nature portfolio

Peer Review File

The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics

Corresponding Author: Dr Marie Lopez

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

20th Jun 2023

Dear Dr Lopez,

Thank you very much for your patience when your Article, "The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics" is sent out for peer review. This paper has now been seen by 2 reviewers. 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 to fully address all these concerns. Among other revisions, it is important to substantially improve the evaluation and benchmarking of the models, and show their advances and benefits over existing models/methods. 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 4 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:

Reviewer #1:
Remarks to the Author:
The authors have proposed a deep learning model, the Nucleotide Transformer, for unsupervised pre-training of sequence models, and applied it to various prediction tasks. This study also assessed the impact of using diverse genome datasets for pre-training transformer models on DNA sequences. The authors trained multiple models on human reference genomes, 1000 genomes, and 850 genomes from other species like fungi, invertebrates, etc. The authors demonstrated that this approach matches or outperforms the baselines on 12 out of 18 tasks and up to 15 after fine-tuning. The method demonstrated similar performance to DeepSEA or DeepSTARR when fine-tuned on corresponding datasets. Although the authors did a commendable software engineering job, making it possible to train such large models, there are some weaknesses that need to be addressed. These include a lack of strong evidence for advancing over the current state-of-the-art, as well as demonstrating the advantage of training on diverse genomes and unsupervised pretraining, as I detailed below.

Major points:

1. In terms of novelty, the nucleotide transformer approach is very similar to DNABERT. Although the Nucleotide Transformer model is larger and trained on sequences beyond the human reference genome, it is still unclear whether its performance is better. Therefore, a comparison with DNABERT is needed.

2. The authors reported that their model matched or outperformed 15/18 baselines on performance. However, it should be noted that these comparisons do not include many state-of-the-art models. Furthermore, the model generally did not surpass stronger baselines, DeepSEA and DeepSTARR, and lacks comparison with current state-of-the-art methods.

For instance, the authors mentioned the Enformer but did not compare the results of the Nucleotide Transformer with those of Enformer in predicting DNASE/ATAC-seq, CAGE, Histone modifications, etc. Similarly, Sei is a recent model reported to outperform DeepSEA. Given the comparison between Nucleotide Transformer and DeepSEA, it is likely that Sei will outperform Nucleotide Transformer on these tasks. Other examples of state-of-the-art models include SpliceAI for splicing variant prediction and BPNet for basepair resolution Chip-Nexus predictions. While it may not be necessary to compare the Nucleotide Transformer to all these methods, some comparison would be needed.

3. There is a lack of evidence that multispecies training or training on multiple genomes provides an advantage in performance after considering issues in the comparisons. The main problem with this comparison is that models trained on human genomes should not be evaluated on non-human datasets (most of the 18 datasets contain some non-human data). Most notably, the performance advantage of the multi-species model in Figure 2a mostly comes from yeast histone mark datasets. Given the significant difference in biology between yeast and humans, it is not very meaningful to compare human models using yeast datasets. The observation that human genome-trained models do not perform well on non-human datasets does not constitute a strong argument for the advantage of the multi-species model.

If we exclude the yeast datasets, there is no obvious advantage from the model trained on multi-species genomes. If we exclude other non-human data, it is not unlikely that no advantage can be observed. Additionally, it is unclear whether a 1000 genomes-trained model outperforms the model trained on only human genome data for models with the same number of parameters, if we exclude the yeast datasets.

It is also preferable to use whole chromosome-based holdouts similar to what the authors used for variant effect prediction comparisons, rather than 10-fold cross-validation without taking spatial adjacency into consideration.

4. The authors have not shown strong evidence that unsupervised pre-training improves performance. It should be noted that, unlike language modeling, which has a vast amount of unlabeled data, the human genome is well annotated, and most datasets have genome-wide coverage. Therefore, the benefit of unsupervised pre-training cannot be assumed. This work showed that the pre-trained model matches but does not outperform previous models trained in a supervised fashion, especially for stronger baselines trained on bigger datasets. This still leaves open the question of the benefit of unsupervised training. One potential opportunity for unsupervised pretraining is, maybe it can improve performance when a limited number of positives in training data are available. It is also notable that a recent preprint (https://www.biorxiv.org/content/10.1101/2022.08.22.504706v2) showed that an unsupervised model can predict variant effect on species without experimental data.

5. For the evaluation of reconstruction accuracy on variants, if I understand correctly, does not seem to be the appropriate setup. Since the model trained on 1000 Genomes can simply memorize allele frequencies, showing that it reconstructs variants from a different dataset does not necessarily show that the model learned to use sequence in a non-trivial way to predict variants. Evaluation on holdout chromosomes is better for this evaluation.

Minor points

1. I am not sure if the maximum attention percentages in some layers reaching 100% for enhancers is particularly meaningful, as all element types seem to have some layers reaching 100%.

2. Comparison with DeepSEA in Figure 4d does not seem appropriate. For example, the nucleotide transformer was fine-tuned on each dataset, but DeepSEA was not. DeepSEA is also not intended for predicting coding mutations or splicing variants, which are the majority of mutations in ClinVar or HGMD.

3. It seems that the definition of p_a(f) should be the proportion of attention scores greater than, rather than less than, the threshold. The statistical test for p_a(f) seems to be not very meaningful, given that any threshold can be significant. Given the sample size, I don't think a p-value is needed for this analysis.

4. More information could be given on how the 1000 genomes dataset was constructed. It should also be noted that the reconstructed genome does not contain all variants.

Reviewer #2:
Remarks to the Author:
In their paper, the authors introduce the Nucleotide Transformer as a foundational model for genomics. The authors describe the Nucleotide Transformer, pitched as a foundation model for genomics. Nucleotide Transformer follows the BERT architecture used for training large language models (LLM). Nucleotide Transformer models are trained in a way similar to BERT, asking the model to predict masked tokens, in this case, 6-mers. This idea is that sequence embedding learned from NT can be used for downstream genomic tasks with probing or fine-tuning.

Building on the success of LLMs, there are recent a deluge of papers on similar ideas, building foundation models for genomics, utilizing sequence only information such as DNABert, BIGBERT, GenSLM, GPN (from Yun Song's group),… or a combination of sequence or functional annotations such as Enformer, BPNet, …. So neither method nor idea is new here. Instead, the authors focus on exploring the impact of sizes of the model and diversity of the datasets. Specifically, authors leveraged genomics sequences from multiple species and attempted to show the merit of training from diverse species.

While potentially interesting, the presented work is an incremental contribution building on existing works. There are major concerns about the design of the model, as well as the validity and effectiveness of the conclusions drawn from the papers.

My major comments are:

At the most basic level, I am not fully convinced that a naive application of BERT type of models, developed for NLP, to DNA sequences is a productive approach. In NLP, the tokens have meanings and are naturally defined. It is unclear what "token" means in the context of DNA sequence. There is no such thing as simple functional units in DNA, especially for noncoding sequences. The human genome is littered with many different kinds of functional elements buried in a sea of nonfunctional sequences. Other than coding sequences or RNA genes, it has been very difficult to properly define and model other types of functional elements. Demarking the boundary of non coding functional elements is even harder. There is also the issue of "grammar" or "rules" with the DNA sequences. Despite years of research, it is still not clear what they are and how general the rules are. So I would like to have some understanding of what kind of knowledge is learned by the Nucleotide Transformer. It is important for authors to look deeper into the interpretation side to find out what their models have learned. Is that information supported by prior biological knowledge?

One of the main points that the authors argue for is the large models with millions of parameters. Big models have shown utility for LLMs. But I am not convinced it is a good idea for training models for DNA sequences. Language or vision are compositional. The number of training samples is unbounded. This kind of diversity is crucial for training LLMs. By contrast, the diversity associated with the human genome is limited, even accounting for genetic variants. With only 3 billion letters in the human genome, a large model with hundreds of millions of parameters can simply memorize the genome. Can authors provide

a convincing argument and experimental support on why larger models are better for DNA sequences?
There is insufficient review of prior works and discuss how the current work differs from prior ones. Authors may improve their introduction by briefly mentioning existing task specific (Enformer, BPNet, …) and general nucleotide language models (DNABERT, …). Then, they can describe their added values compared to their proposed model. For example, if their model size or multi-species training is the major difference from state of the art models. Additionally, authors should compare their contributions against state of the art by conducting further benchmarks in the result section to emphasize the added values of their work.

I have some concerns regarding the evaluations and experimental results. First, evaluation tasks should be expanded to include recent functional genomic data from ENCODE or Epigenomic projects. The authors used the dataset of DeepSEA, which was curated several years ago. There are more recent and richer datasets that the authors should evaluate. Second, the improvement of their models compared to the so-called “SOTA” are marginal in most cases. So this raises the important question on the value of the proposed models. This is also in stark contrast to the results from LLMs, where foundation models offer more substantial improvements.

Authors utilized 6-mer based tokenization. In natural language modeling, word-based tokenization is optimal because words are the smallest meaningful units constructing a sentence. However, DNA grammar does not contain units like words and instead contains regulatory motifs. Earlier genomics models prior to deep learning such as gkmSVM [1] and kmer-SVM [2] k-mer based methods are widely used. However, convolution based models have become the standard because convolutional deep learning models such as DeepSea, Basenji, or Enformer have superior performance. because convolutional layers learn to tokenize regulatory elements from raw DNA sequences. In the manuscript, the authors do not provide any justification for 6-mer based tokenization over the use of convolution. Their architecture choice has major computational cost and limits the receptive field of the model to 6,000 bp.

Authors need to compare alternative architectures such as a model with convolution layers followed by attention layers. Thereby, they may increase the receptive field and reduce the model training cost. If their current architecture outperforms the alternative architectures then their model choice would be justified. Also, masking-based semi-supervised training is compatible with convolutional models; for example, authors may utilize intermediate layer masking after convolutional layers but before attention layers.

SVM bases models have a number of ways similar to final probing layers of the proposed model. Thus, authors may compare probing and gkmSVM despite the method may not be the state of the art.

The state of the art models repeatedly show that increasing the model receptive field significantly improves the performance because sequence context is critical to understand DNA grammar [7, 8]. Their 6 kb sequence length is smaller compared to state of the art models in genomics. Authors may try different lengths such as 1kb, 2kb, .etc, and show that model performance converges with 6 kb. Alternatively, Authors may increase the receptive field with convolution as discussed in the previous comment.

Authors claim that an increasing number of parameters in the model is helpful; however, authors haven’t performed benchmarks based on different parameter numbers so they did not demonstrate this point. Moreover, state of the art models such as Enformer contain significantly far fewer parameters. Also, authors may show their large models are helpful by benchmarking against state of the art models such as Enformer.

Language modeling methods for DNA such as DNABert and BigBERT were previously proposed; yet, a comparison against those models is lacking in the paper.
The overall benchmarking is not sufficient. Methods they are benchmarking against are not SOTA. We suggest that the following methods need to be benchmarked. Authors should compare their model against SpliceAI for splicing-related acceptor and donor prediction. For transcription factor binding, compare to newer methods besides DeepSEA: Catchitt, DeepGRN, FactorNet,Anchor, and compare histone mark prediction against Basenji, and BPNet. Also, they can include deep learning models such as Enformer in QTL variant evaluation.

Authors must include confidence intervals while reporting performance in all analyses. Without confidence intervals, authors cannot claim that their model is better or worse than alternative models.

Does transfer learning from embedding really help? Random embedding performance, just sequence based model performance without embedding over training epochs needs to be reported. (Similar as in ref [9])

Authors may visualize model interpretation tools such as visualize model gradients per input base pair and convince the reader by showing their model learning know regulatory motifs such as splice dinucleotides AG for acceptor and GT donor, TATA box for promoter site, AATAAA poly(A)-signal for alternative polyadenylation sites. However, learning those regulatory elements is trivial after fine-tuning. It would be interesting to see if the pre-trained model with fine-tuning can capture those regulatory elements.

Minor comments:
The model is trying to remember the context of a 6kb region during training. Does this provide any information on the interactions?

Fig 1a, not clear what sizes and colors mean. What are the models, what is the dataset is better to be clearly labeled.

Better to draw the model in a figure.

Details of the model (variables, detailed architecture) are not clearly presented. Better use tables/figures/equations to describe the details.

Not clear what fig 2b is. What is the meaning of the lighter and darker color?

Fig 4b is difficult to read to similar blue tones. Please use a colorblind friendly palette. Also, It better shows quantitative results as well.

References
[1] gkmSVM https://academic.oup.com/bioinformatics/article/32/14/2205/1743153
[2] kmer-SVM https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt519
[3] Basenji: https://genome.cshlp.org/content/28/5/739.full
[4] BigBert: https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf
[5] DNABert: https://doi.org/10.1093/bioinformatics/btab083
[6] Basenji: https://genome.cshlp.org/content/28/5/739.full
[7] Enformer: https://www.nature.com/articles/s41592-021-01252-x
[8] SpliceAI: https://www.cell.com/cell/pdf/S0092-8674(18)31629-5.pdf
[9] BERTMHC: https://www.biorxiv.org/content/biorxiv/early/2020/11/24/2020.11.24.396101.full.pdf

Version 1:

Decision Letter:

13th Dec 2023

Dear Dr Lopez,

Thank you very much for your patience when your Article, "The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics" is sent out for peer review. The paper has now been seen by Reviewer 1. (We have contacted Reviewer 2 and sent multiple reminders but haven't received their review report.) As you will see from their comments below, although Reviewer 1 still raised 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 address these concerns. Among other revisions, please sufficiently discuss the comparison results with the existing supervised models, add caveats and tone down appropriately. In the point-by-point response letter, please also include your previous responses to the comments from Reviewer 2 in the 1st round of review. In the next round of review, we will assign a new reviewer as an arbiter to weigh in.

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 weeks. 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:

Reviewer #1:
Remarks to the Author:
In the revision, the authors have performed more comparisons with similar pretrained models like DNABERT and HyenaDNA, and added additional benchmarks comparing with Enformer and SpliceAI. However, many of the main concerns raised by reviewers remained inadequately addressed. Some of the new results still suffer from the issues that we pointed out such as training on the human genome and evaluating on yeast data. Below I reiterate two major concerns for this manuscript, and many other main points in my first review were inadequately addressed in my opinion.

1. The current 18-task benchmark provides an inadequate and misleading representation of performance. It is not clear why models trained on human genomes should be evaluated on yeast data, as these two organisms have drastically different biology. The rest of the benchmarks also mostly include mostly small datasets (e.g. enhancers) or easy tasks that do not represent current challenges in the field (e.g. promoter prediction). Since these benchmarks are used for most of the results in this manuscript, it is difficult to draw meaningful conclusions.

