the regions of the genome that are not exons (or enhancers or UTRs) has no biological use-case, and is computationally 'easy' since a predictor that predicts all negatives will score very well at this task.

The authors should simply report the actual SOV score (or IoU or any standard *region-based* segmentation measure). If they need to argue that it needs be modified for their problem, they can present that as a supplementary analysis along with clear explanation/justification for why they need to modify it in the context of their use-case.

2. The authors have not discussed the ablation results for the SOV:
"UNet large one-hot (252M)" obtains SOV nearly as high as the best performing models. Does this mean that most of the segmentation accuracy can come from the UNet head? If so, the authors should comment on this interesting finding.

Version 3:

Decision Letter:

21st Sep 2025

Dear Dr Pierrot,

We very much appreciate your patience when we check the revised version of your Article, "Annotating the genome at single-nucleotide resolution with DNA foundation models". I am pleased to inform you that this paper has now been accepted for publication in Nature Methods. The received and accepted dates will be 27th Mar 2024 and 21st Sep 2025. This note is intended to let you know what to expect from us over the next month or so, and to let you know where to address any further questions.

We also greatly appreciate your suggestion of cover arts for this paper. Our art editor and the journal editorial team will consider and discuss all cover art suggestions we receive from all the authors and let you know as soon as possible if we decide to use your cover art suggestion.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Methods 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. It is extremely important that you let us know now whether you will be difficult to contact over the next month. If this is the case, we ask that you send us the contact information (email, phone and fax) of someone who will be able to check the proofs and deal with any last-minute problems.

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.

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 or the NIH public access policy) then you should select the gold OA route, and we will direct you to the compliant route where possible. Because authors warrant under our subscription licensing terms that they haven't committed to licensing any version of their article under a licence inconsistent with the terms of our agreement – including the applicable embargo period – publication under the subscription model isn't suitable for authors whose funders require no embargo.

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.

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

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.

Please note that you and any of your coauthors will be able to order reprints and single copies of the issue containing your article through Nature Portfolio's reprint website, which is located at http://www.nature.com/reprints/author-reprints.html. If there are any questions about reprints please send an email to author-reprints@nature.com and someone will assist you.

Please feel free to contact me if you have questions about any of these points. Thank you very much again for publishing your paper at Nature Methods!

Best regards,

Lin Tang, PhD
Senior Editor
Nature Methods

Summary of the revision:

We sincerely thank all four reviewers for their thoughtful and constructive feedback on our manuscript. We are particularly grateful for the positive overall assessment of our work, together with the detailed and insightful comments and suggestions provided. These observations have guided us in revising the manuscript extensively, and we believe the changes have significantly enhanced its clarity, scope, and overall quality.

The main changes and additions are:

- Revised abstract, introduction and discussion to place this work within the broader field of genome annotation
- We have re-ran all validation analyses for the different models with (1) updated test sets accounting for orthologous regions, (2) multiple data splits and (3) additional segmentation metrics
- Extensive benchmarking against state-of-the-art tools specialized in the different domains of gene annotation, splicing detection and prediction of regulatory elements.
- Extensive model ablations and comparisons with competitive architectures
- Addition of two models using as DNA encoder the longer-range genomics models Enformer and Borzoi (SegmentEnformer and SegmentBorzoi), extending our segmentation models to 500kb sequences and with improved performance on regulatory elements

We have addressed all other comments by revisions to text and figures (see track changes in blue), and have detailed all changes in our point-by-point response below.

We would like to reinforce the novelty of our approach in leveraging DNA foundation models to annotate various genomic elements, not only in the human genome but also in less-well characterized species. Our revised manuscript demonstrates SegmentNT's state-of-the-art performance across multiple genomics tasks, including annotating gene and gene elements, detecting splice sites and predicting gene regulatory elements. Additionally, we show that SegmentNT can be combined with various DNA foundation models - such as Enformer and Borzoi - to segment sequences up to 524kb, significantly enhancing performance on certain genomic elements. We have expanded our multi-species analyses to demonstrate SegmentNT's improved generalization across species on all tested tasks compared to existing methods. Overall, SegmentNT's versatility supports a wide range of genomic applications, offering both insights into the genomic code and accurate predictions on key genomics tasks.

Reviewers' Comments:

Editor: Reviewers 1 and 2 co-reviewed the paper.

We have combined Reviewers 1 and 2 responses below since the comments were the same.

Reviewer #1 and #2 (Remarks to the Author):

This manuscript introduces SegmentNT, a deep learning model that combines a pre-trained DNA foundation model (Nucleotide Transformer) with a segmentation architecture (U-Net) to predict various types of genomic elements at single-nucleotide resolution in DNA sequences. SegmentNT can predict the base-resolution positions of 14 elements, including genes, regulatory regions, and splice sites, and explores extensions to expand the input sequence length up to 50kb. The authors demonstrate the value of using pre-trained DNA foundation models and techniques to extend the model's context length. They also developed a multispecies version of SegmentNT that can generalize to annotate genomic elements in other animal and plant species beyond humans. While the innovation and effectiveness of SegmentNT are nicely demonstrated, there are several major and minor concerns:

We thank the reviewer for acknowledging the innovation of our work and for all the constructive comments that allowed us to improve the manuscript further, very much appreciated.

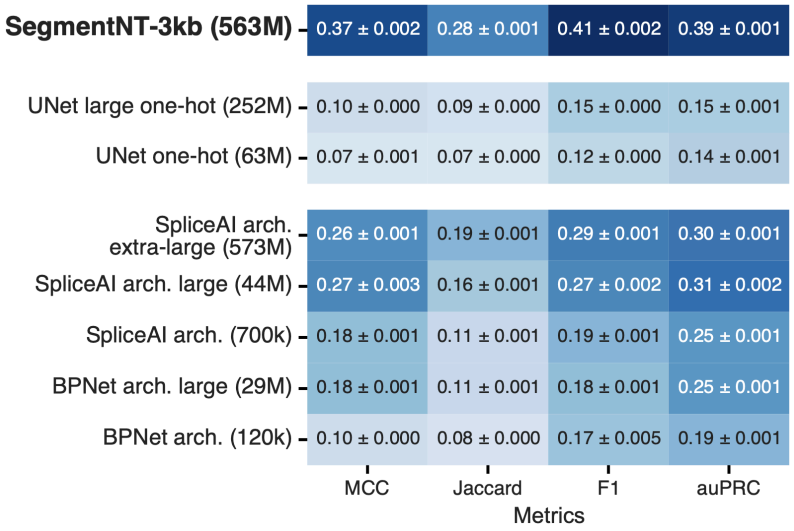
Major concerns:

• There is a lack of discussion about traditional genome annotation tools and how SegmentNT advances the field. For instance, traditional genome annotation approaches can be de novo, alignment-based (CACTUS), or combinations of the two (eg. Comparative Annotation Toolkit, CAT). They can also depend on functional assays, such as chromhmm. Placing this work's contribution within the broader field is important for people to understand how Segment NT is advancing upon previous methods.

We thank all reviewers for highlighting that our introduction and discussion sections were limited regarding the genome annotation field and how SegmentNT fits in it. We have now expanded these sections taking into account all reviewers feedback (see page 2 and 16), and have also performed extensive benchmarking on gene annotation tasks to demonstrate the capabilities of SegmentNT.

• SegmentNT’s performance for splice donor and acceptor in Fig 1e appears the highest. However, the ablations show that UNet performs poorly when using one-hot sequences. UNet was never established as a meaningful model for gene annotations (even though it makes sense); UNet is clearly beneficial when using NT representations as input. On the other hand, SpliceAI, which is not a UNet, but takes one-hot sequences performs much better on predicting splice annotations. Perhaps a better supervised baseline would be to take spliceAI and predict all 14 tasks.

We agree with the reviewer that including additional supervised baselines that have been shown to provide strong performance on annotation tasks from one-hot sequences is important for the sake of a fair comparison. As such, we have now added two different CNN supervised baselines that proved successful in modeling single-nucleotide data: (1) SpliceAI that was developed to predict splice sites (Jaganathan Cell 2019) and (2) BPNet that was developed to predict transcription-factor binding (Avsec Nat Genetics 2021). We used the published architectures and trained them to predict all 14 tasks across the 10 different test set folds (see main results below and in revised Fig 1e). Please note that here, for the sake of this ablation, we have re-used the same neural architecture as SpliceAI and BPnet but not existing weights as we re-trained them in the same conditions and on the same data as for segmentNT. Note also that we had to adapt the very final layer of the network to provide 14 predictions per nucleotide to be comparable. As the reviewer suspects, we observed higher performance over the UNet baselines, but still far from the SegmentNT results. However, while this gap in performance could be explained by the difference in architecture, it could also be explained by the difference in terms of numbers of parameters between the original SpliceAI/BPNet and SegmentNT. To investigate the second hypothesis, we scaled both BPNet and SpliceAI to larger counts of parameters up to 573M which is slightly superior to segmentNT 563M. We observe that while scaling increases performance, it is not enough to compete with SegmentNT, thus supporting the superiority of our architecture and pre-training methodology. We have now added these additional, stronger supervised baselines that enrich our evaluation pipeline and further demonstrate the effectiveness of the SegmentNT framework, many thanks.



• SpliceAI-10k's performance in the original study was 0.98 PR-AUC. In this paper, it is 0.96 for acceptor and 0.94 for donor. Why is there a discrepancy?

This discrepancy is due to small adaptations in the dataset to allow a direct comparison with SegmentNT settings. Namely, we consider sequence windows of 30kb (instead of 10kb in the original publication) and compute the predictions on the whole window (instead of not predicting

for the flanking +/- 5kb sequence). We have also removed sequences with Ns due to the constraint of the SegmentNT architecture. We have added these details in the Methods section (see page 26).

• I would not draw the same conclusion as "These results show that the segmentation capabilities of SegmentNT can be used to predict complex gene rearrangements directly from the sequence, which should be a useful tool for the interpretation of sequence and structural variants that can affect gene regulation and disease." from a PCC of 0.24. I think further evidence is needed to support this claim.

Following the different reviewers' comments, we have revised the manuscript to focus more on the segmentation aspects, adding extensive analyses and baselines on gene annotation, splicing prediction and species generalization. We agree with the reviewer that further evidence is needed to fully validate the use of SegmentNT for mutation analyses. As such, we have removed the analyses of this isoform mutation dataset and will leave the extension to mutations for follow-up studies, many thanks.

• For the MST1R dataset testing isoform change prediction, SpliceAI should be included in the comparison as a relevant baseline.

As mentioned in the response above, we have removed this analysis to focus the paper more on the different segmentation aspects.

• For the Enformer evaluation, 7 ATAC-seq profiles were selected, but it's unclear which ones. Would it make more sense to use Enformer's representations and place a UNet over that as a direct head-to-head comparison with NT embeddings, especially since Enformer's bin size is not base resolution and the promoter prediction task demands it? On that note, how are edge effects handled due to binned predictions versus base-res labels?

After seeing the different comments of all reviewers regarding the Enformer evaluation, we fully agree that it would be more interesting and useful to evaluate Enformer on the SegmentNT settings and have revised the manuscript extensively on this aspect.