Moreover, in the newly added comparison, the authors choose to fine-tune the Enformer model for these tasks. It is not a meaningful comparison as this dataset contains mostly non-human data while Enformer was trained on human and mouse data. In contrast, the Enformer's dataset contains around 7000 tasks, which is much more comprehensive than the ones the authors used here. The authors are encouraged to compare with Enformer on these datasets.

2. There is a lack of performance improvement over strong baselines (e.g. BPNet, Sei, Enformer) developed in the past years on tasks these models were trained for. The same can be said for other pretrained models including DNABERT2 and HyenaDNA. While the authors demonstrated a 1% improvement over SpliceAI when evaluating on a reproduced dataset and comparing with the original reported performance, it is achieved with much higher computational cost. SpliceAI was also surpassed by several recent models that were trained without pretraining by a larger margin.

Overall, I think the fair conclusion over existing foundation models trained on genome sequence including Nucleotide Transformer is that there is still a lack of evidence for improving performance over strong supervised models trained on genome-wide datasets. The value of large pre-trained models in genomics is still not well established given the large computational cost and lack of performance improvement.

Reviewer #2:
None

Version 2:

Decision Letter:

27th Mar 2024

Dear Dr Lopez,

Your Article, "The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics", has now been seen by 2 reviewers (Reviewer 1 and Reviewer 3, a new reviewer who have also been consulted on the Reviewer 1's comments in this round of review). As you will see from their comments below, although the reviewers find your work of considerable potential interest, both of them have raised 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.

Although we appreciate your efforts in revising the paper in previous rounds of revision, in light of the comments from the two reviewers (we had consulted Reviewer 3 who overall agrees with the points raised by Reviewer 1), we think it is important to well address these comments. We therefore invite you to revise your manuscript to address these concerns.

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 4 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:

Reviewer #1:
Remarks to the Author:
While I thank the authors for the response and revision, there are still major issues in the main conclusions of the manuscript, which I explain in detail below.

The authors stated in their last response to reviewers - “Our primary objective in this manuscript was to establish methodologies and discoveries related to constructing and assessing language models for nucleic acid sequences. … The Nucleotide Transformer work stands as the first to establish rigorous protocols and propose an initial collection of evaluation datasets in this domain, and as such we believe it is an important contribution to this objective and the field as a whole” - While I appreciate the efforts the authors put into work, I have to say that the evaluation is poorly conducted and the presentation is misleading. If we consider creation of evaluation datasets and rigorous protocols as the main scientific contribution of the Nucleotide Transformer paper, this manuscript falls short due to numerous concerns that were never addressed during previous revisions.

Even though the authors have toned down in a few places in the text, I still have to respectfully disagree that this manuscript shows the promise of the foundation model in most applications the authors chose. On a related note, a recent preprint (https://www.biorxiv.org/content/10.1101/2024.02.29.582810v1) has performed better designed independent evaluation (applied to nucleotide transformer and other genomic language models) than performed in the manuscript. Their conclusion is that "Our findings suggest that current gLMs do not offer substantial advantages over conventional machine learning approaches that use one-hot encoded sequences. This work highlights a major limitation with current gLMs, raising potential issues in conventional pre-training strategies for the non-coding genome.", where gLM refers to genomic language models, including Nucleotide Transformer, GPN, DNABERT2 and Hyena. While this preprint is recent and presents new evidence, it agrees with my assessment of the manuscript that I don't see good promise on most applications the authors have chosen based on the evidence presented.

Major points:

1. This 18 dataset benchmark has poor quality and poorly represents performance. Rather than establishing rigorous practice, its wider adoption is detrimental to the field. Most if not all of the 18 datasets should not be used in a high quality benchmark for the following reasons (dataset numbed based on order in the Supplementary Table 5):

1,2,3,4,7,8,11,13,16,17 Yeast histone marks. These datasets used very old and based on out-dated technology ChIP-chip. Much better datasets exist and should be used instead. The original URL seems to be inaccessible. I would like to point out that the paper [28] itself was published in 2005 before the rise of modern next-generation sequencing technologies and analysis methods.

5 and 12. Splice donor and acceptor: the labels from the Spliceator paper are not experimentally derived and should not be considered as ground truth. They are from gene prediction and "confirmed" by inspecting multi-sequence alignment. These labels can be heavily biased by previous gene prediction algorithms and contain errors due to no experimental confirmation.

14, 15, 18. Promoter dataset: The negative sequences were constructed by a random substitution procedure that does not respect dinucleotide frequency, which makes the dataset trivial to predict as negatives won't satisfy the statistics of natural sequences.

9, 10. Enhancer dataset: I cannot access the referenced paper because my library does not subscribe to the journal "Biophysical Chemistry".

Thus, most if not all of these datasets have poor quality and also do not represent current challenges in the field of predictive models for genomic sequences. It is indeed true that many recent preprints have adopted this benchmark, however, many adopters likely have not closely understood the source and quality of these datasets and their results obtained are misleading and not representative of true challenges in the field. The authors should discourage the wider usage of this dataset and establish better practices in model evaluations.

The author also claimed that the choice for small datasets were made due to computational cost. I find it strange that the authors spend a lot of computational cost on pretraining the nucleotide transformer, but they are falling short in what is the most important: evaluation and application of the model. Choosing inferior benchmarks to save computational costs does not make sense to me.

2. In addition to the 18 dataset benchmark, the section "The Nucleotide Transformer model learned to reconstruct human genetic variants " analysis is not meaningful and not suitable as an evaluation. It was pointed out in the original review. To explain this more clearly, the authors trained models with and without using 1000 Genomes variants, and the authors evaluated the models on an "independent" human population dataset. However, the "independent" dataset is expected to contain a high degree of overlap in variants contained in 1000 genomes. Showing that the perplexity / likelihood is better when trained on 1000 Genomes with more parameters than just training on reference genome is trivial. The reason is that any model that memorizes the variants and their frequencies in 1000 Genomes can do so, and one can do it without using a deep learning model. Thus, the evaluation is trivial and does not necessarily represent any real gain in knowledge.

3. The problems with proposed evaluation datasets continue if we take into account usage of the 10-fold cross-validation procedure. In my first review, I pointed out the usage of chromosomal holdout for eliminating some spatial adjacency leakage

of information. The authors simply acknowledged that they did not do it in all tasks but did not make any changes.

4. Establishing rigorous protocols also fails short in the comparison with Enformer. As we stated in the previous review, the proposed benchmarking of Enformer is not meaningful as it is not evaluating on what Enformer was designed to do.

While the authors claim (without any citations) that they found several studies using the Enformer torso as a pre-trained model to be further fine-tuned, it is unclear to me why it should be considered as a good practice, nor is it clear that the authors applied a proper process to finetune Enformer. Moreover, it does not contribute to either understanding of Nucleotide Transformer performance or establishing rigorous benchmarking. The last authors' claim is "self-supervised models outperform this technique showcasing the advantage of self-supervised training to build models that can be quickly adapted to various domains". This claim is unconvincing unless the authors show the performance of the Nucleotide Transformer on the Enformer dataset.

In my opinion, if comparison with Enformer remains to be in the current state, it should be removed considering that the authors do not compare Nucleotide Transformer with Enformer at a task that Enformer is designed for.

5. The authors' claim about the superiority of the multi-species model is based on the same 18 datasets which do not constitute solid evidence. The benefit of multi species pretraining is thus far from conclusive. The reasons for the range of genomes included in multispecies training also remain unclear and never justified. The practice of evaluating on datasets from species with drastically different biology (e.g., training on human genome and evaluating on yeast) is highly misleading and does not represent a rigorous protocol for evaluation of multi-species models. The authors' statement in the response, that multi-species model also outperforms human-only model on the human subset of the 18 dataset, is not convincing due to the reasons pointed out above. An appropriate analysis requires much more extensive and high-quality evaluation datasets and careful design (for example, with a more extensive and high-quality set of human benchmark datasets, comparing the performance of training on only the human genome, all primate genomes, all mammal genomes, all vertebrate genomes, etc, and the same goes for evaluating the performance on other species.) .

6. Another problem lies in the inadequate evidence that unsupervised pretraining improve performance. In the rebuttal, the authors pointed out that the pre-trained model outperforms random initialization. However, this is not an evidence because randomly initialized k-mer representation is not expected to function such it does not even properly represent k-mer similarities. As shown in previous works like "Investigation of the BERT model on nucleotide sequences with non-standard pre-training and evaluation of different k-mer embeddings", pretraining on random sequences provided nearly the same benefit as pretraining on real genomic sequences. Improved performance through pretraining for k-mer based architecture can simply be due to the initialization of k-mer representation, which is not need for other modeling choices. More generally speaking, the need for pretraining can be specific to the NT architecture, thus does not constitute an evidence for the value of pretraining for an application.

In the rebuttal, the second piece of evidence the authors point to is the Enformer "comparison". Since this is done on the 18-task benchmark which was not what Enformer was intended for, this adds little to no value for demonstrating the value of pretraining.

Therefore there is no solid evidence to establish that pretraining has any performance advantage.

Finally, it is important to note the differences of "foundation model" in genomic sequence applications compared to NLP when considering its potential: (1) genomic sequences generated from random mutagenesis process (with and without selection) and thus consist of a much lower signal to noise ratio than natural language; (2) genomes are usually fully labeled by genome-wide measurements and there is no shortage of labels (i.e., there is no high proportion of unlabeled datasets like in the NLP); (3) different species' genomes are shaped by different biological processes, making cross species knowledge transfer difficult or even infeasible depending on differences of the species and evaluation should be performed with great care and biological knowledge.

Reviewer #3:
Remarks to the Author:
The manuscript presenting the Nucleotide Transformer seems, on the surface, to be impressive and comes out as improved compared to the first submitted manuscript (which I have not seen). In particular I like the addition of the NT-v2 models which are much faster to fine-tune and will very likely broaden the application of the method. Second, the addition of more models for benchmarking and, third, I appreciate the discussion by the authors of the un/self-supervised vs. supervised training with NT and other DNA transformer models. The point of the NT model is to create models that can be finetuned for other tasks (which any supervised model cannot), not necessarily proving to be the best at all specific tasks.

However, I do also find that several challenges have not been resolved:

Most severe:
1. There are issues with the benchmarking tasks. As mentioned by the other reviewers they are mainly old tasks and datasets - but is what is being used in the field. Is it up to the authors to come up with new (proper) benchmark tasks? For a publication for Nature Methods I think it is – within limits.

2. For any supervised prediction task, a held-out test set, for instance a chromosome, is important for evaluating any prediction task. Was this done properly for all the benchmark tasks?
3. When comparing against the Enformer (which was the latest added benchmark): if the authors want to claim that NT is better than Enformer they should a) do fine-tuning of the Enformer for the 18 tasks, b) show NT vs. Enformer on an Enformer task.

Less severe:
4. Including variants from the 1000G project will make it possible to learn/memorize the allele frequencies and many of these variants will also be in any individual from an “independent” test set. If number of model parameters increased as well then it truly does not show that much (then it basically shows that it could memorize the frequencies).
5. The selection of species for the multi-species training seems quite random - this has also been the case for other DNA language models papers. What is the reasoning for choosing the specific species? It would be interesting to test what models trained on different clades would learn, ie. all primates, all mammals, or any other biologically interesting set of taxa.

I am left with a feeling that overall, the work is interesting, however that some key issues should be resolved and/or acknowledged to merit publication in Nature Methods.

Remarks on code availability:
I have not reviewed the code as the adoption of NT in various other benchmarks indicate that it is quite high quality. Application of models from hugging face are generally straightforward.

Version 3:

Decision Letter:

Our ref: NMETH-A51881C

12th Aug 2024

Dear Dr. Lopez,

We very much apologize for the longer peer review process than we would hope. Thank you for submitting your revised manuscript "The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics" (NMETH-A51881C). Since we are not able to receive review comments from Reviewer 1, in addition to Reviewer 3, we had decided to invite two new reviewers (Reviewers 4 and 5, who co-review) to evaluate the revised manuscript. The paper has now been seen by Reviewers 3, 4 and 5 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 (no need to perform additional analysis) to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

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

Reviewer #3 (Remarks to the Author):

I think the authors have improved the manuscript with the added benchmarks and provided a good rebuttal on the points raised. To me I think they have adequately shown that pre-trained DNA foundation models can be used as a starting point for many DNA and gene structure focused models. As I commented in my first round, I do not think that the authors have to prove that NT is better than all supervised DNA/gene prediction models, but rather that it is a very potent starting point. I think they have achieved that.

Reviewer #4 (Remarks to the Author):

In this paper, the authors trained an unsupervised large language model called Nucleotide Transformer (NT). They created several different versions of NT with varying numbers of parameters (50M to 2.5B) as well as training sets (3,202 human genomes and 850 genomes from species across different phyla), later probing and fine-tuning the models on different downstream tasks. They were benchmarked with the state-of-the-art supervised models in the field, including Enformer and deepSTARR. The authors provide rigorous benchmarking that evaluates the impact that the varying numbers of parameters and training datasets make on performance and biological information that the model is learning. They have addressed the previous reviewer's concerns and in doing so have made a strong and thorough analysis of their model's performance and the biological insights it provides, leaving only minor points to consider.

It's important to recognize that this model was developed before many of the methods it's being compared to, such as HyenaDNA and DNABERT2, and numerous language models have emerged since then. While these models aren't a universal solution for all genomic sequence-based prediction tasks, as smaller, specialized models might perform better in some cases, they offer significant advantages. The primary benefit lies in their nature as "foundation models" - from a single, self-supervised pretraining, they can be fine-tuned for a wide range of tasks. Although the pretraining representations of NT have been shown to be less predictive via probing than specialized supervised models, when fine-tuned (using parameter-efficient methods on just 0.1% of parameters), they can achieve comparable performance. This approach is particularly valuable in situations where developing a custom model for a specific dataset would require extensive architecture and hyperparameter search, which can be a significant burden for many researchers.
In contrast, supervised models typically have architectures optimized for specific datasets or tasks, potentially limiting their applicability to other regulatory genomic tasks. The extent to which pretrained supervised models can also serve as foundation models remains to be fully explored. Even if the NT family of models doesn't achieve state-of-the-art performance on all genomic prediction tasks, its impact on the field is already evident and warrants prompt publication. While questions about its capabilities will persist, these should be addressed in follow-up studies, allowing the scientific community to build upon these models and explore further applications.
Major concerns:
None.

Minor points:
1. Potential data leakage with Enformer and Sei. Informer and Sei were trained on random data splits and so the performance observed by them may be inflated. This should be noted.

2. Due to the largely similar performance between NT-multispecies and NT-1000G, it doesn't seem like a clear-cut takeaway that NT-multispecies is the better model in all cases and that training on multispecies is always better than training on multiple human genomes. The language could be softened to show that each model is beneficial in different contexts, and instead of driving home that NT-multispecies is better further insight into when each model should be applied to maximize their potential would be beneficial.

3. Different colors should be used for the graphs representing NT-1000G (2.5B) and NT-Multispecies (2.5B), as they are too similar to distinguish in certain plots, especially in the perplexity plots and AUC plots.

4. If word count becomes an issue, I would recommend moving this section to supplements: "The Nucleotide Transformer model learned to reconstruct human genetic variants" as it only performs a relative comparison against other NT models with no baseline. Also, these don't consider haplotype structure.

Reviewer #4 (Remarks on code availability):

Nucleotide transformer weights were harnessed with ease for downstream applications.

Editor: Reviewer #5 co-reviews with Reviewer #4.

Version 4:

Decision Letter:

19th Oct 2024

Dear Dr Lopez,

I am pleased to inform you that your Article, "The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics", has now been accepted for publication in Nature Methods. The received and accepted dates will be 1st Mar 2023 and 19th Oct 2024. 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.

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.

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

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

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

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

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

Nucleotide Transformer Rebuttal

Summary of the revision

We are grateful to the two reviewers for their insightful comments and helpful suggestions on our manuscript. We have revised the manuscript to address all comments and suggestions and improve the overall clarity. The main changes and additions are:

- Systematic comparison of our model to five different pre-trained models: DNABERT-1, DNABERT-2, HyenaDNA (1kb and 32kb) and Enformer
- Two additional controls for our fine-tuning approach: probing from raw tokens and fine-tuning from randomly initialized checkpoint
- Additional benchmark datasets and comparison with baselines such as SpliceAI
- Additional interpretation analyses of pre-trained and fine-tuned Nucleotide Transformer models
- Development of a new version of Nucleotide Transformer models (NT-v2) that achieve the same performance while being ten times smaller and having a context length twice as large at 12kb
- Example notebooks to apply these models to any downstream task available on HuggingFace[1]

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

Reviewer #1:

Remarks to the Author:

> The authors have proposed a deep learning model, the Nucleotide Transformer, for unsupervised pre-training of sequence models, and applied it to various prediction tasks. This study also assessed the impact of using diverse genome datasets for pre-training transformer models on DNA sequences. The authors trained multiple models on human reference genomes, 1000 genomes, and 850 genomes from other species like fungi, invertebrates, etc. The authors demonstrated that this approach matches or outperforms the baselines on 12 out of 18 tasks and up to 15 after fine-tuning. The method demonstrated similar performance to DeepSEA or DeepSTARR when fine-tuned on corresponding datasets. Although the authors did a commendable software engineering job, making it possible to train such large models, there are some weaknesses that need to be addressed. These include a lack of strong evidence for advancing over the current state-of-the-art, as well as demonstrating the advantage of training on diverse genomes and unsupervised pretraining, as I detailed below.

We thank the reviewer for the summary and all the constructive comments that allowed us to improve the manuscript further. We thank the reviewer for bringing to our attention that we have not sufficiently clearly explained that the main goal of this work was not to show a strong improvement over the state-of-the-art of every genomics model, but rather to develop a flexible approach that is very adaptable across many diverse genomics tasks, and introduce proper benchmarks that can pave the way for the development of foundational models in genomics. Indeed, since we made our model and benchmark datasets available in March, they were already used as references for the development of further models such as DNABERT-2 and HyenaDNA, and our HuggingFace datasets have been downloaded up to 14k times in one month.

[1] https://huggingface.co/docs/transformers/notebooks#pytorch-bio

We have now improved clarity in the manuscript to emphasize the foundation's ability of the model, methodology and benchmarks. As described in detail below, we have also added more extensive comparisons to other pre-trained models and baselines to demonstrate the advantage of unsupervised training. We also provide notebooks to allow anyone to quickly fine-tune the model on their datasets.

Major points:

> 1. In terms of novelty, the nucleotide transformer approach is very similar to DNABERT. Although the Nucleotide Transformer model is larger and trained on sequences beyond the human reference genome, it is still unclear whether its performance is better. Therefore, a comparison with DNABERT is needed.

We agree with the reviewer that a comparison with other pre-trained models such as DNABERT would be helpful. To provide an extensive comparison and benchmarking of different unsupervised training approaches, we have now added to the manuscript the fine-tuned performance on the 18 downstream tasks of five different pre-trained models: DNABERT-1 [1], DNABERT-2 [2], HyenaDNA (1kb and 32kb context length models; [3]) and Enformer (which was trained as a supervised model on several genomics tasks; [4]). For a fair comparison, we have used the same parameter-efficient fine-tuning approach for every model and used the same cross-validation scheme. Compared with DNABERT and Enformer, our multispecies 2.5B model achieved the best performance in all tasks (see Fig. 2a,b, Supplementary Table 6,8, and below). DNABERT-1 and Enformer achieved this top performance only for the prediction of promoters with TATA-box motif, while DNABERT-2 matched top performance in 9 out of the 18 tasks. Compared with HyenaDNA-1kb the results were more comparable: our 2.5B model has lower performance in 5 tasks, matches performance in other 5 tasks, and shows better performance in 8 out of the 18 tasks. Interestingly, while HyenaDNA was pre-trained on the human reference genome, 6 of the 8 tasks where our multisplecies 2.5B model performs better are from human-related datasets, highlighting again the advantage of pre-training on a diverse set of genome sequences. In addition, the longer-input HyenaDNA-32kb model had worse performance in all 18 tasks, showing that there is a trade-off between increasing the input context length and performance in high-resolution tasks, such as the chromatin/splicing-related genomics tasks used here. We have added these new results in the revised main text, methods, figures and tables, many thanks.
We have also created an interactive leader-board with the results of all models across every task for an easier comparison, accessible at HuggingFace[2]. This is the most extensive benchmark of genomics foundational models so far and should serve as the main reference for future studies.

Model	500M HR	500M 1000G	2.5B 1000G	2.5B MS	DNABERT-1	DNABERT-2	Enformer	hyenadna-tiny-1k	hyenadna-small-32k
H3	0.718 (± 0.014)	0.736 (± 0.012)	0.754 (± 0.008)	**0.793 (± 0.013)**	0.763 (± 0.013)	**0.785 (± 0.012)**	0.724 (± 0.018)	**0.781 (± 0.015)**	0.747 (± 0.017)
H3K14ac	0.372 (± 0.006)	0.381 (± 0.011)	0.453 (± 0.008)	0.538 (± 0.009)	0.403 (± 0.012)	0.515 (± 0.009)	0.284 (± 0.024)	**0.608 (± 0.02)**	0.405 (± 0.015)
H3K36me3	0.448 (± 0.012)	0.468 (± 0.008)	0.53 (± 0.012)	**0.618 (± 0.011)**	0.474 (± 0.009)	0.591 (± 0.005)	0.345 (± 0.019)	**0.614 (± 0.014)**	0.479 (± 0.015)
H3K4me1	0.361 (± 0.015)	0.38 (± 0.009)	0.418 (± 0.012)	**0.541 (± 0.005)**	0.396 (± 0.005)	0.512 (± 0.008)	0.291 (± 0.016)	0.512 (± 0.008)	0.387 (± 0.017)
H3K4me2	0.27 (± 0.015)	0.26 (± 0.016)	0.278 (± 0.015)	0.324 (± 0.014)	0.282 (± 0.013)	0.333 (± 0.013)	0.207 (± 0.021)	**0.455 (± 0.028)**	0.276 (± 0.01)
H3K4me3	0.236 (± 0.018)	0.235 (± 0.008)	0.311 (± 0.012)	0.408 (± 0.011)	0.258 (± 0.01)	0.353 (± 0.021)	0.156 (± 0.022)	**0.55 (± 0.015)**	0.291 (± 0.025)
H3K79me3	0.566 (± 0.016)	0.562 (± 0.015)	0.574 (± 0.007)	0.623 (± 0.01)	0.578 (± 0.007)	0.615 (± 0.01)	0.498 (± 0.013)	**0.669 (± 0.014)**	0.567 (± 0.013)
H3K9ac	0.451 (± 0.011)	0.479 (± 0.01)	0.491 (± 0.009)	0.547 (± 0.011)	0.505 (± 0.009)	0.545 (± 0.009)	0.415 (± 0.02)	**0.586 (± 0.021)**	0.472 (± 0.01)
H4	0.75 (± 0.011)	0.755 (± 0.011)	0.787 (± 0.006)	**0.808 (± 0.007)**	0.784 (± 0.007)	**0.797 (± 0.008)**	0.735 (± 0.023)	0.763 (± 0.012)	0.761 (± 0.017)
H4ac	0.332 (± 0.017)	0.342 (± 0.014)	0.408 (± 0.009)	0.492 (± 0.014)	0.359 (± 0.01)	0.465 (± 0.013)	0.275 (± 0.022)	**0.564 (± 0.011)**	0.379 (± 0.012)
enhancers	**0.502 (± 0.026)**	**0.509 (± 0.033)**	**0.546 (± 0.029)**	**0.545 (± 0.028)**	0.495 (± 0.016)	**0.525 (± 0.026)**	0.454 (± 0.029)	**0.52 (± 0.031)**	0.489 (± 0.018)
enhancers types	**0.429 (± 0.018)**	**0.395 (± 0.027)**	0.432 (± 0.03)	**0.444 (± 0.022)**	0.367 (± 0.034)	**0.423 (± 0.018)**	0.312 (± 0.043)	**0.403 (± 0.056)**	0.352 (± 0.04)
promoter all	0.952 (± 0.002)	0.951 (± 0.003)	0.965 (± 0.002)	**0.975 (± 0.002)**	0.961 (± 0.001)	0.972 (± 0.002)	0.955 (± 0.002)	0.959 (± 0.002)	0.956 (± 0.002)
promoter no tata	0.952 (± 0.002)	0.951 (± 0.002)	0.967 (± 0.001)	**0.977 (± 0.002)**	0.962 (± 0.002)	0.972 (± 0.002)	0.955 (± 0.003)	0.959 (± 0.002)	0.954 (± 0.003)
promoter tata	0.942 (± 0.006)	0.936 (± 0.006)	0.957 (± 0.007)	**0.959 (± 0.004)**	**0.956 (± 0.006)**	**0.955 (± 0.006)**	**0.959 (± 0.006)**	**0.944 (± 0.011)**	0.939 (± 0.008)
splice sites acceptors	0.962 (± 0.003)	0.965 (± 0.002)	0.98 (± 0.002)	**0.986 (± 0.002)**	NaN	0.975 (± 0.002)	0.915 (± 0.007)	0.959 (± 0.003)	0.96 (± 0.007)
splice sites all	0.97 (± 0.002)	0.968 (± 0.001)	0.976 (± 0.002)	**0.982 (± 0.002)**	0.975 (± 0.002)	0.939 (± 0.003)	0.847 (± 0.005)	0.956 (± 0.004)	0.962 (± 0.004)
splice sites donors	0.971 (± 0.002)	0.971 (± 0.001)	0.979 (± 0.001)	**0.987 (± 0.001)**	NaN	0.963 (± 0.002)	0.906 (± 0.008)	0.947 (± 0.007)	0.957 (± 0.006)

> 2. The authors reported that their model matched or outperformed 15/18 baselines on performance. However, it should be noted that these comparisons do not include many state-of-the-art models. Furthermore, the model generally did not surpass stronger baselines, DeepSEA and DeepSTARR, and lacks comparison with current state-of-the-art methods.

For instance, the authors mentioned the Enformer but did not compare the results of the Nucleotide Transformer with those of Enformer in predicting DNASE/ATAC-seq, CAGE, Histone modifications, etc. Similarly, Sei is a recent model reported to outperform DeepSEA. Given the comparison between Nucleotide Transformer and DeepSEA, it is likely that Sei will outperform Nucleotide Transformer on

[2]HuggingFace: https://huggingface.co/spaces/InstaDeepAI/nucleotide_transformer_benchmark

these tasks. Other examples of state-of-the-art models include SpliceAI for splicing variant prediction and BPNet for basepair resolution Chip-Nexus predictions. While it may not be necessary to compare the Nucleotide Transformer to all these methods, some comparison would be needed.

We agree with the reviewer that the manuscript would benefit from more comparisons with additional baselines. We note that in addition to the 18 main downstream tasks used in the paper, we already included in the original manuscript comparisons with DeepSTARR, a state-of-the-art method for enhancer predictions, and DeepSEA, a pioneer method and dataset of chromatin features. We have now added extensive comparisons of Nucleotide Transformer models with SpliceAI [5] for predicting splicing donor and acceptor sites, as suggested by the reviewer (see new Fig. 2d, 5d,e). We adapted the Nucleotide Transformer 2.5B model to deliver nucleotide-level splice site predictions (Fig. 5d) and achieved a top-k accuracy of 95% and a precision-recall AUC of 0.98 with our 6kb-context model (Fig. 2d). Notably, this performance matches the state-of-the-art SpliceAI-10k, which was trained on 15kb input sequences, and outperformed SpliceAI when tested on 6kb input sequences. We have also tested our Nucleotide Transformer v2 500M model that can take 12kb-long input sequences and observed an improved performance of 1% to a top-k accuracy of 96% and a precision-recall AUC of 0.98, an improved performance over SpliceAI (Fig. 5e).

In addition, we have also compared our efficient fine-tuning approach with full-model fine-tuning and observed no significant improvement for chromatin and splicing predictions and a small 3% improvement for enhancer activity predictions (Supplementary Fig. 2), supporting the use of our efficient fine-tuning Nucleotide Transformer approach. We hope this reviewer agrees that we did a substantial effort in the manuscript to provide extensive benchmarks for both the pre-training and finetuning aspects of our models. Comparisons with additional methods will be added in future work.

3. There is a lack of evidence that multispecies training or training on multiple genomes provides an advantage in performance after considering issues in the comparisons. The main problem with this comparison is that models trained on human genomes should not be evaluated on non-human datasets (most of the 18 datasets contain some non-human data).

Most notably, the performance advantage of the multi-species model in Figure 2a mostly comes from yeast histone mark datasets. Given the significant difference in biology between yeast and humans, it is not very meaningful to compare human models using yeast datasets. The observation that human genome-trained models do not perform well on non-human datasets does not constitute a strong argument for the advantage of the multi-species model.
If we exclude the yeast datasets, there is no obvious advantage from the model trained on multi-species genomes. If we exclude other non-human data, it is not unlikely that no advantage can be observed.

Additionally, it is unclear whether a 1000 genomes-trained model outperforms the model trained on only human genome data for models with the same number of parameters, if we exclude the yeast datasets.