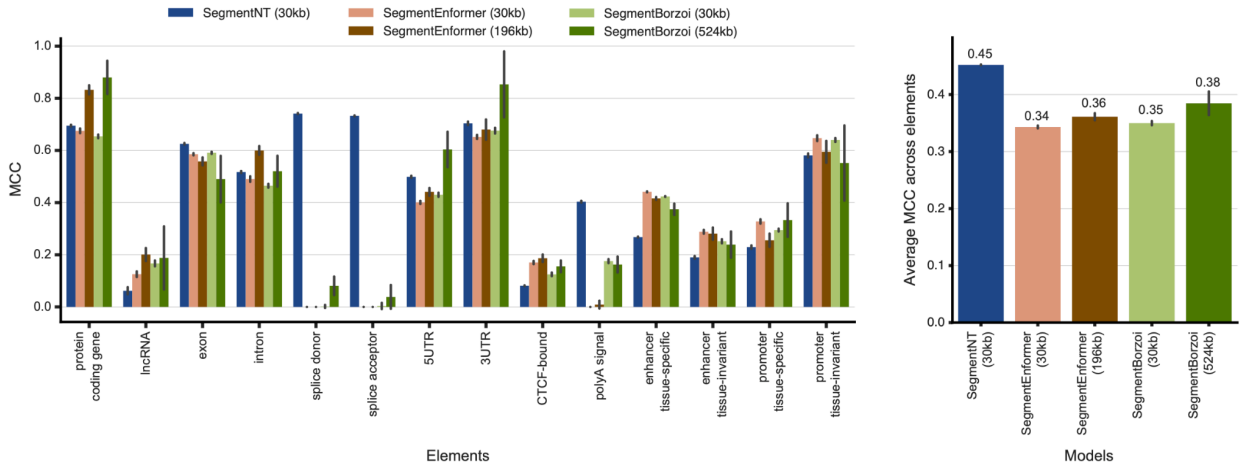
Our initial idea was to compare SegmentNT with published models that can be used to identify promoter and enhancer regions. Since we could not find any model that could be used with good performance for testing in a whole chromosome dataset (outside the standard curated datasets used for model train and testing), we explored how one could use Enformer predictions as surrogates of regulatory activity, which indeed demonstrated good performance. However, we agree that that is not the main use case of Enformer and these prediction results can be misleading.

Therefore, in this new analysis, we decided to investigate the Enformer under a new angle. As shown in our previous NT paper (Dalla-Torre et al 2023), the Enformer model can also be used as a powerful feature extractor. As our SegmentNT framework is general, we can train it using

other feature extractors than NT as the DNA encoder. As such, we have now evaluated Enformer (and Borzoi) as competitive DNA encoders of Nucleotide Transformer by placing a UNet on top of the Enformer/Borzoi representations, calling them SegmentEnformer and SegmentBorzoi, respectively (see Methods). This also allowed us to test the performance of models that take longer context windows as 196kb (Enformer) and 524kb (Borzoi). In addition to taking longer context, both models have been trained on chromatin accessibility, transcription factor binding and gene expression data and as such have probably learned to extract features that would be useful to better annotate some types of elements such as the regulatory elements. Note, however, that while NT has been trained to produce representations over 6-mers, the Enformer and Bozroi were trained respectfully at 128bp and 32bp resolution. Therefore, we adapted the U-NET module to upsample accordingly and we might expect a loss in performance over the detection of small elements.

We finetuned the whole Enformer/Borzoi+UNet networks on our segmentation dataset using either the same 30kb input sequences or the model's original input length (see main results below). In summary, this showed that, as expected, Enformer and Borzoi allow improved performance on regulatory elements but cannot solve high-resolution tasks such as the prediction of splice sites and polyA signals. On gene elements they have lower performance when using 30kb sequences as input, but the performance on some elements increases when using their longer-range inputs - namely protein-coding genes and introns, longer elements that previously benefited from increased sequence length (see Fig 2a). We have added these results as new main Figure 3.

We think this result showcases how the SegmentNT framework can also be extended to additional foundation models to achieve longer sequence contexts and superior performance on particular genomic elements. We will also release these two additional models, segmentEnformer and segmentBorzoi and believe that they will be useful to the community especially to researchers interested in longer context annotation or regulatory genomics. Many thanks for this suggestion!



• SegmentNT's predictions are given at base resolution. I suppose the annotations with hard boundaries make sense for gene features, such as exons, but setting hard boundaries for regulatory elements is a bit more arbitrary. The predictions, at least for regulatory elements, seem to resemble broad distributions. So, are the performance gaps coming from edge effects of annotations or spurious predictions in random locations? How would a Gaussian filter applied to the CRE annotations perform? Perhaps if the variance of the filter were systematically varied, this could help calibrate how much the metrics are influenced by poorly predicted edge effects.

It is indeed true that regulatory elements have artificial boundaries coming from the need to define their genomic positions, which we have used in this study. To investigate in more detail if the mispredictions in these elements are coming from edge effects of annotations or spurious predictions in random locations, we have splatted nucleotides into 3 classes and calculated the enrichment of mispredictions in each class: edge if they are closer than 50nt from a region boundary, interior if further away but inside a positive labeled region, and part of spurious regions otherwise (see new Supplementary Table 4). Overall, we observed among the mispredicted nucleotides an enrichment at edge nucleotides but also at interior regions. For all regulatory element classes, the enrichment of mispredictions inside elements was higher than at the edges, showing that the performance gaps do not come from poorly predicted edge effects due to undefined hard boundaries, nor from spurious predictions in random locations. We mention that result in the results section (see page 7) and have added a new Supplementary Table 4 with all metrics and details. Lastly, we note that, as mentioned above, we have also developed the SegmentEnformer and SegmentBorzoi models that show significantly improved performance specifically on regulatory elements and can be used for these tasks (see more details in answers to respective reviewer comments).

• The performance is summarized for each element across all loci. It could also be informative to look at performance averaged across annotation tasks within loci. This way, one can monitor performance within each loci and then further investigate the loci that are poorly predicted to understand SegmentNT's limitations.

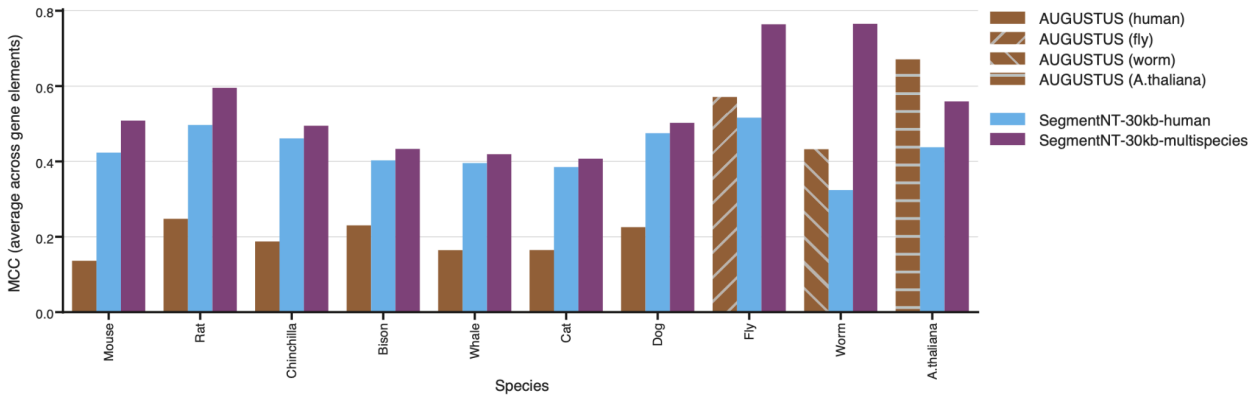
We agree with this reviewer that analysing the performance across loci can provide further insights into SegmentNT's performance. We have inspected its performance across loci but could not find a fair way of summarising it with appropriate metrics due to the large unbalancing of labels between regions, in particular because most regions are intergenic and do not have any labels. In addition, when inspecting different regions with various properties, we did not observe any specific patterns that we could highlight in the manuscript. To circumvent that, we have made available in the revised manuscript a genome browser session with the labels and predictions of SegmentNT along the test chromosomes that should enable the reader to evaluate the performance across regions of interest. We hope this satisfies the reviewer's concerns and appreciate this comment.

• Although it may be out of scope for the paper, it would be interesting to see if the multi-tasking nature helps or hurts performance. Perhaps task ablations?

We thank the reviewer for the insightful suggestion. While we agree that exploring the impact of the multi-tasking nature through task ablations would be an interesting direction, conducting these additional experiments would entail substantial computational resources. Given that this investigation falls outside the primary focus of the paper, as the reviewer acknowledged, we opted not to pursue these experiments.

• The generalization across species should be benchmarked against existing state-of-the-art gene finder tools (if possible).

We agree and have now expanded the comparison with established gene annotation pipelines also across species, for the ones possible. In this case we compared our results with AUGUSTUS, a state-of-the-art *ab initio* (purely sequence-based) gene finder method developed for human but also multiple species. As recommended by AUGUSTUS authors, we used the human model for annotation of mammalian species, and additional specialized AGUSTUS models for the other species available (fly, worm and Arabidopsis; see legend). In summary, we observed that our human model outperforms AUGUSTUS on gene annotation in all mamalian species. In addition, our multispecies model further outperforms our human model and AUGUSTUS itself in all species except for Arabidopsis, with SegmentNT being a single model compared with different AUGUSTUS model versions per species. This shows the superiority of SegmentNT for gene finder and annotation in human but also across multiple species. We present these results in new Fig 7i and Supplementary Fig 9 (see also below), displaying the average metric across the gene elements per species. Many thanks for the suggestion.



Minor concerns:

• How is the probability of belonging to each of the 14 genomic elements at the nucleotide level calculated?

Our output layer predicts for each nucleotide, 2 logits per genomic element. This means that the output tensor of our model has a shape (batch_size, sequence_length, num_elements, 2) where typically num_elements = 14. These logits are then passed through a softmax layer, applied over the last tensor dimension, that returns a tuple (p, 1-p) per nucleotide and per genomics

element where p is the probability of the nucleotide belonging to the element. We have clarified this in the Methods section.

• There is a data leakage possibility for gene elements due to orthologs and gene duplications.

We agree with this reviewer and have improved our data splits to address this concern. More precisely, we have removed from the test set of the human and all other species datasets the genes with homologous/orthologous genes in the human training chromosomes. We observed minor differences, showing that the results were not driven by data leakage, and have updated all results and figures accordingly.

• It would be nice to see performance under different metrics, not just MCC, such as AUPR and AUROC.

We agree and have now added for all models their performance for three additional metrics: area under the precision recall curve (auPRC), Jaccard similarity and the F1-score. All metrics showed consistent results (see new Supplementary Fig 2 and Fig 1e).

• The pictures in Figs 5a and 5b are too small.

We have removed the pictures since they were just illustrative and we could not find an aesthetically elegant way of having all of them.

Reviewer #2 (Remarks on code availability):

Code is accessible and easy to use. README file provides good tutorial and information to reproduce the models.

We appreciate the reviewer's note that the code was easily accessible. It was important for us to make our work and models as widely accessible as possible.