We agree that we should evaluate further the human-trained models and the multispecies model on the different types of tasks. To make this comparison more clear in the manuscript, we have now divided the tasks based on species and task category (yeast chromatin profiles, human regulatory elements and human splicing tasks) and compared the different models. This shows that the multispecies 2.5B model achieves the best performance in every task - both human- and non-human ones (see new Fig. 2a,b, Supplementary Table 6). In addition, the multispecies model was also the best on prioritizing pathogenic variants (Fig. 4c,d). These results show that leveraging sequence variation across species, and likely their degree of conservation, is beneficial to build robust pre-trained sequence models.

We also evaluated the performance of the Human-ref and 1000G models (both with 500M parameters) on the human datasets to evaluate the importance of the diversity of sequences. Indeed, both models showed similar performance across the different human-data downstream tasks (Fig. 2a, Supplementary Table 6). We have adapted Figure 2 and discussed better these results in the revised manuscript.

> It is also preferable to use whole chromosome-based holdouts similar to what the authors used for variant effect prediction comparisons, rather than 10-fold cross-validation without taking spatial adjacency into consideration.

We agree with the reviewer that it is better to use chromosome-based holdouts for cross-validation. However, to have a fair comparison to the baselines we used the same test set as in the original publications. For example for DeepSTARR and DeepSEA the train and test datasets are indeed splitted based on chromosomes. We have clarified this in the methods.

> 4. The authors have not shown strong evidence that unsupervised pre-training improves performance. It should be noted that, unlike language modeling, which has a vast amount of unlabeled data, the human genome is well annotated, and most datasets have genome-wide coverage. Therefore, the benefit of unsupervised pre-training cannot be assumed.

We thank the reviewer for pointing out that we have not clearly shown the value of unsupervised pre-training. We have supported this point through multiple approaches in the revised manuscript. (1) We have compared the performance between fine-tuning from pre-trained models or from randomly initialized models, showing that pre-training leads to significantly better performance (see new Supplementary Table 6,8). (2) We have compared our unsupervised pre-trained model with fine-tuning the Enformer model, which was pre-trained in a supervised fashion on almost 7,000 genomics tasks from human and mouse (including TF binding, chromatin features and gene expression). This showed that Enformer fine-tuned models can perform well on tasks related to the pre-trained tasks (promoters, enhancers, and histone marks) but are not able to generalize to unrelated tasks such as splicing (see new Fig. 2a,b). (3) We have clarified in the manuscript the computational benefits of our efficient fine-tuning strategy PEFT on a new task compared with the training of supervised models from scratch for every different task, being cheaper, faster and requiring only 0.1% of the total number of parameters allowing to fine-tune even our biggest models in less than 15 minutes on a single GPU.

We think that this additional evidence strongly supports the benefits of unsupervised pre-training, many thanks.

> This work showed that the pre-trained model matches but does not outperform previous models trained in a supervised fashion, especially for stronger baselines trained on bigger datasets. This still leaves open the question of the benefit of unsupervised training.

We agree with the reviewer that we don't show in this manuscript a strong improvement over previous supervised models, but we want to clarify that we did not focus here on maximising the performance of our fine-tuned models on these downstream tasks. Instead, the main goal of this work was to develop a flexible approach that enables it to tackle many diverse genomics tasks, which contrasts with supervised models that can only be trained on one task at a time and differ in architecture and parameters between every task. This is a key advantage over previous models trained in a supervised fashion, making the training of future models much easier. To demonstrate that, we have also added a comparison with fine-tuning the Enformer model on the different tasks to highlight the foundation's ability of our model.

Finally, we note that genomics language models are very recent, and we expect that further research on new architectures, larger models and more diverse pre-training data will lead to strong improvements in genomics predictions tasks, as it is being observed for older problems in NLP and more recently in protein prediction tasks. We also expect that such foundational models will be more useful for datasets for which there is no genome-wide data available, such as the activity of regulatory elements in vivo or the functional impact of genetic variants. Our work will serve as a reference point in that direction. We have elaborated on this in the revised manuscript (see page 14).

> One potential opportunity for unsupervised pretraining is, maybe it can improve performance when a limited number of positives in training data are available. It is also notable that a recent preprint showed that an unsupervised model can predict variant effect on species without experimental data.

We agree with the reviewer that unsupervised pre-trained models should have stronger benefits for smaller datasets. We have demonstrated this on the prediction of deleterious variants, where datasets are relatively small. Here, our fine-tuned models either slightly outperformed or closely matched the performance of other models based on conservation or functional features (Fig. 4d).
Additionally, an important advantage of unsupervised pre-trained models over supervised models is their zero-shot predictive value, as also demonstrated in the paper cited by the reviewer. Indeed, we show that our zero-shot scores have a very good predictive value for prioritizing functional variants (AUCs across different variant types from 0.7 to 0.8), even matching or surpassing other methods on eQTLs and meQTLs tasks (see Fig. 4c,d). We have discussed this in the revised manuscript (see page 14).

> 5. For the evaluation of reconstruction accuracy on variants, if I understand correctly, does not seem to be the appropriate setup. Since the model trained on 1000 Genomes can simply memorize allele frequencies, showing that it reconstructs variants from a different dataset does not necessarily show that the model learned to use sequence in a non-trivial way to predict variants. Evaluation on holdout chromosomes is better for this evaluation.

We thank the reviewer for pointing out that this was not sufficiently clearly explained. The goal of this analysis is to test the imputation capability of our pre-trained language model - i.e. to assess if a model pre-trained in genomes from a given diverse population (and respective allele frequencies) can generalize to reconstruct the variants of unseen human genome sequences. Similar to imputation methods, this needs to be evaluated in the same genomic regions seen during pre-training, but in genomes with different haplotypes, as we did here for an independent dataset of genetically diverse human genomes. Evaluating this capability in a holdout chromosome would likely not work because the model wouldn't be able to predict variants and haplotype if not from previously observed LD. We have clarified this point in the respective section of the main text (page 7), many thanks.

Minor points:

> 1. I am not sure if the maximum attention percentages in some layers reaching 100% for enhancers is particularly meaningful, as all element types seem to have some layers reaching 100%.

We understand the reviewer's comment. Layers reaching 100% do not occur for all element types, but they do occur for 3'UTR, promoter and TF binding sites, in addition to enhancers. We think this information is meaningful but have added the other elements. In addition, we have also highlighted now how the different element types are captured by different heads and layers of the same model (see page 8). We have also improved this results section and respective Fig. 3f,g to highlight better these attention interpretation analyses.

> 2. Comparison with DeepSEA in Figure 4d does not seem appropriate. For example, the nucleotide transformer was fine-tuned on each dataset, but DeepSEA was not. DeepSEA is also not intended for predicting coding mutations or splicing variants, which are the majority of mutations in ClinVar or HGMD.

We thank the reviewer for pointing out that our explanations had been unclear. It is indeed true that in Fig 4d we fine-tuned the Nucleotide Transformer models on each variant dataset. However, regarding DeepSEA, we used the DeepSEA (Beluga version)-predicted Disease Impact Score (DIS) (REF). This score is based on a logistic regression model that prioritizes likely disease-associated mutations on the basis of the predicted transcriptional or post-transcriptional regulatory effects of these mutations (See Zhou et. al, 2019), and is used in the respective paper to predict the functional impact of genetic variants in gene expression, including eQTLs. Thus we think that it is a relevant method for comparison in the variant prediction tasks where we evaluate our models.
In addition, we agree with the reviewer that the DeepSEA scores are related mostly to transcriptional effects, which explains the lower performance of DeepSEA in ClinVar and HGMD tasks. We have added it for consistency, to evaluate all methods across the 4 variant tasks (eQTL, meQTL, ClinVar, HGMD). We have now clarified how the DeepSEA scores were calculated and used in the revised main

text and methods for clarity.

> 3. It seems that the definition of $p_a(f)$ should be the proportion of attention scores greater than, rather than less than, the threshold. The statistical test for $p_a(f)$ seems to be not very meaningful, given that any threshold can be significant. Given the sample size, I don't think a p-value is needed for this analysis.

We thank the reviewer for pointing out this typo in the formula and have corrected it.
Regarding the statistical test, we still think it's a informative metric to identify heads that significantly attend to the genomic elements and have decided to keep it for completeness.

> 4. More information could be given on how the 1000 genomes dataset was constructed. It should also be noted that the reconstructed genome does not contain all variants.

We have revised the Methods section to describe how the 1000 genomes dataset was constructed, many thanks.

Reviewer #2:

Remarks to the Author:

> In their paper, the authors introduce the Nucleotide Transformer as a foundational model for genomics. The authors describe the Nucleotide Transformer, pitched as a foundation model for genomics. Nucleotide Transformer follows the BERT architecture used for training large language models (LLM). Nucleotide Transformer models are trained in a way similar to BERT, asking the model to predict masked tokens, in this case, 6-mers. This idea is that sequence embedding learned from NT can be used for downstream genomic tasks with probing or fine-tuning.
> Building on the success of LLMs, there are recent a deluge of papers on similar ideas, building foundation models for genomics, utilizing sequence only information such as DNABert, BIGBERT, GenSLM, GPN (from Yun Song's group),... or a combination of sequence or functional annotations such as Enformer, BPNet, ... So neither method nor idea is new here.
>
> Instead, the authors focus on exploring the impact of sizes of the model and diversity of the datasets. Specifically, authors leveraged genomics sequences from multiple species and attempted to show the merit of training from diverse species. While potentially interesting, the presented work is an incremental contribution building on existing works. There are major concerns about the design of the model, as well as the validity and effectiveness of the conclusions drawn from the papers.

We thank the reviewer for the summary of our work and the context in the field, and for all the constructive comments that allowed us to improve the manuscript further. We thank the reviewer for pointing out that we have not sufficiently clearly explained how our work differs from recent foundation models for genomics. Although there have been similar methods developed during the past two years (DNABERT, BigBird, GenSLMs, GPN, HyenaDNA), we are the first to (1) systematically evaluate different model sizes (50M up to 2.5B parameters) and different pre-training datasets (the human reference genome, a collection of 3,202 diverse human genomes, and 850 genomes from several species), and (2) to introduce proper benchmarks that can pave the way for the development of foundational models in genomics. Indeed, since we made our work and benchmarks available in March, they were already used as references for the development of further models such as DNABERT-2 and HyenaDNA, and our HuggingFace datasets have been downloaded up to 14k times in one month.

To further improve our work and establish it as the reference genomics LLM model paper in the field, we have added in the revised manuscript an extensive comparison and benchmarking to five different pre-trained models (DNABERT-1 [1], DNABERT-2 [2], HyenaDNA (1kb and 32kb context length models; [3]) and Enformer (which was pre-trained as a supervised model on several genomics tasks; [4])), two of them posted online during this revision. Furthermore, we have developed a new version

of Nucleotide Transformer models that are able to achieve the same performance than the 2.5B model while being 10 times smaller and having a context length twice as large at 12kb.

We have now improved clarity in the manuscript to emphasize how our work differs from other genomics LLMs, including our methodology and benchmarks. In addition, we also provide notebooks to allow anyone to quickly fine-tune our model on their datasets. Many thanks.

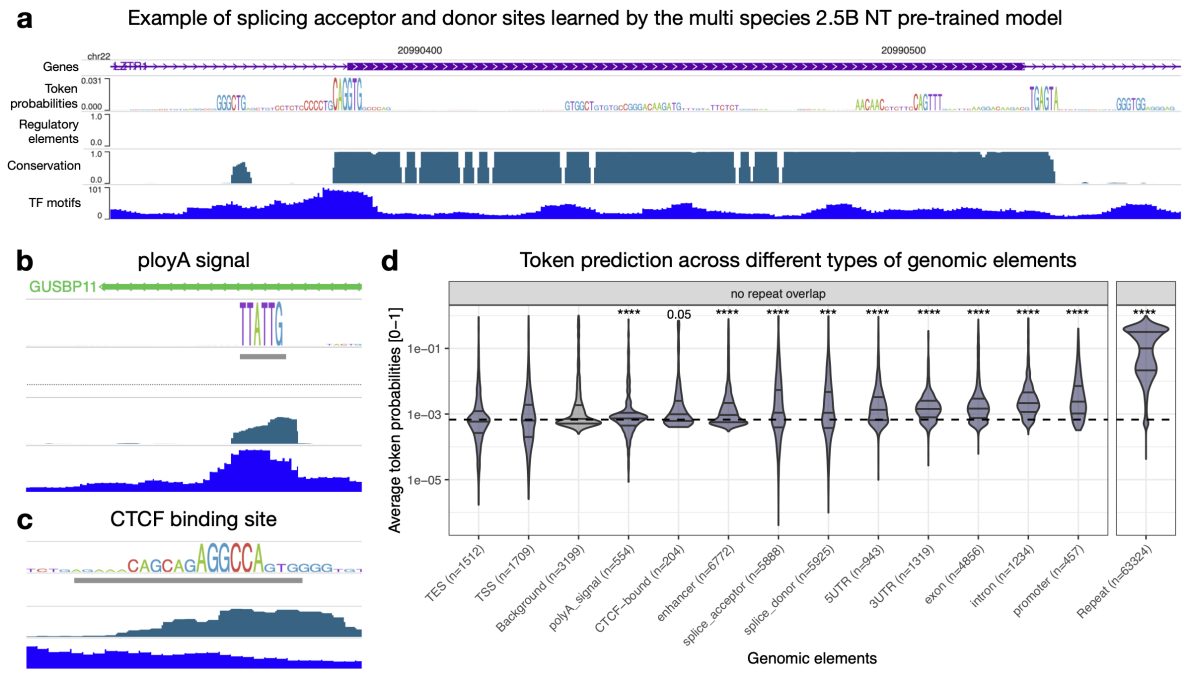
My major comments are:

> At the most basic level, I am not fully convinced that a naive application of BERT type of models, developed for NLP, to DNA sequences is a productive approach. In NLP, the tokens have meanings and are naturally defined. It is unclear what "token" means in the context of DNA sequence. There is no such thing as simple functional units in DNA, especially for noncoding sequences. The human genome is littered with many different kinds of functional elements buried in a sea of nonfunctional sequences. Other than coding sequences or RNA genes, it has been very difficult to properly define and model other types of functional elements. Demarking the boundary of non coding functional elements is even harder. There is also the issue of "grammar" or "rules" with the DNA sequences. Despite years of research, it is still not clear what they are and how general the rules are. So I would like to have some understanding of what kind of knowledge is learned by the Nucleotide Transformer. It is important for authors to look deeper into the interpretation side to find out what their models have learned. Is that information supported by prior biological knowledge?

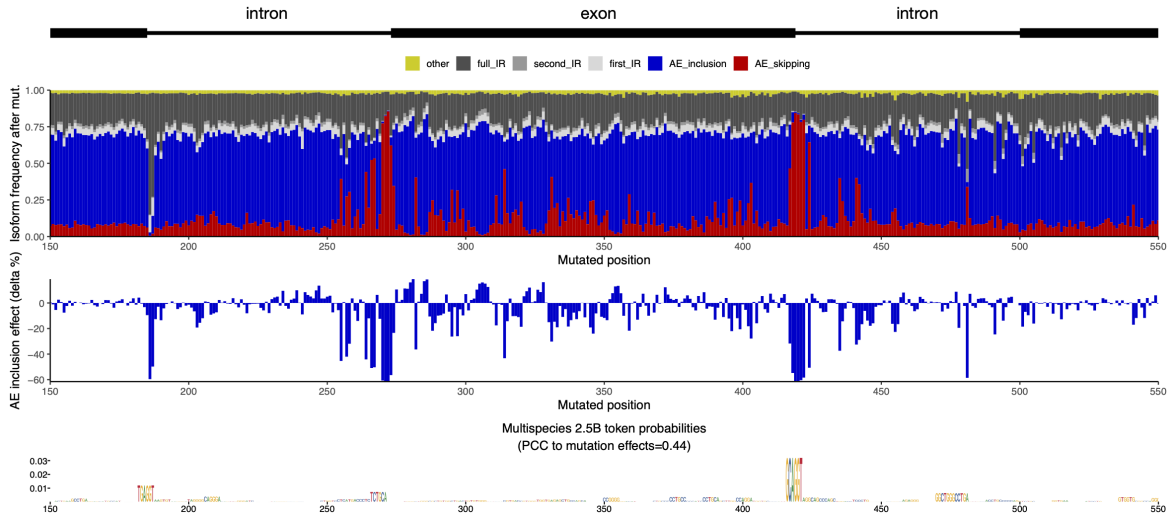
We agree with the reviewer that the interpretation of the sequence-rules learned by the Nucleotide Transformer is relevant for our understanding of how these LLMs encode genomic sequences. We note that interpretation of such models remains very challenging and is an active field of research. Still, we have addressed this in the original manuscript using several state-of-the-art techniques. We showed that the DNA sequences of 5 different types of genomic elements (3'UTR, CDS, intron, intergenic, 5'UTR) are encoded in the model through unique and very distinct sequence embeddings at the different layers (see Fig 3d and Supplementary Fig. 6). In addition, we performed extensive analyses of the attention maps to understand which sequence regions are captured and used by the attention layers (see page 8, Fig. 3f,g and Supplementary Fig. 7-15). We have now improved these analyses to show that the model attention specifically targets the different types of genomic elements throughout its different layers, such as 5'UTRs, exons, introns, 3'UTRs, open chromatin regions, promoters, enhancers, transcription factor (TF) binding regions and CTCF regions. These different element types are captured by different heads and layers, where the maximum attention percentages were highest for the largest models, reaching for instance a value of almost 100% attention for enhancers in the 1000G 2.5B model.

We have also added now additional analyses to interpret the pre-trained models at higher resolution (i.e. more local sequence features). We have looked at token probabilities across the genome and different types of genomics elements as a measure of sequence constraint and importance learned by the language model. Specifically, we calculated the 6-mer token probabilities (based on masking each token at a time) for every 6kb chunk in chr22. Analysing this token reconstruction metric revealed that in addition to repetitive elements, that are well reconstructed by the model as expected, the pre-trained model learned a variety of gene structure and regulatory elements. For instance, in panel (a) we show a screenshot of an exon where the model predicts well the acceptor and donor splicing sites. In panels (b) and (c) we show examples of polyA signal and CTCF binding sites, respectively. The original nucleotides are shown with their height proportional to the token probabilities. We have also created an interactive browser session with token probabilities across chromosome 22 on the WashU Epigenome Browser[3] to allow further exploration by the readers. We summarised the token probabilities across different types of genomics elements and show that the pre-trained model significantly learned about the structure and sequence features of different types of elements (panel (d)).

[3]WashU Epigenome Browser: https://shorturl.at/jov28



Furthermore, we compared our token reconstruction predictions with an experimental saturation mutagenesis splicing assay of exon 11 of the gene MST1R (data from Braun et al. [6]). The top panel shows the measured impact in the different splicing isoforms of mutating every nucleotide position, with the effect in Alternative Exon (AE) 11 inclusion showed in the middle panel. Note how the expected important positions (such as splicing acceptor and donor) match the positions with the strongest mutation impact. Importantly, we observed a significant correlation between these experimental mutation effects and the token predictions by our multispecies 2.5B pre-trained model (Pearson Correlation Coefficient (PCC) = 0.44). Our model learns well constraints at the different splicing junctions, but also a region in the middle of the second intron that is indeed important for the splicing of this exon. This is a strong validation of the biological knowledge acquired by our model during unsupervised pre-training.



We have integrated these new analyses and results in the revised manuscript (see Supplementary Fig. 16). These results illustrate how the transformer models have learned to recover gene structure and functional properties of genomic sequences and integrate them directly into its attention mechanism. We hope this reviewer agrees that we have done our best to interpret the biological knowledge encoded in these models. We acknowledge that the interpretation of such models remains challenging and an intense field of research. Future studies will allow us to improve their interpretation and reveal novel

biological principles.

> One of the main points that the authors argue for is the large models with millions of parameters. Big models have shown utility for LLMs. But I am not convinced it is a good idea for training models for DNA sequences.Language or vision are compositional. The number of training samples is unbounded. This kind of diversity is crucial for training LLMs. By contrast, the diversity associated with the human genome is limited, even accounting for genetic variants. With only 3 billion letters in the human genome, a large model with hundreds of millions of parameters can simply memorize the genome.

Genomics LLMs learn statistical relationships between DNA sequence-features and sequence types from the vast amount of sequences they've seen during training. The goal is not to memorize the training data, but to generalize from the sequence relationships learned during training to unseen sequences or downstream tasks. Indeed, we have monitored overfitting during pre-training by having a held-out validation set of sequences and observed no overfitting. In addition, while memorization by LLMs is an open research area in NLP and can indeed be an issue for zero-shot evaluations, it will not lead to overfitting for downstream prediction tasks like the ones shown in this manuscript. Even if the model does memorize all DNA sequences, as the data is unlabeled, it does not help to perform the downstream tasks. Actually, most probably overfitting would affect the downstream performance.

> Can authors provide a convincing argument and experimental support on why larger models are better for DNA sequences?

We thank the reviewer for pointing out that this was not sufficiently clearly explained. We have assessed the importance of model size by training four distinct language models of different sizes, ranging from 500M up to 2.5B parameters. We assessed the ability of each model on 18 different genomic prediction datasets and observed that the largest models outperformed their smaller counterparts (Fig 2a,b). In addition, we have added in the revised manuscript a second version of the Nucleotide Transformer models, with a new architecture. Again, we have tested this new model architecture with different sizes ranging from 50M to 500M parameters and show that the largest model achieves the best performance (see Fig. 51-c). These results match observations in other fields such as NLP, where larger training datasets and model sizes have been shown to improve performance [7].
Regarding the experimental support, all our models of different sizes were evaluated in 18 different downstream tasks whose datasets derive from experimental work. Our work is the first to evaluate the importance of model size in genomics LLMs. We have revised the manuscript to clarify the different points described above and hope the reviewer is convinced that our flexible LLM approach is promising to model all the different types of DNA sequences.

> There is insufficient review of prior works and discuss how the current work differs from prior ones. Authors may improve their introduction by briefly mentioning existing task specific (Enformer, BPNet, ...) and general nucleotide language models (DNABERT, ...). Then, they can describe their added values compared to their proposed model. For example, if their model size or multi-species training is the major difference from state of the art models. Additionally, authors should compare their contributions against state of the art by conducting further benchmarks in the result section to emphasize the added values of their work.

We thank this reviewer for this suggestion that helped us improve the clarity of the manuscript. We have extended the review of previous work in the introduction and mentioned additional task-specific models (BPNet, Sei, ORCA) and pre-trained language models that were released during the revision of our manuscript (DNABERT2, GENA-LM and HyenaDNA) (see page 2). We note that we have already mentioned the pioneers supervised Enformer model and unsupervised language model DNABERT1 in the previous manuscript version.
In addition, we have emphasized throughout the manuscript the foundation's ability of the model, its methodology, the different model sizes and training sets evaluated, and the systematic benchmark datasets and analyses that we provide.
We also agree with the reviewer that the manuscript would benefit from more comparisons with additional baselines. We already included in the manuscript comparisons with DeepSTARR, a state-

of-the-art method for enhancer predictions, and DeepSEA, a pioneer method and dataset of chromatin features. We have now added extensive comparisons of Nucleotide Transformer models with SpliceAI [5] for predicting splicing donor and acceptor sites, as suggested by the reviewer (see new Fig. 2d, 5d,e). We adapted the Nucleotide Transformer 2.5B model to deliver nucleotide-level splice site predictions (Fig. 5d) and achieved a top-k accuracy of 95% and a precision-recall AUC of 0.98 with our 6kb-context model (Fig. 2d). Notably, this performance matches the state-of-the-art SpliceAI-10k, which was trained on 15kb input sequences, and outperformed SpliceAI when tested on 6kb input sequences. We have also tested our Nucleotide Transformer v2 500M model that can take 12kb-long input sequences and observed an improved performance of 1% to a top-k accuracy of 96% and a precision-recall AUC of 0.98, an improved performance over SpliceAI (Fig. 5e).

As suggested by both reviewers, and to provide an extensive comparison and benchmarking of different unsupervised training approaches, we have now added to the manuscript the fine-tuned performance on the 18 downstream tasks of five different pre-trained models: DNABERT-1 [1], DNABERT-2 [2], HyenaDNA (1kb and 32kb context length models; [3]) and Enformer (which was trained as a supervised model on several genomics tasks; [4]). We have used the same parameter-efficient fine-tuning approach for every model and the same cross-validation scheme to enable direct comparison of the results. Compared with DNABERT and Enformer, our multispecies 2.5B model achieved the best performance in all tasks (see Fig. 2a,b, Supplementary Table 6,8, and below). DNABERT-1 and Enformer achieved this top performance only for the prediction of promoters with TATA-box motif, while DNABERT-2 matched top performance in 9 out of the 18 tasks. Compared with HyenaDNA-1kb the results were more comparable: our 2.5B model has lower performance in 5 tasks, matches performance in other 5 tasks, and shows better performance in 8 out of the 18 tasks. Interestingly, while HyenaDNA was pre-trained on the human reference genome, 6 of the 8 tasks where our multisplecies 2.5B model performs better are from human-related datasets, highlighting again the advantage of pre-training on a diverse set of genome sequences. In addition, the longer-input HyenaDNA-32kb model had worse performance in all 18 tasks, showing that there is a trade-off between increasing the input context length and performance in high-resolution tasks, such as the chromatin/splicing-related genomics tasks used here.
We have created an interactive leader-board with the results of all models across every task for an easier comparison, accessible at HuggingFace[4]. This is the most extensive benchmark of genomics foundational models so far and should serve as the main reference for future studies.

Thank you for all the suggestions.

Model	500M HR	500M 1000G	2.5B 1000G	2.5B MS	DNABERT-1	DNABERT-2	Enformer	hyenadna-tiny-1k	hyenadna-small-32k
H3	0.718 (± 0.014)	0.736 (± 0.012)	0.754 (± 0.008)	**0.793 (± 0.013)**	0.763 (± 0.013)	**0.785 (± 0.012)**	0.724 (± 0.018)	**0.781 (± 0.015)**	0.747 (± 0.017)
H3K14ac	0.372 (± 0.006)	0.381 (± 0.011)	0.453 (± 0.008)	0.538 (± 0.009)	0.403 (± 0.012)	0.515 (± 0.009)	0.284 (± 0.024)	**0.608 (± 0.02)**	0.405 (± 0.015)
H3K36me3	0.448 (± 0.012)	0.468 (± 0.008)	0.53 (± 0.012)	**0.618 (± 0.011)**	0.474 (± 0.009)	0.591 (± 0.005)	0.345 (± 0.019)	**0.614 (± 0.014)**	0.479 (± 0.015)
H3K4me1	0.361 (± 0.015)	0.38 (± 0.009)	0.418 (± 0.012)	**0.541 (± 0.005)**	0.396 (± 0.005)	0.512 (± 0.008)	0.291 (± 0.016)	0.512 (± 0.008)	0.387 (± 0.017)
H3K4me2	0.27 (± 0.015)	0.26 (± 0.016)	0.278 (± 0.015)	0.324 (± 0.014)	0.282 (± 0.013)	0.333 (± 0.013)	0.207 (± 0.021)	**0.455 (± 0.028)**	0.276 (± 0.01)
H3K4me3	0.236 (± 0.018)	0.235 (± 0.008)	0.311 (± 0.012)	0.408 (± 0.011)	0.258 (± 0.01)	0.353 (± 0.021)	0.156 (± 0.022)	**0.55 (± 0.015)**	0.291 (± 0.025)
H3K79me3	0.566 (± 0.016)	0.562 (± 0.015)	0.574 (± 0.007)	0.623 (± 0.01)	0.578 (± 0.007)	0.615 (± 0.01)	0.498 (± 0.013)	**0.669 (± 0.014)**	0.567 (± 0.013)
H3K9ac	0.451 (± 0.011)	0.479 (± 0.01)	0.491 (± 0.009)	0.547 (± 0.011)	0.505 (± 0.009)	0.545 (± 0.009)	0.415 (± 0.02)	**0.586 (± 0.021)**	0.472 (± 0.01)
H4	0.75 (± 0.011)	0.755 (± 0.011)	0.787 (± 0.006)	**0.808 (± 0.007)**	0.784 (± 0.007)	**0.797 (± 0.008)**	0.735 (± 0.023)	0.763 (± 0.012)	0.761 (± 0.017)
H4ac	0.332 (± 0.017)	0.342 (± 0.014)	0.408 (± 0.009)	0.492 (± 0.014)	0.359 (± 0.01)	0.465 (± 0.013)	0.275 (± 0.022)	**0.564 (± 0.011)**	0.379 (± 0.012)
enhancers	**0.502 (± 0.026)**	**0.509 (± 0.033)**	**0.546 (± 0.029)**	**0.545 (± 0.028)**	0.495 (± 0.016)	**0.525 (± 0.026)**	0.454 (± 0.029)	**0.52 (± 0.031)**	0.489 (± 0.018)
enhancers types	**0.429 (± 0.018)**	**0.395 (± 0.027)**	0.432 (± 0.03)	**0.444 (± 0.022)**	0.367 (± 0.034)	**0.423 (± 0.018)**	0.312 (± 0.043)	**0.403 (± 0.056)**	0.352 (± 0.04)
promoter all	0.952 (± 0.002)	0.951 (± 0.003)	0.965 (± 0.002)	**0.975 (± 0.002)**	0.961 (± 0.001)	0.972 (± 0.002)	0.955 (± 0.002)	0.959 (± 0.002)	0.956 (± 0.002)
promoter no tata	0.952 (± 0.002)	0.951 (± 0.002)	0.967 (± 0.001)	**0.977 (± 0.002)**	0.962 (± 0.002)	0.972 (± 0.002)	0.955 (± 0.003)	0.959 (± 0.002)	0.954 (± 0.003)
promoter tata	0.942 (± 0.006)	0.936 (± 0.006)	0.957 (± 0.007)	**0.959 (± 0.004)**	**0.956 (± 0.006)**	**0.955 (± 0.006)**	**0.959 (± 0.006)**	**0.944 (± 0.011)**	0.939 (± 0.008)
splice sites acceptors	0.962 (± 0.003)	0.965 (± 0.002)	0.98 (± 0.002)	**0.986 (± 0.002)**	NaN	0.975 (± 0.002)	0.915 (± 0.007)	0.959 (± 0.003)	0.96 (± 0.007)
splice sites all	0.97 (± 0.002)	0.968 (± 0.001)	0.976 (± 0.002)	**0.982 (± 0.002)**	0.975 (± 0.002)	0.939 (± 0.003)	0.847 (± 0.005)	0.956 (± 0.004)	0.962 (± 0.004)
splice sites donors	0.971 (± 0.002)	0.971 (± 0.001)	0.979 (± 0.001)	**0.987 (± 0.001)**	NaN	0.963 (± 0.002)	0.906 (± 0.008)	0.947 (± 0.007)	0.957 (± 0.006)

> I have some concerns regarding the evaluations and experimental results. First, evaluation tasks should be expanded to include recent functional genomic data from ENCODE or Epigenomic projects. The authors used the dataset of DeepSEA, which was curated several years ago. There are more recent and richer datasets that the authors should evaluate.