Reviewer #3 (Remarks to the Author):

Summary

—----------

This manuscript introduces SegmentNT, a deep learning model that leverages a pre-trained DNA language model (Nucleotide Transformer, NT) and a 1D U-Net architecture to predict the location of various genomic elements within DNA sequences at single-nucleotide resolution. The authors demonstrate SegmentNT's ability to predict the location of both gene elements (exons, introns, splice sites, etc.) and regulatory elements (promoters, enhancers) in the human genome, claiming superior performance compared to baseline models and generalization capabilities to other species. They further highlight the model's efficiency and its potential for analyzing sequence variants and their impact on transcript isoforms.

General Remarks

—----------

This manuscript presents SegmentNT, a deep learning model for genome annotation that leverages a pre-trained DNA language model. While the exploration of this approach is timely given the increasing interest in foundation models for genomics, the manuscript currently falls short of demonstrating its practical value and significance to the broader biological community. The authors primarily focus on showcasing the technical feasibility of using a DNA language model for annotation, but lack a clear articulation of how SegmentNT addresses the limitations of existing methods and advances the state-of-the-art in genome annotation. The benchmarking strategy is currently relying on artificially constrained comparisons, a misuse of existing tools, and could focus more on known tools and relevant specialized methods. We are left with a premature set of investigations, whose documentation has perhaps the merit to attract the attention to a traditional computational biology area that will benefit from LLMs, but with the drawback that it does not advance the state-of-the-art and will confuse further newcomers to computational biology coming from other areas of machine learning by providing a set of flawed benchmarks and tasks. Overall, the study requires a clear focus to establish the practical utility of the approach, to provide more rigorous comparative analyses, and to clearly articulate its contribution to addressing key challenges regarding genomics research.

We thank the reviewer for highlighting that our work is timely, and for the constructive feedback that allowed us to improve the manuscript and to demonstrate the practical value and significance of SegmentNT.

Major Points

—----------

1. The motivation as currently presented is too limited to be of broad interest and this is reflected in many analyses presented by the author. Currently, the main aim of the paper seems to be to prove that DNA-LMs can „also“ do genome annotation. This is interesting from a methodological perspective. However, to be impactful, the work must be placed into the context of the general literature on genome annotation, clearly identify where current tools are lacking and explain how a DNA-LM based approach would improve upon state-of-the-art problems of interest to biologists. Specifically, the study entirely ignores how genome annotation (see pipelines e.g. used by the Ensembl annotation team) is performed. Those pipelines and the tools they build upon have a single-nucleotide resolution. Ignoring this literature is concerning, starting from the fact that Segment-NT takes the output of these pipelines as ground truth. The limitations of these tools and pipelines are not spelled out.

We thank this reviewer for pointing out that we did not make clear the scope of SegmentNT and how it advances state-of-the-art problems of interest to biologists. This point was common to all reviewers and thus we have combined their feedback to extensively revise the manuscript to accommodate their comments. Namely, we have expanded the different introduction/results/discussion sections with more details about current models and pipelines,

their limitations and how SegmentNT compares with them; and we have added extensive benchmarking on gene annotation tasks to demonstrate the capabilities of SegmentNT. We hope the revised version conveys better the advances of SegmentNT in both the DNA-LMs and the genome annotation fields, many thanks for your comment.

2. Related to this, the manuscript lacks comprehensive benchmarking against established gene annotation tools. To demonstrate SegmentNT's practical value, the authors should compare its predictions to established gene annotation pipelines such as for example AUGUSTUS or MAKER, using standard evaluation metrics and datasets for gene prediction accuracy. Additionally, while tools like Deepromoter are mentioned in the introduction, a comparison against these specialized methods is missing.

We agree that our previous manuscript version had limited benchmarking against established gene annotation tools. We have now expanded the comparison with established gene annotation pipelines. Since MAKER is a pipeline that includes various models that do not use sequence only but also require other features and experimental data, we decided to compare SegmentNT with AUGUSTUS itself, the state-of-the-art *ab initio* (purely sequence-based) method, which is also used within MAKER, to make an apple-to-apple comparison.

We evaluated all AUGUSTUS models in different datasets and settings of gene annotation tasks, including segments of only genes and predicting only the main isoform or all isoforms, or in whole-chromosome setting (see detailed results in results section “Comparison with established gene annotation tools”, Fig 4 and Supplementary Fig 5). In summary, SegmentNT is competitive with AUGUSTUS for the easier setting of segmenting the main isoform of various genes, but it outperforms AUGUSTUS when predicting all confidently annotated gene isoforms. Within the test set settings of SegmentNT, where the model needs to segment the gene regions along the whole test set chromosomes, SegmentNT achieved superior performance across all gene elements by a large margin, with both higher recall and precision. This result agrees with the limitations of this type of *ab initio* tools when scanned through the whole genome, in contrast to their very high accuracy when segmenting a genic region. These results substantially extend the scope and relevance of our manuscript , many thanks for the suggestion.

Finally, as suggested by the reviewer, we have also added comparisons with DeePromoter for the prediction of promoter regions. Since the web-server for promoter prediction developed by the authors (https://home.jbnu.ac.kr/NSCL/deepromoter.htm) was no longer available, we retrained their model using the pytorch implementation provided at https://github.com/egochao/DeePromoter/. We managed to reproduce the results in their test set and thus have the correct model checkpoints to evaluate on our benchmark (see more details in Methods). We evaluated DeePromoter in a setting where we slide the model over the whole test chromosomes and evaluate if promoter regions are correctly predicted by the model over background regions. Despite the strong performance of DeePromoter on their test set, we observed poor generalization to the context of genomics sequences, likely due to the problematic design of their test set (constructed only based on shuffling parts of promoter regions) (Fig 6). DeePromoter (auPRC=0.01) performed worse than SegmentNT (0.39) but also

worse than another binary promoter classifier trained on a more general positive/negative dataset (0.25), showing the limitation of the model itself and not the sliding window approach. In summary, we have added additional comparisons between SegmentNT and competitive approaches for promoter and enhancer identification (including our new SegmentEnformer and SegmentBorzoi models) through the whole genome (see new main Fig 6).

3. Statistical significance testing is largely absent. The authors should perform appropriate statistical tests to determine whether the reported performance differences between models are statistically significant.

We agree with this reviewer and have now added error bars and confidence intervals to all models and metric comparisons (see Fig 1c,e, 2a,b, new Supplementary Fig 2 and Supplementary Tables 2 and 3). In addition, we have compared the performance of all models using statistical testing by performing paired t-tests using their performance across the 10 different data samplings of the test set (see new Supplementary Fig 4). These results provide stronger support to our conclusions, many thanks!

4. Currently, SegmentNT is benchmarked against ablations or competing methods that are artificially constrained. SegmentNT is benchmarked against a version that is randomly initialized. This comparison does not account for the amount of compute spent on pretraining the Nucleotide Transformer: “the version with random initialized NT showed much slower convergence and have not converged yet even after 68M training sequences (34B tokens), a training more than three times longer”. The authors should train the model until convergence if they really want to make a fair point about the effect of pre-trained DNA language models.

We agree and have now trained the randomly initialized model until convergence. When comparing both models we observed that self-supervised pre-training on genomes allows our SegmentNT model to converge 7 times faster to an asymptotic performance twice greater across all 14 genomic elements: average MCC 0.37 compared with 0.16 for the version with a random initialized NT. We have added these and additional model ablations and baselines to Fig 1e and Supplementary Tables 1 and 2.

5. When comparing the supervised Enformer to SegmentNT, the comparison would only be fair if both had the same U-Net-head on top of their embeddings. Currently, the authors use the prediction over 7 selected ATAC-seq (or DNase?) profiles as promoter-proxy, neglecting that many regions in the genome can be accessible, including enhancers, transposable elements and potentially parts of coding sequences, and thus accessibility is not an exclusive mark for promoters. Encode, for example, categorizes promoters as „Elements (cCREs) putatively labeled as promoters defined by a signature of high DNase and H3K4me3 signal and that lie within 200 bp of an annotated GENCODE transcription start site (TSS).“ which indicates that accessibility alone is insufficient.

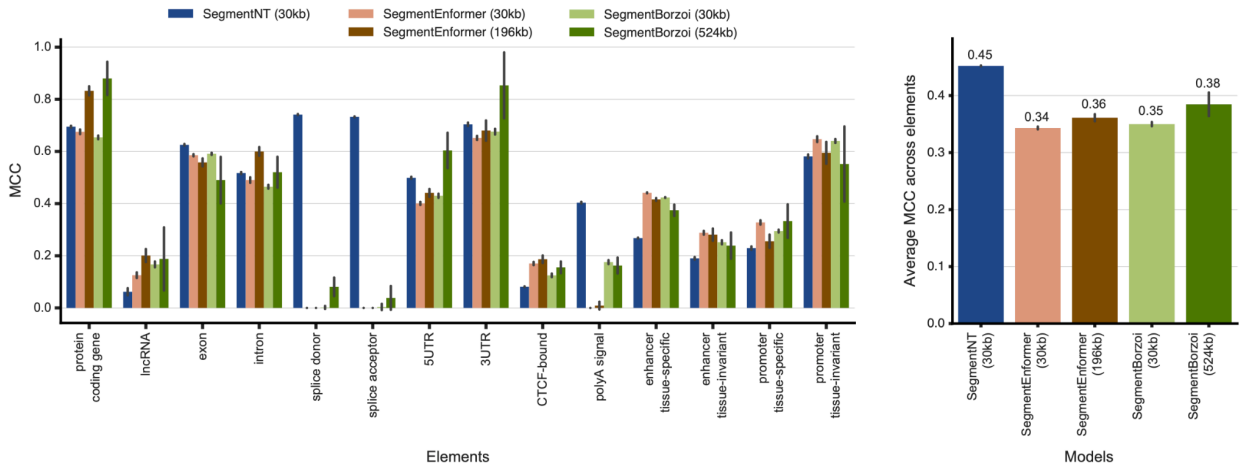
After seeing the different comments of all reviewers regarding the Enformer evaluation, we fully agree that it would be more interesting and useful to evaluate Enformer on the SegmentNT settings and have revised the manuscript extensively on this aspect.

Our initial idea was to compare SegmentNT with published models that can be used to identify promoter and enhancer regions. Since we could not find any model that could be used with good performance for testing in a whole chromosome dataset (outside the standard curated datasets used for model train and testing), we explored how one could use Enformer predictions as surrogates of regulatory activity, which indeed demonstrated good performance. However, we agree that that is not the main use case of Enformer and these prediction results can be misleading.

Therefore, in this new analysis, we decided to investigate the Enformer under a new angle. As shown in our previous NT paper (Dalla-Torre et al 2023), the Enformer model can also be used as a powerful feature extractor. As our SegmentNT framework is general, we can train it using other feature extractors than NT as the DNA encoder. As such, we have now evaluated Enformer (and Borzoi) as competitive DNA encoders of Nucleotide Transformer by placing a UNet on top of the Enformer/Borzoi representations, calling them SegmentEnformer and SegmentBorzoi, respectively (see Methods). This also allowed us to test the performance of models that take longer context windows as 196kb (Enformer) and 524kb (Borzoi). In addition to taking longer context, both models have been trained on chromatin accessibility, transcription factor binding and gene expression data and as such have probably learned to extract features that would be useful to better annotate some types of elements such as the regulatory elements. Note, however, that while NT has been trained to produce representations over 6-mers, the Enformer and Bozroi were trained respectfully at 128bp and 32bp resolution. Therefore, we adapted the U-NET module to upsample accordingly and we might expect a loss in performance over the detection of small elements.