We agree with the reviewer that there have been more recent and comprehensive genomics chromatin datasets. We chose to use the original DeepSEA dataset because (1) its 919 classification tasks provide a manageable dataset size but still diverse set of transcription factor, DNA accessibility and histone modification tasks across different cell types, and (2) because it has been used in genomics

[4]HuggingFace: https://huggingface.co/spaces/InstaDeepAI/nucleotide_transformer_benchmark

as a reference dataset across the last years, with performance metrics reported across many different methodologies. While we could have done the benchmarks in a larger dataset version, we expected that the performance difference to the baselines would be the same, and decided that it would not be worth the extended computational cost of such experiments.

> Second, the improvement of their models compared to the so-called "SOTA" are marginal in most cases. So this raises the important question on the value of the proposed models. This is also in stark contrast to the results from LLMs, where foundation models offer more substantial improvements.

We thank the reviewer for bringing to our attention that we have not sufficiently clearly explained that the main goal of this work was not to show a strong improvement over the state-of-the-art of every genomics model, but rather to develop a flexible approach that is very adaptable across many diverse genomics tasks, and introduce proper benchmarks that can pave the way for the development of foundational models in genomics. Our foundational models contrast with supervised models that can only be trained on one task at a time and differ in architecture and parameters between every task. This is a key advantage over previous models trained in a supervised fashion, making the training of future models much easier.
To elaborate on this, we have added a comparison with fine-tuning the Enformer model on the different tasks to highlight the foundation's ability of our model. To further evaluate the potential of pre-trained models, we have also fine-tuned the whole Nucleotide Transformer model (not using PEFT as before) on the more challenging downstream tasks (DeepSTARR, DeepSEA and SpliceAI) and showed an improvement over baseline supervised models.

Finally, we note that genomics language models are very recent, and we expect that further research on new architectures, larger models and more diverse pre-training data will lead to strong improvements in genomics predictions tasks, as it is being observed for older problems in NLP and more recently in protein prediction tasks. We also expect that such foundational models will be more useful for datasets for which there is no genome-wide data available, such as the activity of regulatory elements in vivo or the functional impact of genetic variants. Our work will serve as a reference point in that direction. We have improved clarity throughout the manuscript to emphasize the foundation's ability of the model, methodology and benchmarks, as well as the computational benefits of our efficient fine-tuning strategy PEFT on a new task compared with the training of supervised models from scratch for every different task, being cheaper, faster and requiring only 0.1% of the total number of parameters allowing to fine-tune even our biggest models in less than 15 minutes on a single GPU.

> Authors utilized 6-mer based tokenization. In natural language modeling, word-based tokenization is optimal because words are the smallest meaningful units constructing a sentence. However, DNA grammar does not contain units like words and instead contains regulatory motifs. Earlier genomics models prior to deep learning such as gkmSVM and kmer-SVM k-mer based methods are widely used. However, convolution based models have become the standard because convolutional deep learning models such as DeepSea, Basenji, or Enformer have superior performance. because convolutional layers learn to tokenize regulatory elements from raw DNA sequences. In the manuscript, the authors do not provide any justification for 6-mer based tokenization over the use of convolution. Their architecture choice has major computational cost and limits the receptive field of the model to 6,000 bp.

We thank the reviewer for pointing out that we have not provided a clear justification for the use of 6-mer based tokenization. As the reviewer writes in the introduction above, it is still not clear what and how general the "grammar" or "rules" of DNA sequences actually are. However, the regulatory code is commonly compared with a form of language, taking the regulatory motifs as "words" and their interactions as the sequence "grammar" of the regulatory code. The benefit of sequence tokenization is that we don't need to make any assumption about the DNA code structure. Similar to NLP, the tokenization is not done for tokens to have meaning but just to find a trade-off between sequence length and vocab size. The "meaning" is captured later by embeddings during training. In fact, most of the state-of-the-art NLP models don't just take words as tokens but rather use some kind of subword-tokenization algorithms. Similarly, there is probably not a unique token approach that will work the best for every genomics problem, and it will depend on how the embeddings are learned

during training.
Here we have decided to take a tokenization approach based on a trade-off between sequence length (up to 6kb) and embedding size. We have experimented with different k-mer sizes and observed the highest performance with 6-mers (data not shown). Similarly, DNABERT also found 6-mer based tokenization to achieve the best performance for DNA tasks (see DNABERT paper). We have added more details regarding our tokenization approach in the revised manuscript.
In addition, we have developed a new version of Nucleotide Transformer models (NT-v2), by leveraging contemporary architectural advancements and extending training duration, that are able to achieve the same performance than the 2.5B model while being 10 times smaller and having a context length twice as large at 12kb.

> Authors need to compare alternative architectures such as a model with convolution layers followed by attention layers.
> Thereby, they may increase the receptive field and reduce the model training cost. If their current architecture outperforms the alternative architectures then their model choice would be justified. Also, masking-based semi-supervised training is compatible with convolutional models; for example, authors may utilize intermediate layer masking after convolutional layers but before attention layers.

The comparison with an architecture based on convolutional layers is a very interesting point given the success of convolutional models in genomics, and could indeed justify our model choice. While developing a new pre-trained model based on convolutional layers and train it from scratch is out of scope of this manuscript, following the reviewer's suggestion, we have compared our unsupervised pre-trained model (purely based on embeddings and attention) with fine-tuning the Enformer model, which combines convolution layers with attention and was pre-trained in a supervised fashion on almost 7,000 genomics tasks from human and mouse (including TF binding, chromatin features and gene expression). Comparing both approaches on our 18 downstream tasks showed that Enformer fine-tuned models can perform well on tasks related to the pre-trained tasks (promoters, enhancers, and histone marks) but are not able to generalize to unrelated tasks such as splicing (see new Fig. 2a,b). Still, our multispecies 2.5B model achieved the best performance in all tasks, matched by Enformer only for the prediction of promoters with TATA-box motif. Given that Enformer is considered a state-of-the-art model in the field, this analysis is of great interest, and we have added it in the revised manuscript, many thanks.

> SVM bases models have a number of ways similar to final probing layers of the proposed model. Thus, authors may compare probing and gkmSVM despite the method may not be the state of the art.

We thank the reviewer for this suggestion. However, as the reviewer mentions, SVM-based models have been outperformed by deep learning models, such as the ones we use as baselines in this manuscript. Since we have already expanded our comparisons with 4 additional pre-trained language models in the revised manuscript, we have decided to not include SVM-based models in our comparisons for conciseness.

> The state of the art models repeatedly show that increasing the model receptive field significantly improves the performance because sequence context is critical to understand DNA grammar [7, 8]. Their 6 kb sequence length is smaller compared to state of the art models in genomics. Authors may try different lengths such as 1kb, 2kb, .etc, and show that model performance converges with 6 kb. Alternatively, Authors may increase the receptive field with convolution as discussed in the previous comment.

We agree with the reviewer that the length of the model receptive field is an important aspect for genomics models, where longer sequence context is leading to better results. We have managed to increase it to 6kb for the Nucleotide Transformer V1 models (compared for instance with 512bp for DNABERT-1 and GPN), and to 12kb for the Nucleotide Transformer V2 model thanks to its more efficient architecture. We want to clarify that finding the best possible architecture and sequence length was not the purpose of this study, and that 6kb or 12kb are not said to be the best input lengths. These were still chosen due to computing limitations and we are working on extending them in further

models.
Still, increasing the input context length is not a trivial task, as also demonstrated by our evaluation of HyenaDNA. The HyenaDNA models were developed to have context lengths up to 1 million bases, but our benchmarking results show that the longer-input HyenaDNA-32kb model had worse performance in all 18 tasks than our 2.5B model and the shorter-input HyenaDNA-1kb model (Fig. 2a,b). This shows that there is a trade-off between increasing the input context length and the performance in high-resolution tasks, such as the chromatin/splicing-related genomics tasks used here. This is an area of intense research, and we and others are currently working on efficiently extending the context length of this type of genomics foundational models. We have added this notion to the main text in the revised manuscript (see page 4).

> Authors claim that an increasing number of parameters in the model is helpful; however, authors haven't performed benchmarks based on different parameter numbers so they did not demonstrate this point.

We thank the reviewer for pointing out that this was not sufficiently clearly explained. We have assessed the importance of model size by training four distinct language models of different sizes, ranging from 500M up to 2.5B parameters. We assessed the ability of each model on 18 different genomic prediction datasets and observed that the largest models outperformed their smaller counterparts (Fig. 2a,b). In addition, we have added in the revised manuscript a second version of the Nucleotide Transformer models, with a new architecture. Again, we have tested this new model architecture with different sizes ranging from 50M to 500M parameters and show that the largest model achieves the best performance (see Fig. 5a-c, Supplementary Table 6,8). These results match observations in other fields such as NLP, where larger training datasets and model sizes have been shown to improve performance [7]. We have emphasized this analysis in the revised manuscript, many thanks.

> Moreover, state of the art models such as Enformer contain significantly far fewer parameters.
> Also, authors may show their large models are helpful by benchmarking against state of the art models such as Enformer.

We thank the reviewer for pointing out that we have not sufficiently clearly explained the difference between the number of parameters in finetuned unsupervised language models and supervised models from scratch. It is true that our pre-trained models have a large number of parameters (the largest having 2.5 billion), however with our efficient fine-tuning strategy PEFT the number of parameters that are finetuned are only a very small proportion of 0.1% (2.5 million parameters for the 2.5B model). This contrasts with the high number of parameters required to train large supervised models from scratch such as Enformer (250 million parameters). We have improved clarity throughout the manuscript to emphasize the foundation's ability of the model and the computational benefits of our efficient fine-tuning strategy PEFT on a new task compared with the training of supervised models from scratch (such as Enformer), being cheaper, faster and requiring only 0.1% of the total number of parameters allowing to fine-tune even our biggest models in less than 15 minutes on a single GPU.

The comparison between our unsupervised language model approach and the Enformer supervised model approach is a very good idea and should be very informative for the field about the value and foundational ability of large unsupervised pre-trained models. To address this, we have compared our unsupervised pre-trained model with fine-tuning the Enformer model, which was pre-trained in a supervised fashion on almost 7,000 genomics tasks from human and mouse (including TF binding, chromatin features and gene expression). Comparing both approaches on our 18 downstream tasks showed that Enformer fine-tuned models can perform well on tasks related to the pre-trained tasks (promoters, enhancers, and histone marks) but are not able to generalize to unrelated tasks such as splicing (see new Fig. 2a,b). Still, our multispecies 2.5B model achieved the best performance in all tasks, matched by Enformer only for the prediction of promoters with TATA-box motif (Fig. 2a,b, Supplementary Table 6). We have added this and other comparisons with different pre-trained language models (detailed in our response to the comment below) in the revised manuscript, many thanks.

> Language modeling methods for DNA such as DNABert and BigBERT were previously proposed; yet,

> a comparison against those models is lacking in the paper. The overall benchmarking is not sufficient. Methods they are benchmarking against are not SOTA. We suggest that the following methods need to be benchmarked. Authors should compare their model against SpliceAI for splicing-related acceptor and donor prediction. For transcription factor binding, compare to newer methods besides DeepSEA: Catchitt, DeepGRN, FactorNet,Anchor, and compare histone mark prediction against Basenji, and BPNet. Also, they can include deep learning models such as Enformer in QTL variant evaluation.

We agree with the reviewer that the manuscript would benefit from more comparisons with other pre-trained language models and additional baselines.
Following the suggestions of both reviewers, we have added to the revised manuscript the fine-tuned performance on the 18 downstream tasks of five different pre-trained models: DNABERT-1 [1], DNABERT-2 [2], HyenaDNA (1kb and 32kb) [3] and Enformer (which was trained as a supervised model on several genomics tasks; [4]). As described above, compared with DNABERT, HyenaDNA-32kb and Enformer, our multispecies 2.5B model achieved the best performance in all tasks (see Fig. 2a,b, Supplementary Table 6,8). DNABERT-1 and Enformer achieved this top performance only for the prediction of promoters with TATA-box motif, while DNABERT-2 matched top performance in 9 out of the 18 tasks. Compared with HyenaDNA-1kb the results were more comparable: our 2.5B model has lower performance in 5 tasks, matches performance in other 5 tasks, and shows better performance in 8 out of the 18 tasks. This is the most extensive benchmark of genomics foundational models so far and will serve as the main reference for future studies.