We finetuned the whole Enformer/Borzoi+UNet networks on our segmentation dataset using either the same 30kb input sequences or the model's original input length (see main results below). In summary, this showed that, as expected, Enformer and Borzoi allow to achieve improved performance on regulatory elements but cannot solve high-resolution tasks such as the prediction of splice sites and polyA signals. On gene elements they have lower performance when using 30kb sequences as input, but the performance on some elements increases when using their longer-range inputs - namely protein-coding genes and introns, longer elements that previously benefited from increased sequence length (see Fig 2a). We have added these results as new main Figure 3.

We think this result showcases how the SegmentNT framework can also be extended to additional foundation models to achieve longer sequence contexts and superior performance on particular genomic elements. We will also release these two additional models, segmentEnformer and segmentBorzoi and believe that they will be useful to the community especially to researchers interested in longer context annotation or regulatory genomics. Many thanks for this suggestion!



6. Related to the previous point, for regulatory elements such as promoters/enhancers, it is somewhat unclear whether there is an added value to predicting a label such as „enhancer“ compared to predicting biochemical marks such as accessibility, histone acetylation or transcription factor binding, as is done by Enformer. In the end, the biochemical marks represent the biological mechanism whereas annotations such as „enhancer“ reflect a human attempt to systematize these mechanisms. The existence of enhancer RNAs, for example, implies that the distinction between enhancer and promoter is not always clearly definable. As such, it is not obvious whether a model that predicts human labeling of regulatory elements really can provide a benefit over a model such as Enformer that directly predicts biochemical tracks, particularly for applications in human genetics, where ENCODE already is near exhaustive. Potentially the authors could address this by showing that their labeling of regulatory elements provides some benefit in the cross-species setting, e.g. by using it to predict when the orthologues of enhancers cease to be functional [Kaplow, et al. Science 2023].

That’s an interesting point. We agree that those are two different perspectives. On one hand, predicting biochemical marks directly (e.g. Enformer) allows higher functional resolution of a given sequence; but these models are always trained on a specific subset of cell types and tracks and cannot generalize to unseen cell types, tracks and species. In addition, it is hard to systematically consolidate these biological mechanisms into classes of genomic regions that can be studied. Another possibility, that we favored in this work, was to consider genomic elements as discrete regions that were already annotated and labeled based on those biochemical marks. Predicting these discrete labels offers additional value on two main fronts: (1) such labels encapsulate multiple biochemical signals, which may collectively represent functional distinctions that are less apparent when evaluating individual marks in isolation; (2) predicting these consolidated labels allows to generalize across species better than predicting biochemical data that is more tissue/species-specific, as mentioned by the reviewer. We agree that it is important to compare our setting with the one from Enformer and have expanded on this in the discussion section (see page 16), many thanks.

7. The application of Enfomer is flawed. The authors use a reimplementation of it. There is no data showing that the reimplantation is correct (how much do the predictions differ). It is referred to a non-reviewed preprint which does not show this either. More critically, Enformer takes 200kb as input. The authors run Enformer with only 10kb padding the rest of the sequence. It is therefore very misleading to label such predictions as “Enformer” predictions.

Regarding the first point of reimplementing Enformer in our Jax codebase we have validated its exact reproduction by showing an absolute difference of less than 1e-5 in all activations as well as in the predictions between the original implementation and ours. We have added these details to the methods (see page 24). In addition, the preprint referred by the reviewer has now been peer-reviewed and accepted, further validating the choice of this implementation.

Regarding the benchmarking of Enformer we agree with this and the other reviewers that the way we benchmarked against Enformer was not the most appropriate. As mentioned above, we have now revised this and evaluated Enformer (and Borzoi) on the segmentation task directly using the U-Net head on top of their embeddings, as suggested by all reviewers. We have discussed the results in another comment of this reviewer above and point to the new Figure 3 for further details.

8. Prometre and enhancer results are cherry-picked. The authors apply their approach to “the two best classes of regulatory elements”. Promoter tissue-invariant and enhancer tissue-specific (Fig 2a,b). Comprehensive results should be shown.

We have revised these analyses in three main aspects. (1) We have removed Enformer from the baselines, as discussed above, and added DeePromoter as a published baseline for promoters. (2) Since the baselines were trained to predict promoters and enhancers globally, not divided by tissue-specific or tissue-invariant as in SegmentNT, in this analysis we have now combined these two classes into a single promoter and enhancer classes and evaluated all models on these labels. (3) Finally, we have also benchmarked SegmentEnformer and SegmentBorzoi on these tasks as they showed improved performance on regulatory elements. See the new details in the Methods and the new results section and Figure 6.

9. The comparison between SegmentNT and SpliceAI is conceptually flawed. SpliceAI is developed to predict splice sites within precursor RNAs as clearly stated in the original publication. Genome annotation is not done by running SpliceAI on the plus strand of a genome. Nonetheless, the authors apply SpliceAI outside transcribed regions and on the reverse strand of genes. The authors then claim the superiority of their tool specifically on genes encoded in the minus strand (Figure 4b). This is extremely misleading.

We thank this reviewer for pointing out that we did not explain sufficiently clearly the comparison with SpliceAI. We indeed acknowledge that SpliceAI is developed to predict splice sites within precursor RNAs and have evaluated our models against SpliceAI (and now also Pangolin, as an additional baseline) within those settings, where our models show competitive performance. Since SegmentNT is more general than that and can be used on any sequence of interest, not

only the ones with gene sequences, we also compare it with the splicing models when running it along the whole genome. Here, we differentiate cases with genes encoded in the positive and negative strand since the splicing models were always trained with precursor RNAs in the positive orientation, and because when evaluating the models in a given random sequence without prior gene annotation there is no way of knowing in which orientation the gene is. For example, in cases of annotating new genomes without reference gene annotations, our approach allows us to find the genes and splice sites irrespective of their orientation and prior annotations. In the initial manuscript version, we mentioned the results for each strand separately for completeness, but we were clear that the splicing models only work for genes in the positive strand, and that was the test set scenario we used to claim that SegmentNT was superior on whole-genome setting.

Still, we understand that it can be misleading and have improved the splicing analyses extensively (considering all reviewers' comments), adding Pangolin as a new baseline, adding more metrics and precision-recall curves, and removing the result of genes in the negative orientation (see updated Fig 5 and Supplementary Fig 6).

10. For splice site prediction within precursor RNAs, Pangolin should be compared as well as it is multi-species and claims to outperform SpliceAI.

Many thanks for the suggestion. We agree and have added Pangolin to all splicing analyses, including the different human test set scenarios and the analyses across multiple species (see updated Fig 5 and Supplementary Fig 6). Many thanks.

11. For splice site prediction, the performance of SegmentNT on non-coding RNA splice sites should be shown as some of the performance could be driven by some correlative signal from coding sequence.

Thank you for this comment. Following up on the different reviewer comments we have extended our splicing analyses in humans and across multiple species, as well as incorporated additional baselines like Pangolin. This consolidated our results and showed that SegmentNT achieves state-of-the performance on splice site prediction in human sequences and outperforms current models in other species' genomes.

As suggested by the review, we have also added the performance of SegmentNT models together with SpliceAI and Pangolin baselines on non-coding RNA splice sites as suggested (see new Supplementary Fig 6c). Here we considered our whole-genome test sets with lncRNAs in the positive strand, and removed regions with protein-coding gene splice sites to focus on non-coding RNA. We observed that both SegmentNT and Pangolin achieve lower performance than SpliceAI on this non-coding RNA splice sites dataset. As mentioned by the reviewer, the high performance could be related to some correlative signal from coding sequence. An additional explanation for the superior performance of SpliceAI over both SegmentNT and Pangolin could be its use of novel splice junctions commonly observed in the GTEx cohort within the SpliceAI dataset, adding more diversity for non-canonical splice sites in

the training. This is an interesting observation that we added to the revised manuscript (see page 10 and new Supplementary Fig 6c) and that we will evaluate in future works, many thanks.

12. Any single metrics summarizing confusion tables has issues and MCC is not immune to this. Moreover, confusion tables depend on the cutoffs applied. Precision-recall curves in case of highly imbalanced classes, or ROC curves otherwise, should be shown to assess whether methods are uniformly better than others or only at some cutoffs. The 0.5 cutoff should be indicated on those curves for SegmentNT and the recommended cutoffs of other tools (SpliceAI has three recommended levels). For very imbalanced tasks, we recommend reporting the top-k accuracy, where k is the actual number of positives, see the SpliceAI publication and reference therein.

We agree and have now added for all models their performance for three additional metrics: area under the precision recall curve (auPRC), Jaccard similarity and the F1-score. All metrics showed consistent results (see new Supplementary Fig 2 and Fig 1e). For the splicing analyses, we have also added the top-k accuracy and precision-recall curves where we indicate the different cutoffs for each model, as suggested (see Fig 5 and Supplementary Fig 6). Many thanks.

13. Regarding the mutagenesis screen for splicing, the authors should compare to other models including Pangolin, SpliceAI and MMSplice - the latter having been developed for this.

Following the different reviewers' comments, we have revised the manuscript to focus more on the segmentation aspects, adding extensive analyses and baselines on gene annotation, splicing prediction and species generalization. As such, we have removed the analyses of this isoform mutation dataset and will leave the extension to mutations for follow-up studies.

14. In the multi-species scenario: since genes can be shared across species by homology and species have different numbers of chromosomes, the strategy of helding out chromosomes is subject to data leakage. The authors should assess this issue discuss and implement a strategy to circumvent it. This problem can particularly affect models like the one proposed which has a very high number of parameters in striking contrast to state-of-the-art genome annotation tools (generalized HMMs).

This is an important point to improve our evaluation and prevent data leakage. We have now removed from the test set of all species the genes with orthologous genes in the human training chromosomes. We have re-ran all validation analyses for the different models with the updated test sets across all species, including the 10 different data splits. Overall we observed minor differences, showing that the results were not driven by data leakage, and have updated all results and figures accordingly. Many thanks.

15. The manuscript claims that SegmentNT species generalization improves when trained on multiple species. However, the base model, NTv2, is in both cases already a multispecies (self-supervised) model and should have captured the semantics of different species. DNA

language models have previously been shown to capture regulatory motifs across evolution and not only on one target species [Nguyen et al,bioRxiv (2024); Karollus et al Genom Biol 2024]. It should be explored how much the SegmentNT U-Net is overfit to single species, if NTv2 has not learned these species well enough to contribute to genomic annotation tasks, or if worse performance is mostly an artifact of worse genome annotations for distant species.