For the additional baselines, we already included in the manuscript comparisons with baselines for 18 different downstream tasks, as well as DeepSTARR, a state-of-the-art method for enhancer predictions, and DeepSEA, a pioneer method and dataset of chromatin features. We agree that additional comparisons are useful for the readers and have now added comparisons of Nucleotide Transformer with SpliceAI ([5]), for predicting splicing donor and acceptor sites (see new Fig. 2d and 5d,e). This showed that the Nucleotide Transformer model predicts splice sites with 95% top-k accuracy and 0.98 precision-recall AUC, which is comparable with the state-of-the-art SpliceAI method. In addition, we tested our Nucleotide Transformer v2 500M model that can take 12kb-long input sequences and observed an improved performance of 1% to a top-k accuracy of 96% and a precision-recall AUC of 0.98, an overall improved performance over SpliceAI (Fig. 5e).
In addition, we have also compared our efficient fine-tuning approach with full-model fine-tuning and observed no significant improvement for chromatin and splicing predictions and a small 3% improvement for enhancer activity predictions (Supplementary Fig. 2), supporting the use of our efficient fine-tuning Nucleotide Transformer approach.
We acknowledge that there are additional methods that would be interesting to add to these comparisons, but also that we cannot provide every comparison in the manuscript. We hope this reviewer agrees that we did a substantial effort in the manuscript to provide extensive benchmarks for both the pre-training and finetuning aspects of our models. Comparisons with additional methods will be added in future work.

> Authors must include confidence intervals while reporting performance in all analyses. Without confidence intervals, authors cannot claim that their model is better or worse than alternative models.

We apologize if it was not clear but we have included confidence intervals based on 10-fold cross-validation for the predictions of every downstream task (see Fig 2). We have now also included the confidence intervals in Supplementary Table 6. We have also created a leaderboard table with the results of all folds for ours and all other benchmarked language models, accessible at HuggingFace[5]. This will allow the readers to have access to all the results and different summary statistics metrics, and will be augmented over time with future models, serving as an arena for genomics language models.

> Does transfer learning from embedding really help? Random embedding performance, just sequence based model performance without embedding over training epochs needs to be reported. (Similar as in ref [9]).

[5]HuggingFace: https://huggingface.co/spaces/InstaDeepAI/nucleotide_transformer_benchmark

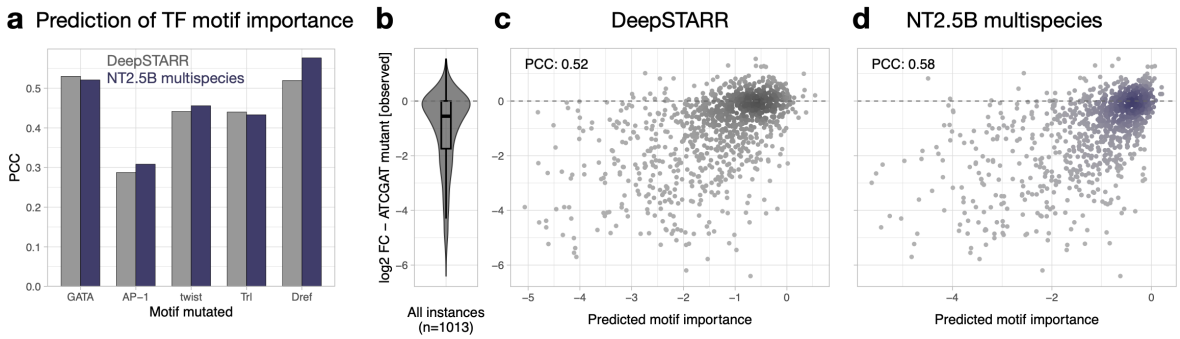
We thank the reviewer for this great suggestion. We have now compared the performance between fine-tuning from pre-trained models or from randomly initialized models, showing that pre-training leads to significantly better performance (see new Supplementary Table 8). For completeness, we have also compared with probing from raw tokens, showing a strong increase in performance in all tasks (Supplementary Table 8). These new results strongly support our unsupervised pre-training approach, and we have added them to the revised manuscript (see page 3), many thanks.

> Authors may visualize model interpretation tools such as visualize model gradients per input base pair and convince the reader by showing their model learning know regulatory motifs such as splice dinucleotides AG for acceptor and GT donor, TATA box for promoter site, AATAAA poly(A)-signal for alternative polyadenylation sites. However, learning those regulatory elements is trivial after fine-tuning. It would be interesting to see if the pre-trained model with fine-tuning can capture those regulatory elements.

We agree with the reviewer that the interpretation of additional sequence-rules learned by the Nucleotide Transformer model is relevant for the readers. The visualization of the regulatory motifs and features as mentioned by the reviewer is a very good idea. We have added additional analyses to interpret the pre-trained and fine-tuned models in the revised manuscript. As described above, and to do not repeat all the details (please see our detailed response above to #reviewer1 first major comment), we have now looked at token reconstruction probabilities across the different types of genomics elements and regulatory motifs as a measure of sequence constraint and importance learned by the model. Specifically, we calculated the 6-mer token probabilities (based on masking each token at a time) for every 6kb window in chr22 (available as a WashU Epigenome Browser session[6] for further investigation by the readers). This showed that the pre-trained model learned a variety of gene structure and regulatory elements such as acceptor and donor splicing sites, polyA signals, CTCF binding sites and others (Supplementary Fig. 16a-d). Furthermore, we compared our token predictions with an experimental saturation mutagenesis splicing assay of exon 11 of the gene MST1R (data from Braun et al. [6]). This revealed a significant correlation between the experimental mutation effects and the token predictions by our multispecies 2.5B pre-trained model (Pearson Correlation Coefficient (PCC) = 0.44; Supplementary Fig. 16e), which learned well constraints at the different splicing junctions but also at a region in the middle of the second intron important for the splicing of this exon. This is a strong validation of the biological knowledge acquired by the Nucleotide Transformer pre-trained model.

For the interpretation of the fine-tuned models, we have chosen the enhancer activity fine-tuned model since there is an extensive experimental motif-mutation enhancer dataset available in the DeepSTARR paper [8]. Here, we took the multispecies 2.5B model full-finetuned on the DeepSTARR enhancer activity data and asked if the model learned about TF motifs and their relative importance for enhancer activity. We used a dataset of experimental mutation of hundreds of individual instances of five different motif types across hundreds of enhancer sequences [8] and tested the accuracy of our model in predicting those mutation effects. Compared with the state-of-the-art enhancer activity DeepSTARR model, our model achieves similar performance for four TF motifs and demonstrates higher performance for the Dref TF motif (Supplementary Fig. 17 and below). These results illustrate how the both pre-trained and fine-tuned models have learned about gene structure and functional properties of genomic sequences.

[6] WashU Epigenome Browser: https://shorturl.at/jov28



We have integrated these new analyses and results in the revised manuscript (see Supplementary Figs. 16 and 17 and page 8). Many thanks for the suggestions for more detailed model interpretation that have improved the manuscript.

Minor comments:

> The model is trying to remember the context of a 6kb region during training. Does this provide any information on the interactions?

We have evaluated this by analysing the attention maps. Indeed the model attention specifically targets the different types of genomic elements throughout its different heads and layers, such as 5'UTRs, exons, introns, 3'UTRs, open chromatin regions, promoters, enhancers, transcription factor (TF) binding regions and CTCF regions. We have improved the attention analysis to highlight this (see page 8, Fig. 3f,g and Supplementary Fig. 7-15). It will be interesting in the future to further explore interactions between pairs of elements, particularly now that we managed to build models trained on 12kb context, where an input sequence can contain many of the elements mentioned above.

> Fig 1a, not clear what sizes and colors mean. What are the models, what is the dataset is better to be clearly labeled.
> Better to draw the model in a figure.

We have revised this figure and now show in Fig. 1 a revised model cartoon, together with a summary of model size, perception field and downstream task performance for all foundational models tested, many thanks.

> Details of the model (variables, detailed architecture) are not clearly presented. Better use tables/figures/equations to describe the details.

We apologize if it was not clear but we have provided all model hyperparameters and details in Supplementary Table 1, in addition to their description in the Methods section.

> Not clear what fig 2b is. What is the meaning of the lighter and darker color?.

We have simplified this figure and revised its legend to add the new results and improve clarity. We now provide a summary performance metric for each model in single bars.

> Fig 4b is difficult to read to similar blue tones. Please use a colorblind friendly palette. Also, It better shows quantitative results as well.

We agree and have updated this panel.

References

[1] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021.

[2] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, "Dnabert-2: Efficient foundation model and benchmark for multi-species genome," *arXiv preprint arXiv:2306.15006*, 2023.

[3] E. Nguyen, M. Poli, M. Faizi, A. Thomas, C. Birch-Sykes, M. Wornow, A. Patel, C. Rabideau, S. Massaroli, Y. Bengio, *et al.*, "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution," *arXiv preprint arXiv:2306.15794*, 2023.

[4] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley, "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature methods*, vol. 18, no. 10, pp. 1196–1203, 2021.

[5] K. Jaganathan, S. K. Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, *et al.*, "Predicting splicing from primary sequence with deep learning," *Cell*, vol. 176, no. 3, pp. 535–548, 2019.

[6] S. Braun, M. Enculescu, S. T. Setty, M. Cortés-López, B. P. de Almeida, F. R. Sutandy, L. Schulz, A. Busch, M. Seiler, S. Ebersberger, *et al.*, "Decoding a cancer-relevant splicing decision in the ron proto-oncogene using high-throughput mutagenesis," *Nature communications*, vol. 9, no. 1, p. 3315, 2018.

[7] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, , *et al.*, "Scaling language models: Methods, analysis insights from training gopher," *arXiv preprint arXiv:2112.11446*, 2022.

[8] B. P. de Almeida, F. Reiter, M. Pagani, and A. Stark, "Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers," *Nature Genetics*, vol. 54, no. 5, pp. 613–624, 2022.

Nucleotide Transformer Rebuttal

Summary of the revision

We are grateful to the two reviewers for their insightful comments and helpful suggestions on our manuscript. We have revised the manuscript to address all comments and suggestions and improve the overall clarity. The main changes and additions are:

- Systematic comparison of our model to five different pre-trained models: DNABERT-1, DNABERT-2, HyenaDNA (1kb and 32kb) and Enformer
- Two additional controls for our fine-tuning approach: probing from raw tokens and fine-tuning from randomly initialized checkpoint
- Additional benchmark datasets and comparison with baselines such as SpliceAI
- Additional interpretation analyses of pre-trained and fine-tuned Nucleotide Transformer models
- Development of a new version of Nucleotide Transformer models (NT-v2) that achieve the same performance while being ten times smaller and having a context length twice as large at 12kb
- Example notebooks to apply these models to any downstream task available on HuggingFace[1]

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

Reviewer #1 - September 19th 2023 :

Remarks to the Author:

> The authors have proposed a deep learning model, the Nucleotide Transformer, for unsupervised pre-training of sequence models, and applied it to various prediction tasks. This study also assessed the impact of using diverse genome datasets for pre-training transformer models on DNA sequences. The authors trained multiple models on human reference genomes, 1000 genomes, and 850 genomes from other species like fungi, invertebrates, etc. The authors demonstrated that this approach matches or outperforms the baselines on 12 out of 18 tasks and up to 15 after fine-tuning. The method demonstrated similar performance to DeepSEA or DeepSTARR when fine-tuned on corresponding datasets. Although the authors did a commendable software engineering job, making it possible to train such large models, there are some weaknesses that need to be addressed. These include a lack of strong evidence for advancing over the current state-of-the-art, as well as demonstrating the advantage of training on diverse genomes and unsupervised pretraining, as I detailed below.

We thank the reviewer for the summary and all the constructive comments that allowed us to improve the manuscript further. We thank the reviewer for bringing to our attention that we have not sufficiently clearly explained that the main goal of this work was not to show a strong improvement over the state-of-the-art of every genomics model, but rather to develop a flexible approach that is very adaptable across many diverse genomics tasks, and introduce proper benchmarks that can pave the way for the development of foundational models in genomics. Indeed, since we made our model and benchmark datasets available in March, they were already used as references for the development of further models such as DNABERT-2 and HyenaDNA, and our HuggingFace datasets have been downloaded up to 14k times in one month.

[1] https://huggingface.co/docs/transformers/notebooks#pytorch-bio

We have now improved clarity in the manuscript to emphasize the foundation's ability of the model, methodology and benchmarks. As described in detail below, we have also added more extensive comparisons to other pre-trained models and baselines to demonstrate the advantage of unsupervised training. We also provide notebooks to allow anyone to quickly fine-tune the model on their datasets.

Major points:

> 1. In terms of novelty, the nucleotide transformer approach is very similar to DNABERT. Although the Nucleotide Transformer model is larger and trained on sequences beyond the human reference genome, it is still unclear whether its performance is better. Therefore, a comparison with DNABERT is needed.

We agree with the reviewer that a comparison with other pre-trained models such as DNABERT would be helpful. To provide an extensive comparison and benchmarking of different unsupervised training approaches, we have now added to the manuscript the fine-tuned performance on the 18 downstream tasks of five different pre-trained models: DNABERT-1 [1], DNABERT-2 [2], HyenaDNA (1kb and 32kb context length models; [3]) and Enformer (which was trained as a supervised model on several genomics tasks; [4]). For a fair comparison, we have used the same parameter-efficient fine-tuning approach for every model and used the same cross-validation scheme. Compared with DNABERT and Enformer, our multispecies 2.5B model achieved the best performance in all tasks (see Fig. 2a,b, Supplementary Table 6,8, and below). DNABERT-1 and Enformer achieved this top performance only for the prediction of promoters with TATA-box motif, while DNABERT-2 matched top performance in 9 out of the 18 tasks. Compared with HyenaDNA-1kb the results were more comparable: our 2.5B model has lower performance in 5 tasks, matches performance in other 5 tasks, and shows better performance in 8 out of the 18 tasks. Interestingly, while HyenaDNA was pre-trained on the human reference genome, 6 of the 8 tasks where our multisplecies 2.5B model performs better are from human-related datasets, highlighting again the advantage of pre-training on a diverse set of genome sequences. In addition, the longer-input HyenaDNA-32kb model had worse performance in all 18 tasks, showing that there is a trade-off between increasing the input context length and performance in high-resolution tasks, such as the chromatin/splicing-related genomics tasks used here. We have added these new results in the revised main text, methods, figures and tables, many thanks.
We have also created an interactive leader-board with the results of all models across every task for an easier comparison, accessible at HuggingFace[2]. This is the most extensive benchmark of genomics foundational models so far and should serve as the main reference for future studies.