We thank this reviewer for pointing out that we did not sufficiently explain the difference between the self-supervised pre-training of NTv2 and supervised training of SegmentNT. In fact, there is no notion of overfitting between NTv2 and SegmentNT, as NTv2 was trained through self-supervised learning and as such has never seen any of the labels used to train SegmentNT. The fact that NTv2 has been trained over genomes from several species is not a limitation but rather a strength as it improves transfer between these self-supervised and supervised learning phases (as we show in our ablation studies). In addition, the performance of the human SegmentNT model on the other species shows that this model can generalize across species, and that the species generalization improves when trained on multiple species.

Minor Points

—----------

16. The name of the tools is a misnomer. Segment-NT is not a segmentation algorithm. It is a single-base multitask classification. Proposing a segmentation algorithm could have been methodologically innovative and may be a safer direction at this stage that the authors could investigate.

We would like to respectfully disagree with this reviewer's opinion. SegmentNT is indeed aligned with segmentation principles - with segmentation being simply the task of assigning a class label to every single pixel of an input image for computer vision, or assigning a class label to every nucleotide of a DNA sequence in this genomics context (see for instance book "Digital Image Processing", Rafael C. Gonzalez & Richard E. Woods, 2018). More precisely, SegmentNT is doing *instance segmentation* - computing a binary mask over entities for each "instance" (here instance refers to an element type and entities to nucleotides; see for example https://www.ibm.com/topics/instance-segmentation). While this is a form of multitask classification, it effectively segments DNA sequences by assigning functional labels to each nucleotide, predicting the boundaries and labels of genomic elements at single-nucleotide resolution. We will clarify these points in the paper. If the reviewer is asking for us to perform *semantic segmentation* instead (attributing one label to each entity), this would limit us as we would have to remove some element types; for instance we couldn't have both protein coding genes and introns/exons.

17. In Fig. 1a, the pictogram incorrectly displays 3-mers as inputs instead of nonoverlapping 6mers.

Fixed.

18. What is the purpose of the U-Net architecture? As described in Fig. 1b, one could just go from (L, 1024) to (L, 6*14) directly. In contrast to Borzoi, which maps from single base-pair input to 32bp bins, the base-model NTv2 should have already learned to mix along the sequence dimension during pretraining, so that pooling in the U-Net might no longer be necessary. An explanation would be interesting for readers. Furthermore, is the NTv2 variant for the human SegmentNT the best option, or would that be the human-only NTv1?

The purpose of adding these additional U-Net layers during fine-tuning is to better capture multi-scale, hierarchical representations along the sequence, which can be crucial for capturing fine-grained patterns in sequence data. This hierarchical feature learning improves localization and contextual awareness, as verified by improved performance when combining NTv2 with the U-Net head. (see Fig. 1e). In addition, this approach makes SegmentNT more flexible since it can be used with any DNA encoder independent of its output dimensions (like the Borzoi case mentioned by the reviewer, that we also added in the revised manuscript, see new Fig 3). We have added these insights to Methods (see page 22).

In addition, we have also added the comparison of SegmentNT using NTv2 500M or the human 2.5B NTv1 model as DNA encoder which showed consistent improved performance for the NTv2 encoder (average MCC of 0.37 vs 0.34; see updated Fig 1e). These results align with the ranking obtained in the original NT study.

19. The emphasis on the speedup (Fig. 2C) is not motivated in the text. It is indeed unclear if inference speed is a concern for genomic annotation models that mostly will be run once per genome.

Although the inference speed is less relevant for segmenting full genomes since it will be run once per genome, as mentioned by the reviewer, this will become important when the use-cases are to segment regions with candidate genetic variants and personalized genomes, or cases such as in-silico sequence design and clinical trial data automated analysis where speed is paramount. For that reason, we focused on inference times per sequence and demonstrated high efficiency for our optimized setup in Jax. We added this notion to the manuscript (see page 12).

20. The manuscript could benefit from the inclusion of negative controls to assess the robustness of the model's predictions. Several recent studies analyzed how foreign DNA is interpreted by human cells. If SegmentNT correctly annotated this foreign DNA, it would provide compelling evidence that the model is using causal principles of genome architecture to make its predictions.

While we consider this an interesting follow-up experiment to our work, we think it is out-of-scope of the current manuscript given the extensive revisions that we have done to address all other reviewer comments.

21. p. 10 and in Discussion, it is argued that it is encouraging that the multispecies model still retained predictive performance for species whose genome underwent large-scale rearrangements. The authors should clarify their argumentation and explain why would ploidy and large-scale rearrangements affect the genomic code (eg, have to do with predicting splice sites, etc)?

We have added an argumentation for why this result is a positive indication to the results and discussion sections. Although the genomic code is not expected to change, as mentioned by the reviewer, such changes increase the sequence diversity of the different types of elements and should make it harder to predict, compared with genomes with less diversification to the ones where SegmentNT was trained.

22. In Discussion, it is stated ""Here we focused on a more challenging task of segmenting genomic elements in DNA sequences at nucleotide resolution." The authors should explain segmenting genomic elements more challenging than other (not named) tasks? In fact, computational biology have developped tools to annotate genomes way before gene expression models could be developed.

We have revised this discussion section to contrast genome segmentation of multiple elements compared with single tasks of classifying short sequences as containing a given type of element"

23. The Discussion should refer to concurrent approaches to transcript prediction based on experimental data (RNA-seq coverage): borzoi, bigRNA. If these methods get peer-reviewed and published they should be part of the benchmark.

We have added this notion to the discussion.

Typos

—-------

24. Everywhere: finetuning -> fine-tuning.
25. Figure 1b: U-Net in all caps
Done

26. Figure S4b,c do these constructs have indeed multiple mutations? Was the model applied to a set of mutations jointly?

Following the different reviewers' comments, we have revised the manuscript to focus more on the segmentation aspects, adding extensive analyses and baselines on gene annotation, splicing prediction and species generalization. As such, we have removed the analyses of this isoform mutation dataset and will leave the extension to mutations for follow-up studies.

27. Figure S4c the authors should denote whether the mutations correspond to a branch point mutation and donor and acceptor site mutation. These are fine examples but already well-modeled since decades.

See response above.

28. Figure 5a, “zero-shot generalization” is misused. The classes are the same than during training (exons, splice sites, etc.). Zero-shot would mean to predict an entire new class (say enhancer by only having learnt from promoters or exons.). The crux of zero-shot learning is to use some auxiliary information about the class [https://arxiv.org/abs/1707.00600] and this is not the approach undertaken here. Generalization to unseen species should be used instead.
29. Discussion: “... of segmenting genom*i*c elements”

Done

Reviewer #3 (Remarks on code availability):

I did not review the code at this stage given the major concerns overall. I do appreciate that code was available.

Thank you for the comment. It was important for us to make our work and models as widely accessible as possible.

Reviewer #4 (Remarks to the Author):

Overall, it is very nice to see a new, important application of pre-trained language models in genomics (and not just more benchmarks). Comparisons with spliceAI and generalization to other species are very interesting. Commend the authors for including randomly initialized baselines, but several of the comparisons need to be done more carefully (e.g., fine-tune the head only).

Although not related to the work here, thanks to some of the authors for the mRNA vaccine. It's an honor to review a paper from some of the people who helped to protect us from covid-19.

Thanks very much for your patience with the slow review turnaround. This was a very interesting paper with many areas for improvement so it took some time to think about it.

We thank the reviewer for highlighting the relevance of our work in demonstrating that pre-trained language models can be used to solve important applications in genomics. We also appreciate the positive feedback about the tasks we considered in this study, generalization to different species, and the different baselines we provide. Thank you very much for your words and for all the constructive feedback that allowed us to improve this manuscript further. We addressed all comments as explained below.

Essential

1. The authors need to explain more clearly how they are evaluating segmentations. Naively, MCC is not a great measure for segmentation because it's not comparable between segments of different lengths and treats each nucleotide as a separate prediction. (E.g., if half of the sequence is gene and the other half is non-coding, it's much easier to get a high MCC if there is only one changepoint than if there are 10 change points. Even more complicated issues if it's not half and half) If the authors have computed MCC using each nucleotide as a separate prediction (not explained in the text, but that's what I guess), they should present at least one other metric that is at the "region" level. Examples of this kind of metric are intersection over union (as in the original Unet paper), Dice coefficients, or (probably nicer to stick with bioinformatics) the classical Sov score that was developed for protein secondary structure segmentation (PMID: 10022357)

We thank this reviewer for pointing out that we did not explain sufficiently clearly the segmentation task of SegmentNT. SegmentNT is doing *instance segmentation* - computing a binary mask over entities for each "instance" (here instance refers to an element type and entities to nucleotides; see for example https://www.ibm.com/topics/instance-segmentation). As such, we have used typical metrics for instance segmentation treating each nucleotide as separate predictions. Following the reviewer's comment, we have now clarified our evaluation in the main text and methods. In addition, we have also added for all models their performance for three additional metrics: Jaccard similarity (intersection over union), area under the precision recall curve (auPRC) and the F1-score. Jaccard similarity addresses that precise question since it measures the intersection over union and it is very similar to Dice coefficients. All metrics showed consistent results (see Fig 1e and new Supplementary Fig 2). Many thanks for this comment that we think helped us improve the robustness of our results and manuscript.

2. Where are the error bars? Not a single performance measure is presented with confidence intervals/estimates of uncertainty. Given the variation in sample sizes for the different types of elements, estimates of performance for some elements must be more confident than others. As in any scientific study, all reported measurements should include estimates of uncertainty. I realize that it is commonplace in the computer science conferences and online machine learning competitions to achieve "state-of-the-art" by simply reporting a better number than the other methods, but in a serious peer-reviewed study, I believe we need to see estimates of uncertainty. (And connected to point 1, the error bars/uncertainty estimates cannot be based on the assumption that each nucleotide is independent. They must be based on the number of regions of each type. That is the real "sample size".)

We agree with this reviewer and have now added error bars and confidence intervals to all models and metric comparisons (see Fig 1c,e, 2a,b, new Supplementary Fig 2 and Supplementary Tables 2 and 3).

Regarding the metrics at "region" level, we do not understand the definition of regions used by the reviewer. As mentioned above, we have used metrics (and error bars/uncertainty estimates)

appropriate for instance segmentation, accounting for intersection over union and average precision. If the reviewer provides a mathematical definition of “region” and metric we would be happy to compute it.

3. The test of the segmentation method on species that are distant from the training data is fantastic, and in my view the most interesting part of the paper. However, the authors should clarify where the "ground truth" is coming from for these genome annotations. For well-studied model systems like arabidopsis, c elegans, e coli or drosophila will have high quality genome annotations that are based on numerous data sources and expert curations. If these are not included in the training data and are evolutionarily distant from the species in the trainin data, this is a true test of out of sample generalization. On the other hand, most other species are studied by fewer molecular biologists, and will only have annotations that are mostly based on bioinformatics predictions. These should not be treated as "ground truth." In these situations we are only comparing the nt-transformer segmentation to classical bioinformatics aprpoaches.

Thank you for this comment, which we agree is important to clarify the origin of the ground-truth labels used in this work. Indeed, some species have been extensively studied and have high-quality genome annotations, while others have annotations that are mostly based on bioinformatics predictions derived from the well-studied species. Since it was out-of-scope in our work to improve the ground-truth annotations for less-characterized species, we used the Ensembl annotations as ground-truth just to compare our models with concurrent approaches. We tried to reduce the impact of the quality of the annotations by using some of the well-characterized species for training and others for testing, while presenting the model comparisons across all other species as well. We agree it is important to mention this limitation that should be addressed in future works (see added it to page 13), many thanks.

4. The comparison to the Unet alone vs. the embedding + Unet is not a fair comparison because of the difference in tokenization. As far as we understand, the Unet alone is on a 4-word vocab, while the embedding + Unet is on a 4096 word vocab. Therefore, as far as we understand it, the number of parameters and the dimensionality of the models in not comparable. The authors need to redo this key control experiment where the UNet is trained on a 4096 word vocab.

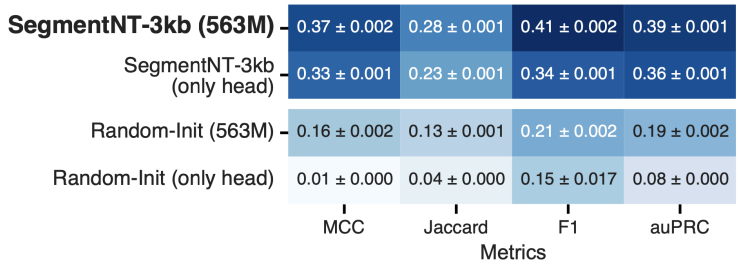
We have added the baseline with the UNet using the 1024 word vocab (the embedding dimension of NTv2, not 4096) and observed that the performance was very similar to the one using one-hot encoded sequences (4-word vocab). See new Fig 1e.

5. The random initialization is a great negative control, but it's not a fair baseline if they didn't let it converge. In addition, they need to 1) use "frozen" embeddings from nt-transformer and only train the segmentation head and 2) use a randomly initialized, but frozen embedder (untrained) and train the segmentation head. The performance differences on these experiments will show the importance of the pre-training task.

We agree and have now trained the randomly initialized model until convergence. When comparing both models we observed that self-supervised pre-training on genomes allows our

SegmentNT model to converge 7 times faster to an asymptotic performance twice greater across all 14 genomic elements: average MCC 0.37 compared with 0.16 for the version with a random initialized NT. We have added these and additional model ablations and baselines to Fig 1e and Supplementary Tables 1 and 2.

In addition, we have added the two additional baselines suggested by the reviewer which indeed nicely show the importance of the pre-training task. See the main results below and in Fig 1e and Supplementary Tables 1 and 2. Many thanks.



6. many overstated claims:
-The authors claims about 0-shot and few-shot learning are misleading. They use 0-shot and few-shot to mean generalization out of species or out of context length, which are both impressive, but are not 0- or few-shot in the sense that the term is used in the field. We believe that 0- and few- shot should only be used to describe performance on a task the model was never trained to do. Since they have fine-tuned the parameters of the embedder on the segmentation task, there is no sense in which this is 0-shot learning (as far as we can understand)..

-“We show improved performance on the detection of splice sites throughout the genome and demonstrate strong nucleotide-level precision”

The generality of this claim is not supported by the data presented. The performance of the model on detecting splice sites is similar to SpliceAI for protein-coding genes

-“We show that SegmentNT achieves high nucleotide accuracy for all elements” Not convincingly true, some elements like lncRNA are still very poorly segmented.

-“SegmentNT accurately detects splice donor and acceptor sites in both strands in any given input DNA sequence” Is not supported by the results. The performance of SegmentNT is high but not that high.

- “a new paradigm on how we analyze and interpret DNA” seems overstated. The authors should restrict themselves to claims that can support with evidence.

Thank you. We agree with these points and have revised these statements accordingly.

7. The attention to detail of the manuscript is at times poor, with many errors which inhibited understanding. These are too numerous to fully account for here, however the below list details some:
• In the abstract the authors claim SegmentNT can “generalize to elements of different human and plant species” which doesn't make sense because there is only one human species.
• Supplementary Fig 3 contains the exact same data as Fig 4b and 4c, despite the text indicating it should contain different values.
• The text refers to Supplementary Fig. 4d-g, however figures 4e, 4f, and 4g do not exist
• Figure 5a caption indicates the held-out test set contains 12 species, however the figure indicates there are actually 17 species. Also misspells “species.”
• Citations 28, 44, 53, 54, and 57 are missing key information. Some citations in the text not correctly ordered when within the same brackets.
• Section A.5 has details which conflict with those elsewhere (different model context length, does not indicate human annotations are included in the finetuning)
• First sentence of Section D.2 refers to the methods section despite being in the methods section

Thank you very much for this extensive revision. We have corrected all those points.

• The link to the Illumina Basespace platform cannot be accessed

This link is accessible after logging in in Illumina’s user website

• Y-axis of Supplementary figure 1b is incorrect
• Question marks in negative region of axes in Supplementary figure 4b) and 4c)
• Various instances of minor inconsistencies in wording across the manuscript. Approach sometimes referred to “SegmentNT” and other times as “Segment-NT”, x-axis and y-axis titles between adjacent plots changed despite showing the same metrics. Models given names not found elsewhere in first sentence of “SegmentNT generalizes to sequences up to 50kb” section.
• Many minor grammatical and formatting errors (e.x. Tetraodon spelt incorrectly in methods section, Time tree of life links not working, etc.)
A thorough re-reading of the manuscript must be performed to eliminate these errors.

Thank you very much for this extensive revision. We have corrected all those points.

Suggestions.
1. In my view the comparison with enformer is problematic. As the authors show, using binary models is not nearly as good as training the segmentation head (Unet) on top. Since Enformer was not trained on a segmentation task, to make a fair comparison the authors should use the bp-level predictions of all the tracks from enformer as a pre-trained representation and then train a unet segmentation head on top. This could also be evaluated on the out-of-sample "out-of-species" test. Otherwise, the authors should drop this analysis or at least acknowledge that it is not really a fair comparison. My sense is that the spliceAI results are strong enough to

make the case for the fine-tuned segmentation model. I don't feel that the enformer results (especially as presented) are necessary.

After seeing the different comments of all reviewers regarding the Enformer evaluation, we fully agree that it would be more interesting and useful to evaluate Enformer on the SegmentNT settings and have revised the manuscript extensively on this aspect.

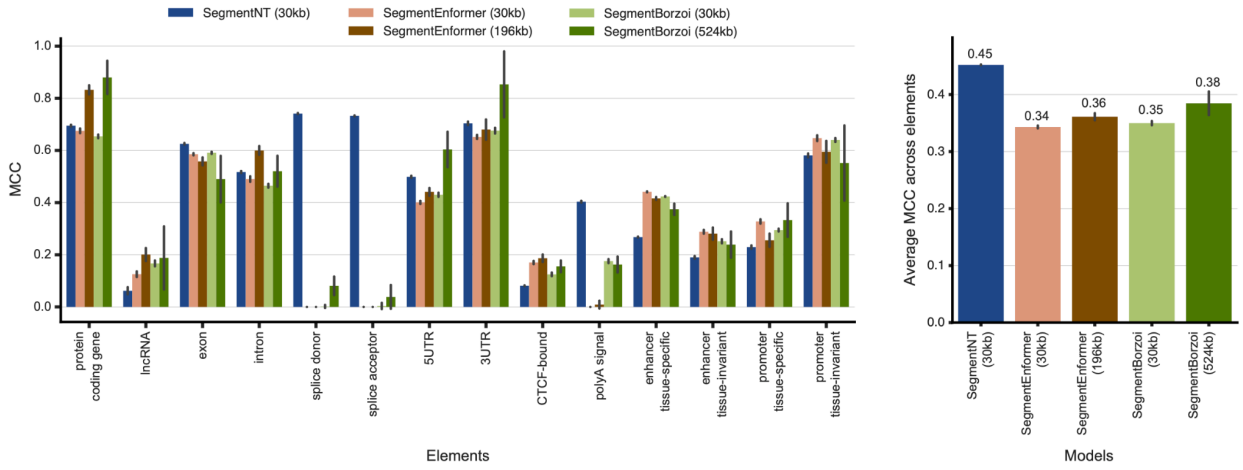
Our initial idea was to compare SegmentNT with published models that can be used to identify promoter and enhancer regions. Since we could not find any model that could be used with good performance for testing in a whole chromosome dataset (outside the standard curated datasets used for model train and testing), we explored how one could use Enformer predictions as surrogates of regulatory activity, which indeed demonstrated good performance. However, we agree that that is not the main use case of Enformer and these prediction results can be misleading.

Therefore, in this new analysis, we decided to investigate the Enformer under a new angle. As shown in our previous NT paper (Dalla-Torre et al 2023), the Enformer model can also be used as a powerful feature extractor. As our SegmentNT framework is general, we can train it using other feature extractors than NT as the DNA encoder. As such, we have now evaluated Enformer (and Borzoi) as competitive DNA encoders of Nucleotide Transformer by placing a UNet on top of the Enformer/Borzoi representations, calling them SegmentEnformer and SegmentBorzoi, respectively (see Methods). This also allowed us to test the performance of models that take longer context windows as 196kb (Enformer) and 524kb (Borzoi). In addition to taking longer context, both models have been trained on chromatin accessibility, transcription factor binding and gene expression data and as such have probably learned to extract features that would be useful to better annotate some types of elements such as the regulatory elements. Note, however, that while NT has been trained to produce representations over 6-mers, the Enformer and Bozroi were trained respectfully at 128bp and 32bp resolution. Therefore, we adapted the U-NET module to upsample accordingly and we might expect a loss in performance over the detection of small elements.

We finetuned the whole Enformer/Borzoi+UNet networks on our segmentation dataset using either the same 30kb input sequences or the model's original input length (see main results below). In summary, this showed that, as expected, Enformer and Borzoi allow improved performance on regulatory elements but cannot solve high-resolution tasks such as the prediction of splice sites and polyA signals. On gene elements they have lower performance when using 30kb sequences as input, but the performance on some elements increases when using their longer-range inputs - namely protein-coding genes and introns, longer elements that previously benefited from increased sequence length (see Fig 2a). We have added these results as new main Figure 3.

We think this result showcases how the SegmentNT framework can also be extended to additional foundation models to achieve longer sequence contexts and superior performance on particular genomic elements. We will also release these two additional models,

segmentEnformer and segmentBorzoi and believe that they will be useful to the community especially to researchers interested in longer context annotation or regulatory genomics. Many thanks for this suggestion!



2. If I am reading the paper correctly, the authors did not use the "largest" of the nt transformer models (in terms of parameters and training data). Seems like they only used a 500M multispecies (this means the model has seen some of the other species genomes? so it is not really held out data?). This should be discussed/explained, especially in the context of the current thinking around "scaling" of "foundation models." If using larger models does not improve performance, this is an important (negative) result about the scaling of current MLM pre-training. If the larger models are just too "big" to use in practice, that's also an important result on the practical side for the computational genomics community.

We did not use the largest NT model from the first version, with 2.5B parameters, but instead used the largest model of the second NT version (NTv2 500M parameters), because that was the most performing model in the NT benchmark (see the NT paper for more details about their architectures). We have added this reason to the Methods.

Regarding the models “having seen some of the other species' genomes”, this should not be an issue since they never saw any labels for those genomes. For the downstream segmentation task, the labeled data from the held-out species was never seen during training. Indeed, that’s the benefit of having a pre-trained phase on unlabeled data - building strong sequence representations independent of labeled data that can be leveraged during fine-tuning. The same happens in other domains like computer vision with models pre-trained on general datasets of unlabeled images and then tested/applied on datasets with labeled images, some very related to the ones used during pre-training. This is a common approach when building and using foundation models across domains.

Minor essential issues.

1. The have provided a Method that can segment a genome. However, code they provided for segmenting the genome in their README.md does not run because of a syntax error (we found the bug and were able to correct, but this is not really our job).

We apologize for this syntax error and have already corrected it. Still, in addition to the code provided in the README, we have also at the time added an example notebook to reproduce model inference and results that hope it is useful (available at https://github.com/instadeepai/nucleotide-transformer/blob/main/examples/inference_segment_nt.ipynb).

2. We cannot replicate the paper's findings because the methods are not described clearly enough (and no codes are provided).

-the output shape in the figure (L x 6 x 14) contradicts what they write in the methods section A.1 (N x K x 2) which we believe is N = 6 x L and K = 14 but there is an extra factor of 2.

We have updated the figure cartoon and added more details to the Methods.

-For human the test and validation set chromosomes are indicated, but not for the other 5 species used to train the multispecies model. This is necessary information for replicating the results.

We have added this information in the respective methods section.

-we cannot understand the experiments where they compared to their binary promoter classifier or to the enformer enhancer predictor. There are so many issues that we can't follow, we will not be able to list them all here.

Thanks for the feedback. We have revised the analyses and methods descriptions (see page 25). We hope it is more clear now.

e.g., "We note that for Enformer we had to pad the
10kb sequences for inference, since the original model predicts scores for 114,688bp."
To our knowledge Enformer takes as input 196,608. We've been padding our sequences to that size. Can you please clarify what you are actually padding your sequences to.

As suggested by the reviewers, we have removed this analysis and now benchmark Enformer on the full segmentation task

e.g., choice of 10 nt window is arbitrary and unjustified.

This was used for a compromise between high resolution but limited number of windows per genome region. We have added this justification to Methods (see page 25).

-there are also problems with the comparison to spliceAI

E.g., the numbers in the text do not match the numbers in supplementary figure 3.

e.g., It's hard to understand why they did the experiment of limiting to 10Kb test sequences.

We have corrected these issues and expanded the comparison with SpliceAI and Pangolin to sequences of 30kb.

-Figure 1a has a labeled detection threshold, however such thresholds are not referred to elsewhere in the manuscript. The threshold needs to be defined in the manuscript or it cannot be reproduced.

We have added the threshold in the figure and in the main text.

-inconsistent use of MCC and PR-AUC in different analyses. We wonder if they are hiding results? OR the paper was done in stages and not well organized. The authors should be consistent.

We have now reported MCC and PR-AUC metrics for all analyses throughout the manuscript and supplementary material.

Summary of the revision:

We are grateful to the three reviewers for the positive assessment of our work. We have revised the manuscript to address their remaining comments and suggestions and improve the overall clarity. We have detailed all changes in our point-by-point response below.

Reviewer #1 (Remarks to the Author):

The authors have addressed all my concerns thoroughly. The revisions, including expanded benchmarking across relevant tasks, stronger baselines incorporating state-of-the-art tools, additional supervised embedding methods, improved data splits to prevent data leakage, and better contextualization within the field, have significantly enhanced the rigor and clarity of the manuscript. I believe it now meets a very high standard and will undoubtedly be impactful for the broader field.

We thank the reviewer for this comment and for appreciating our effort to address all reviewers' comments to the best of our knowledge. Many thanks for the helpful revision.

Reviewer #1 (Remarks on code availability):

I checked to see the organization of the repo and found it to appear quite complete with thorough documentation, though I did not attempt to install and run it.

Thank you for the comments about the code, very appreciated.

Reviewer #3 (Remarks to the Author): -------

General comments
The revision is strong and has addressed most of the concerns. The following points shall still be addressed. We are reviewer 3. We first pick up on a point by reviewers 1 and 2.

We are happy to see that most of the reviewer's concerns were addressed. We address the remaining ones below.

Reviewers 1 and 2:
SpliceAI-10k's performance in the original study was 0.98 PR-AUC. In this paper, it is 0.96 for acceptor and 0.94 for donor. Why is there a discrepancy?

Response: This discrepancy is due to small adaptations in the dataset to allow a direct comparison with SegmentNT settings. Namely, we consider sequence windows of 30kb (instead of 10kb in the original publication) and compute the predictions on the whole window (instead of not predicting for the flanking +/- 5kb sequence). We have also removed sequences with Ns due

to the constraint of the SegmentNT architecture. We have added these details in the Methods section (see page 26).

Reviewer 3: The authors may wish to modify SpliceAI or its usage to better match characteristics of their model (input and output size). However, reporting the performance of a tool when, instead, a modified version of the tool was applied is very misleading. The manuscript may include this modified SpliceAI but should also report the original SpliceAI performance.
Relatedly, the new Fig. 5a shows Pangolin and SpliceAI predictions outside of transcript boundaries. This is misleading as this is outside of the application domain of these tools. As mentioned in another point, the authors want to show that predicting splice sites genome-wide is hard because best-in-class tools have been developed for predicting splice sites within transcript boundaries. Therefore, these tools are suboptimal for de novo annotation of genomes for which transcribed regions are not experimentally determined or accurately predicted. This point is valid but confusing because Pangolin and SpliceAI false positives outside their application domain are conceptually different than those found within. To avoid confusion, the authors could highlight in Fig 5a the gene boundary (e.g. with a grey background outside the transcribed region) and add to the legend that SpliceAI and Pangolin have not been trained for out-of-transcript-boundary applications and make further false positives when applied there.

We thank this reviewer for explaining in detail this concern and for giving helpful suggestions to address it. We agree with the different points and have now made clear the areas where the comparison with SpliceAI is outside their training domain. Namely, we now also report SpliceAI's original performance for comparison, highlight in Fig 5a mispredictions within or outside gene boundaries, and discuss the differences between methods in the main text (see Fig 5a and page 11). Many thanks.

Reviewer 3 (this reviewer):
3. Statistical significance testing is largely absent. The authors should perform appropriate statistical tests to determine whether the reported performance differences between models are statistically significant.

Response: We agree with this reviewer and have now added error bars and confidence intervals to all models and metric comparisons [....]. These results provide stronger support to our conclusions, many thanks!

Reviewer 3: Statistical assessment was more often performed in this revision. However, the authors shall i) take reviewer 4's point 2 into account about bases within regions (e.g. in one annotated exon, one annotated intron, etc.) problematically not being i.i.d. and ii) perform statistical assessments systematically. There are missing statistical assessments in new Fig. 4, new Fig. 5, and new Fig. 7i. Also, the Statistics section of the reporting summary shall be revised. Statistical testing, sample size reporting, etc. are not "not applicable".

We thank the reviewer for this comment. We have i) added region-level metrics (see response below), ii) added the missing statistical assessments in the mentioned figures and iii) revised the reporting summary. Regarding the statistical assessment, we note that results in figures 4a,c,f,g and 5b are on data from published benchmarks and as such we followed the same evaluations therein, which does not allow us to add error bars.

13. Regarding the mutagenesis screen for splicing, the authors should compare to other models including Pangolin, SpliceAI and MMSplice - the latter having been developed for this.

Response: Following the different reviewers' comments, we have revised the manuscript to focus more on the segmentation aspects, adding extensive analyses and baselines on gene annotation, splicing prediction and species generalization. As such, we have removed the analyses of this isoform mutation dataset and will leave the extension to mutations for follow-up studies.

Reviewer 3: Variant effect prediction can indeed be considered out of the scope of a genome annotation method, whereby exploiting correlative signals may be very effective. Consequently, the following statement of the discussion should be replaced by a statement about future work to evaluate the capacity of this approach for variant effect prediction:
" Third, the accuracy of SegmentNT predictions can be leveraged to evaluate the impact of sequence variants on the different types of genomic elements, as we showed for splicing isoforms."

We agree and have revised this sentence, thank you.

Related to this, the last word of this sentence is wrong:
"In addition, the SegmentNT-30kb-multispecies model showed improved performance over its human counterpart for the prediction of splicing variants (Supplementary Fig. 6d)."
It should be:
"In addition, the SegmentNT-30kb-multispecies model showed improved performance over its human counterpart for the prediction of splice sites (Supplementary Fig. 6d)."

We have corrected this sentence.

Reviewer 3: Minor Points
16. The name of the tools is a misnomer. Segment-NT is not a segmentation algorithm. It is a single-base multitask classification. Proposing a segmentation algorithm could have been methodologically innovative and may be a safer direction at this stage that the authors could investigate.

Response: We would like to respectfully disagree with this reviewer's opinion. SegmentNT is indeed aligned with segmentation principles - with segmentation being simply the task of assigning a class label to every single pixel of an input image for computer vision, or assigning a class label to every nucleotide of a DNA sequence in this genomics context (see for instance

book “Digital Image Processing”, Rafael C. Gonzalez & Richard E. Woods, 2018). More precisely, SegmentNT is doing instance segmentation - computing a binary mask over entities for each "instance" (here instance refers to an element type and entities to nucleotides; see for example https://www.ibm.com/topics/instance-segmentation). While this is a form of multitask classification, it effectively segments DNA sequences by assigning functional labels to each nucleotide, predicting the boundaries and labels of genomic elements at single-nucleotide resolution. We will clarify these points in the paper. If the reviewer is asking for us to perform semantic segmentation instead (attributing one label to each entity), this would limit us as we would have to remove some element types; for instance we couldn't have both protein coding genes and introns/exons.

Reviewer 3: Thanks. Other fields, other definitions. A segment, geometrically, is a part of a straight line bounded by two distinct endpoints. It is commonly understood that segmenting a genome amounts to partitioning it into intervals (other example: https://academic.oup.com/bioinformatics/article/22/16/1963/208107). Some classes of algorithms more naturally favour such outputs, where adjacent bases tend to be of the same class. One example is Tiberius, the deep learning follower of AUGUSTUS, which employs an HMM to this end.
This point is not as academic as it seems. It relates to another miscommunication regarding statistical assessment with Reviewer 4. Reviewer 4 underscores in their points #1 and #2 that adjacent bases are not independent and, therefore, that statistical assessment shall not be done at the base level. To address this, the authors should use ground truth regions (e.g. a ground truth whole exon) as one instance instead of considering all bases of a region as independent observations.

We appreciate the reviewer’s constructive feedback and agree that we should complement our manuscript with a region-level metric. Taking into account reviewer 3 and 4’s comments on this, we have now added a new metric that evaluates the correct prediction of segments - namely the classical segment overlap score (SOV) that was developed for protein secondary structure segmentation (as suggested by reviewer 4). See Methods section “Model training and evaluation” for the implementation, taken from reference: https://scfbm.biomedcentral.com/articles/10.1186/s13029-018-0068-7). We display all results in our new figures 1e, 4f-h, S2g and S9 for results. We observed similar results between nucleotide- and region-level metrics strengthening our conclusions. Many thanks.

New minor points:
2) P.9 the main ioform -> the main isoform
3) The discussion should now mention concurrent work published since initial submission, notably Tiberius (Gabriel et al. Bioinformatics 2024)
4) P. 10 “Pangolin compared with SpliceAI on non-coding RNA splices” -> splice sites.

We have corrected these sentences and added Tiberius in the discussion, many thanks.

Code:

We tested the HuggingFace code which worked. It should somewhere be mentioned that the HuggingFace uses FlashZoi, a different implementation of Borzoi and slightly different model, than the Jax implementation upon which the results are based.

Thank you for testing our code and giving this feedback!
We would like to note that we did use the published Borzoi model (from Keras) and ported it directly to Jax, not the one from FlashZoi, for all the analyses in our paper. For porting it to HuggingFace we refactored the original implementation in keras to a format compatible with our other huggingface model. Here we did use the FlashZoi implementation for the common parts to the published model for easier implementation, but changed it accordingly to the published version when needed - such as the positional embeddings in the attention layers. All this was verified by obtaining an absolute difference of less than 1e-5 in all activations as well as in the predictions between the original implementation and ours. We have added that note on the HuggingFace repo.

Reviewer #4 (Remarks to the Author): -------

We appreciate that the authors have done a lot of work and responded to nearly all of our comments. We especially appreciate the convincing demonstration of the value of pre-training and the exploration of model mis-predictions. Overall, the manuscript is significantly improved. However, we believe there are still some key issues that need to be addressed before the manuscript can be published.

We would like to thank this reviewer for the thorough first revision that helped us improve the manuscript significantly, and for acknowledging our work to address all comments. We address the remaining comments below.

-The code for segment-enformer and segment-borzoi is not available on their huggingface space or github.

We have corrected this during the revision. See updated comment of this reviewer at the bottom.

-We are grateful that the authors decided to address the incorrect use of “the term 0-shot” by removing it, but we still find the term in the figure legend of Figure 2, where the authors report zero-shot generalization of segmentNT. We emphasize that including these “buzzwords” will only serve to further sow confusion in our field. Please avoid. Similarly, use of the term “foundation model” has been recently criticized by Yun Song’s group when referring to genomics models (see “Genomic Language Models: Opportunities and Challenges” https://arxiv.org/html/2407.11435v1). We also suggest that the authors avoid this term in their work and focus on their actual findings related to genome annotation.

We have addressed your concerns by revising the legend of Figure 2 to remove the “zero-shot” term for greater clarity. Regarding the term “foundation model,” we acknowledge the recent

discussions around its use in genomics and have updated the manuscript accordingly. Specifically, we have provided a clear definition of what we mean by "DNA foundation model" in this work in the introduction (see page 2).

-The authors still have not presented a region-based segmentation evaluation statistic appropriate for segmentation. Several were suggested in the first review, but the authors write:

“Regarding the metrics at “region” level, we do not understand the definition of regions used by the reviewer. As mentioned above, we have used metrics (and error bars/uncertainty estimates) appropriate for instance segmentation, accounting for intersection over union and average precision. If the reviewer provides a mathematical definition of “region” and metric we would be happy to compute it.”

First, authors appear to misunderstand the concept of “instance segmentation.” If we take the analogy with images, their classification of individual nucleotides into 14 categories can be considered a multi-label “semantic segmentation.” Instance segmentation would distinguish each specific CDS or UTR. The authors need to correct this confusion in the final manuscript. https://en.wikipedia.org/wiki/Image_segmentation

Second, the authors’ confusion about region-level metrics is analogous to confusing pixel-level metrics and object-level metrics in image segmentation, which is shown as a pitfall in Figure 1a of this review: “Metrics reloaded: recommendations for image analysis validation, Nature Methods volume 21, pages 195–212 (2024)” The authors claim to be segmenting multiple objects of different sizes (e.g., exons, UTRs, splice sites, etc., etc.,). Therefore, an object-level metric for the different regions would give an actual estimate of segmentation performance. As it is now, the authors have only given nucleotide level classification performance. A region-level metric would estimate the agreement between the nucleotides within the predicted segment and the known boundaries. For example, if CDSs go from X:200-300 and X:500-505, and the predictions are from X:190-310 and X:490-515, the intersection over union at the region level is (0.83 + 0.20) / 2 = 0.52. However, if evaluated at the nucleotide level as the authors are doing, we get 100 + 5 / 120 + 35 = 0.68. This shows how pixel/nt-level metrics are strongly biased towards the performance on larger objects/regions. Also note that the true sample size is n=2 exons, while at the nt-level we have n=155. The nucleotide-level metric will be associated with greatly over-estimated confidence (biased variance estimates).

We appreciate the reviewer’s constructive feedback about the various terminologies regarding segmentation, and the additional explanations about region-level metrics. We agree with the reviewer and have revised the manuscript to explicitly say that SegmentNT can be considered as multi-label “semantic segmentation” (see page 3).

In that regard, we agree that we should also complement our manuscript with a region-level metric, as also reinforced by reviewer 3. Taking into account the reviewer’s comments, we have now added a new metric that evaluates the correct prediction of segments - namely the segment overlap score (SOV) that was developed for protein secondary structure segmentation

(as suggested by the reviewer). See our new figures 1e, 4f-h, S2g and S9, and the methods for more detailed description of the SOV score (we used the implementation from https://scfbm.biomedcentral.com/articles/10.1186/s13029-018-0068-7). We observed similar results between nucleotide- and region-level metrics strengthening our conclusions. Many thanks.

-We thank the authors for doing additional baselines to assess the value of pretraining. However, the authors claim to use a 1024-vocabulary, but this is not true. In fact, they are using a 1024 embedding dimension of the 1-hot encoded DNA. The authors should correct this in the manuscript. (To make an actually fair comparison, they could tokenize the DNA into a 6-mer vocabulary because that is the input to the nucleotide transformer. This would also address the difference in length of tokenized input (L vs. L/6) for this experiment.)

We appreciate the reviewer's observation and have corrected the vocabulary description in the manuscript (see Fig 1e and Methods page 23). Regarding the suggestion for an additional ablation study using a 6-mer vocabulary, we understand the intent to provide a more direct comparison. However, the methods we benchmarked against are designed for single nucleotide inputs, not 6-mers. Therefore, while the number of tokens varies between architectures, the actual sequence length in nucleotides—which is the biologically relevant metric—is conserved across our comparisons. See for example SegmentEnformer and SegmentBorzoi that work with single-nucleotide as well, similar to the baselines and ablation models we tested.

Given the extensive benchmarking and ablation studies already conducted and presented in the revised manuscript, we believe that pursuing this particular ablation at this stage would not significantly enhance the core findings.

We have added a clarification to acknowledge the difference in tokenized input length and explain why we focused on single nucleotide comparisons, aligning with the methods we benchmarked against. We believe this addresses the reviewer's point while maintaining the focus on the comprehensive analysis already provided.

We still noticed many cases where detail is lacking and mistakes have not been corrected:
1. We thank the authors for removing the chunks with homologous genes from their test data. The authors should clarify whether this also removes homologous distal regulatory elements, or whether homology could still be inflating performance on these regions.
2. Similarly, the authors “sample” the test set 10 times to obtain estimates of confidence, but the sampling strategy is not explained. Ideally these are non-overlapping partitions.
3. There are also minor writing issues like English spelling errors (inclused vs. included, ioform vs. isoform, etc), surnames in references missing diacritics, e.g., Tomáš Brůna), and the authors refer to “matched U-Net connections” but this is not a term we are familiar with.
4. The authors write “Overall, SegmentNT detects splice donor and acceptor sites in any given input DNA sequence with high accuracy” but the accuracy on non-coding RNAs is still very low (supplementary figure 6, panel c shows auPRC is approximately 0.02). This needs to be corrected.

Many thanks for the detailed comments. We have added missing details and corrected the remaining mistakes.

5. Figure 1 panel b has not been corrected (length L should be N).

In this case the panel b is correct since we use N as the number of nucleotides in the DNA sequence and L as the number of DNA tokens (with roughly $L \approx N/6$) (see Methods). In that case, the input to the model is L and not N. We have clarified that in the legend.

Reviewer #4 (Remarks on code availability):

The codes for all the models were available, although run inference_segment_borzoi.ipynb and inference_segment_enformer.ipynb did not run as expected on colab. Codes should be checked more carefully.

Thank you for testing our code. We noticed a small typo in the tokenizers that were being used which led to different predictions. We have fixed it now and made sure that the code works as expected.

Summary of the revision:

We thank all the reviewers for the suggestions and rigor throughout the different revision rounds. We have revised the manuscript to address their remaining comments. We have detailed all changes in our point-by-point response below.

Reviewer #3 (Remarks to the Author):

The authors have addressed most points satisfactorily, with the exception of the statistical assessments in figs 4a,c,f,g and 5b.
I remain puzzled as to why it is not possible to provide at least a basic statistical evaluation of the observed trends—if nothing else, the variability in their own data should be quantifiable. The explanation that the analysis follows "published benchmarks" is not convincing. If previous studies did not include statistical assessments, that should not preclude this manuscript from meeting basic scientific standards.
I trust that the editors will be more patient and can enforce this fundamental requirement. I do not need to review the revised manuscript again.

We thank the reviewer for the comment and understand the concern. We have now added the statistical assessments in the missing Figures 4a,c,f,g and 5b, as well as Supplementary Figure 6a. We explain the statistical tests in the respective legends and methods. Thank you for the feedback that allowed us to improve the manuscript.

Reviewer #4 (Remarks to the Author):

The authors have addressed all of the issues, and have finally added the SOV score as a region-level metric. Unfortunately, there are still two issues. Thanks to the authors for their patience through so many rounds of reviews.

1. They have implemented the SOV in a perplexing way. They wrote:

"The SOV score is sensitive to the predicted threshold and focuses primarily on true positives. To offer a more balanced evaluation, we computed the SOV score for both positive and negative regions, and averaged these two SOV scores to obtain a robust measure of performance."

Can the authors give a citation/precedent for computing an evaluation on the negative regions? For highly imbalanced data, such as genome annotations, the negative regions are most of the genome for most of the classes. Finding the regions of the genome that are not exons (or enhancers or UTRs) has no biological use-case, and is computationally 'easy' since a predictor that predicts all negatives will score very well at this task.

The authors should simply report the actual SOV score (or IoU or any standard *region-based* segmentation measure). If they need to argue that it needs be modified for their problem, they

can present that as a supplementary analysis along with clear explanation/justification for why they need to modify it in the context of their use-case.

We thank the reviewer for pointing out that it is preferable to still use the standard SOV score for highly unbalanced data. We have revised the Figures 1e, 4f-h, S2g,h and S9b, and respective results and methods sections to incorporate the standard SOV score. All the results and conclusions remained the same. Many thanks.

2. The authors have not discussed the ablation results for the SOV:
"UNet large one-hot (252M)" obtains SOV nearly as high as the best performing models. Does this mean that most of the segmentation accuracy can come from the UNet head? If so, the authors should comment on this interesting finding.

This result was misleading due to the previous implementation of the SOV metric. With the standard metric we now observe a significant improvement of SegmentNT over the "UNet large one-hot (252M)" model as well as the other baselines. Thanks for pointing it out.