Model	500M HR	500M 1000G	2.5B 1000G	2.5B MS	DNABERT-1	DNABERT-2	Enformer	hyenadna-tiny-1k	hyenadna-small-32k
H3	0.718 (± 0.014)	0.736 (± 0.012)	0.754 (± 0.008)	**0.793 (± 0.013)**	0.763 (± 0.013)	**0.785 (± 0.012)**	0.724 (± 0.018)	**0.781 (± 0.015)**	0.747 (± 0.017)
H3K14ac	0.372 (± 0.006)	0.381 (± 0.011)	0.453 (± 0.008)	0.538 (± 0.009)	0.403 (± 0.012)	0.515 (± 0.009)	0.284 (± 0.024)	**0.608 (± 0.02)**	0.405 (± 0.015)
H3K36me3	0.448 (± 0.012)	0.468 (± 0.008)	0.53 (± 0.012)	**0.618 (± 0.011)**	0.474 (± 0.009)	0.591 (± 0.005)	0.345 (± 0.019)	**0.614 (± 0.014)**	0.479 (± 0.015)
H3K4me1	0.361 (± 0.015)	0.38 (± 0.009)	0.418 (± 0.012)	**0.541 (± 0.005)**	0.396 (± 0.005)	0.512 (± 0.008)	0.291 (± 0.016)	0.512 (± 0.008)	0.387 (± 0.017)
H3K4me2	0.27 (± 0.015)	0.26 (± 0.016)	0.278 (± 0.015)	0.324 (± 0.014)	0.282 (± 0.013)	0.333 (± 0.013)	0.207 (± 0.021)	**0.455 (± 0.028)**	0.276 (± 0.01)
H3K4me3	0.236 (± 0.018)	0.235 (± 0.008)	0.311 (± 0.012)	0.408 (± 0.011)	0.258 (± 0.01)	0.353 (± 0.021)	0.156 (± 0.022)	**0.55 (± 0.015)**	0.291 (± 0.025)
H3K79me3	0.566 (± 0.016)	0.562 (± 0.015)	0.574 (± 0.007)	0.623 (± 0.01)	0.578 (± 0.007)	0.615 (± 0.01)	0.498 (± 0.013)	**0.669 (± 0.014)**	0.567 (± 0.013)
H3K9ac	0.451 (± 0.011)	0.479 (± 0.01)	0.491 (± 0.009)	0.547 (± 0.011)	0.505 (± 0.009)	0.545 (± 0.009)	0.415 (± 0.02)	**0.586 (± 0.021)**	0.472 (± 0.01)
H4	0.75 (± 0.011)	0.755 (± 0.011)	0.787 (± 0.006)	**0.808 (± 0.007)**	0.784 (± 0.007)	**0.797 (± 0.008)**	0.735 (± 0.023)	0.763 (± 0.012)	0.761 (± 0.017)
H4ac	0.332 (± 0.017)	0.342 (± 0.014)	0.408 (± 0.009)	0.492 (± 0.014)	0.359 (± 0.01)	0.465 (± 0.013)	0.275 (± 0.022)	**0.564 (± 0.011)**	0.379 (± 0.012)
enhancers	**0.502 (± 0.026)**	**0.509 (± 0.033)**	**0.546 (± 0.029)**	**0.545 (± 0.028)**	0.495 (± 0.016)	**0.525 (± 0.026)**	0.454 (± 0.029)	**0.52 (± 0.031)**	0.489 (± 0.018)
enhancers types	**0.429 (± 0.018)**	**0.395 (± 0.027)**	0.432 (± 0.03)	**0.444 (± 0.022)**	0.367 (± 0.034)	**0.423 (± 0.018)**	0.312 (± 0.043)	**0.403 (± 0.056)**	0.352 (± 0.04)
promoter all	0.952 (± 0.002)	0.951 (± 0.003)	0.965 (± 0.002)	**0.975 (± 0.002)**	0.961 (± 0.001)	0.972 (± 0.002)	0.955 (± 0.002)	0.959 (± 0.002)	0.956 (± 0.002)
promoter no tata	0.952 (± 0.002)	0.951 (± 0.002)	0.967 (± 0.001)	**0.977 (± 0.002)**	0.962 (± 0.002)	0.972 (± 0.002)	0.955 (± 0.003)	0.959 (± 0.002)	0.954 (± 0.003)
promoter tata	0.942 (± 0.006)	0.936 (± 0.006)	0.957 (± 0.007)	**0.959 (± 0.004)**	**0.956 (± 0.006)**	**0.955 (± 0.006)**	**0.959 (± 0.006)**	**0.944 (± 0.011)**	0.939 (± 0.008)
splice sites acceptors	0.962 (± 0.003)	0.965 (± 0.002)	0.98 (± 0.002)	**0.986 (± 0.002)**	NaN	0.975 (± 0.002)	0.915 (± 0.007)	0.959 (± 0.003)	0.96 (± 0.007)
splice sites all	0.97 (± 0.002)	0.968 (± 0.001)	0.976 (± 0.002)	**0.982 (± 0.002)**	0.975 (± 0.002)	0.939 (± 0.003)	0.847 (± 0.005)	0.956 (± 0.004)	0.962 (± 0.004)
splice sites donors	0.971 (± 0.002)	0.971 (± 0.001)	0.979 (± 0.001)	**0.987 (± 0.001)**	NaN	0.963 (± 0.002)	0.906 (± 0.008)	0.947 (± 0.007)	0.957 (± 0.006)

> 2. The authors reported that their model matched or outperformed 15/18 baselines on performance. However, it should be noted that these comparisons do not include many state-of-the-art models. Furthermore, the model generally did not surpass stronger baselines, DeepSEA and DeepSTARR, and lacks comparison with current state-of-the-art methods.

For instance, the authors mentioned the Enformer but did not compare the results of the Nucleotide Transformer with those of Enformer in predicting DNASE/ATAC-seq, CAGE, Histone modifications, etc. Similarly, Sei is a recent model reported to outperform DeepSEA. Given the comparison between Nucleotide Transformer and DeepSEA, it is likely that Sei will outperform Nucleotide Transformer on

[2]HuggingFace: https://huggingface.co/spaces/InstaDeepAI/nucleotide_transformer_benchmark

these tasks. Other examples of state-of-the-art models include SpliceAI for splicing variant prediction and BPNet for basepair resolution Chip-Nexus predictions. While it may not be necessary to compare the Nucleotide Transformer to all these methods, some comparison would be needed.

We agree with the reviewer that the manuscript would benefit from more comparisons with additional baselines. We note that in addition to the 18 main downstream tasks used in the paper, we already included in the original manuscript comparisons with DeepSTARR, a state-of-the-art method for enhancer predictions, and DeepSEA, a pioneer method and dataset of chromatin features. We have now added extensive comparisons of Nucleotide Transformer models with SpliceAI [5] for predicting splicing donor and acceptor sites, as suggested by the reviewer (see new Fig. 2d, 5d,e). We adapted the Nucleotide Transformer 2.5B model to deliver nucleotide-level splice site predictions (Fig. 5d) and achieved a top-k accuracy of 95% and a precision-recall AUC of 0.98 with our 6kb-context model (Fig. 2d). Notably, this performance matches the state-of-the-art SpliceAI-10k, which was trained on 15kb input sequences, and outperformed SpliceAI when tested on 6kb input sequences. We have also tested our Nucleotide Transformer v2 500M model that can take 12kb-long input sequences and observed an improved performance of 1% to a top-k accuracy of 96% and a precision-recall AUC of 0.98, an improved performance over SpliceAI (Fig. 5e).

In addition, we have also compared our efficient fine-tuning approach with full-model fine-tuning and observed no significant improvement for chromatin and splicing predictions and a small 3% improvement for enhancer activity predictions (Supplementary Fig. 2), supporting the use of our efficient fine-tuning Nucleotide Transformer approach. We hope this reviewer agrees that we did a substantial effort in the manuscript to provide extensive benchmarks for both the pre-training and finetuning aspects of our models. Comparisons with additional methods will be added in future work.

3. There is a lack of evidence that multispecies training or training on multiple genomes provides an advantage in performance after considering issues in the comparisons. The main problem with this comparison is that models trained on human genomes should not be evaluated on non-human datasets (most of the 18 datasets contain some non-human data).

Most notably, the performance advantage of the multi-species model in Figure 2a mostly comes from yeast histone mark datasets. Given the significant difference in biology between yeast and humans, it is not very meaningful to compare human models using yeast datasets. The observation that human genome-trained models do not perform well on non-human datasets does not constitute a strong argument for the advantage of the multi-species model.
If we exclude the yeast datasets, there is no obvious advantage from the model trained on multi-species genomes. If we exclude other non-human data, it is not unlikely that no advantage can be observed.

Additionally, it is unclear whether a 1000 genomes-trained model outperforms the model trained on only human genome data for models with the same number of parameters, if we exclude the yeast datasets.

We agree that we should evaluate further the human-trained models and the multispecies model on the different types of tasks. To make this comparison more clear in the manuscript, we have now divided the tasks based on species and task category (yeast chromatin profiles, human regulatory elements and human splicing tasks) and compared the different models. This shows that the multispecies 2.5B model achieves the best performance in every task - both human- and non-human ones (see new Fig. 2a,b, Supplementary Table 6). In addition, the multispecies model was also the best on prioritizing pathogenic variants (Fig. 4c,d). These results show that leveraging sequence variation across species, and likely their degree of conservation, is beneficial to build robust pre-trained sequence models.

We also evaluated the performance of the Human-ref and 1000G models (both with 500M parameters) on the human datasets to evaluate the importance of the diversity of sequences. Indeed, both models showed similar performance across the different human-data downstream tasks (Fig. 2a, Supplementary Table 6). We have adapted Figure 2 and discussed better these results in the revised manuscript.

> It is also preferable to use whole chromosome-based holdouts similar to what the authors used for variant effect prediction comparisons, rather than 10-fold cross-validation without taking spatial adjacency into consideration.

We agree with the reviewer that it is better to use chromosome-based holdouts for cross-validation. However, to have a fair comparison to the baselines we used the same test set as in the original publications. For example for DeepSTARR and DeepSEA the train and test datasets are indeed splitted based on chromosomes. We have clarified this in the methods.

> 4. The authors have not shown strong evidence that unsupervised pre-training improves performance. It should be noted that, unlike language modeling, which has a vast amount of unlabeled data, the human genome is well annotated, and most datasets have genome-wide coverage. Therefore, the benefit of unsupervised pre-training cannot be assumed.

We thank the reviewer for pointing out that we have not clearly shown the value of unsupervised pre-training. We have supported this point through multiple approaches in the revised manuscript. (1) We have compared the performance between fine-tuning from pre-trained models or from randomly initialized models, showing that pre-training leads to significantly better performance (see new Supplementary Table 6,8). (2) We have compared our unsupervised pre-trained model with fine-tuning the Enformer model, which was pre-trained in a supervised fashion on almost 7,000 genomics tasks from human and mouse (including TF binding, chromatin features and gene expression). This showed that Enformer fine-tuned models can perform well on tasks related to the pre-trained tasks (promoters, enhancers, and histone marks) but are not able to generalize to unrelated tasks such as splicing (see new Fig. 2a,b). (3) We have clarified in the manuscript the computational benefits of our efficient fine-tuning strategy PEFT on a new task compared with the training of supervised models from scratch for every different task, being cheaper, faster and requiring only 0.1% of the total number of parameters allowing to fine-tune even our biggest models in less than 15 minutes on a single GPU.

We think that this additional evidence strongly supports the benefits of unsupervised pre-training, many thanks.

> This work showed that the pre-trained model matches but does not outperform previous models trained in a supervised fashion, especially for stronger baselines trained on bigger datasets. This still leaves open the question of the benefit of unsupervised training.

We agree with the reviewer that we don't show in this manuscript a strong improvement over previous supervised models, but we want to clarify that we did not focus here on maximising the performance of our fine-tuned models on these downstream tasks. Instead, the main goal of this work was to develop a flexible approach that enables it to tackle many diverse genomics tasks, which contrasts with supervised models that can only be trained on one task at a time and differ in architecture and parameters between every task. This is a key advantage over previous models trained in a supervised fashion, making the training of future models much easier. To demonstrate that, we have also added a comparison with fine-tuning the Enformer model on the different tasks to highlight the foundation's ability of our model.

Finally, we note that genomics language models are very recent, and we expect that further research on new architectures, larger models and more diverse pre-training data will lead to strong improvements in genomics predictions tasks, as it is being observed for older problems in NLP and more recently in protein prediction tasks. We also expect that such foundational models will be more useful for datasets for which there is no genome-wide data available, such as the activity of regulatory elements in vivo or the functional impact of genetic variants. Our work will serve as a reference point in that direction. We have elaborated on this in the revised manuscript (see page 14).

> One potential opportunity for unsupervised pretraining is, maybe it can improve performance when a limited number of positives in training data are available. It is also notable that a recent preprint showed that an unsupervised model can predict variant effect on species without experimental data.

We agree with the reviewer that unsupervised pre-trained models should have stronger benefits for smaller datasets. We have demonstrated this on the prediction of deleterious variants, where datasets are relatively small. Here, our fine-tuned models either slightly outperformed or closely matched the performance of other models based on conservation or functional features (Fig. 4d).
Additionally, an important advantage of unsupervised pre-trained models over supervised models is their zero-shot predictive value, as also demonstrated in the paper cited by the reviewer. Indeed, we show that our zero-shot scores have a very good predictive value for prioritizing functional variants (AUCs across different variant types from 0.7 to 0.8), even matching or surpassing other methods on eQTLs and meQTLs tasks (see Fig. 4c,d). We have discussed this in the revised manuscript (see page 14).

> 5. For the evaluation of reconstruction accuracy on variants, if I understand correctly, does not seem to be the appropriate setup. Since the model trained on 1000 Genomes can simply memorize allele frequencies, showing that it reconstructs variants from a different dataset does not necessarily show that the model learned to use sequence in a non-trivial way to predict variants. Evaluation on holdout chromosomes is better for this evaluation.

We thank the reviewer for pointing out that this was not sufficiently clearly explained. The goal of this analysis is to test the imputation capability of our pre-trained language model - i.e. to assess if a model pre-trained in genomes from a given diverse population (and respective allele frequencies) can generalize to reconstruct the variants of unseen human genome sequences. Similar to imputation methods, this needs to be evaluated in the same genomic regions seen during pre-training, but in genomes with different haplotypes, as we did here for an independent dataset of genetically diverse human genomes. Evaluating this capability in a holdout chromosome would likely not work because the model wouldn't be able to predict variants and haplotype if not from previously observed LD. We have clarified this point in the respective section of the main text (page 7), many thanks.

Minor points:

> 1. I am not sure if the maximum attention percentages in some layers reaching 100% for enhancers is particularly meaningful, as all element types seem to have some layers reaching 100%.

We understand the reviewer's comment. Layers reaching 100% do not occur for all element types, but they do occur for 3'UTR, promoter and TF binding sites, in addition to enhancers. We think this information is meaningful but have added the other elements. In addition, we have also highlighted now how the different element types are captured by different heads and layers of the same model (see page 8). We have also improved this results section and respective Fig. 3f,g to highlight better these attention interpretation analyses.

> 2. Comparison with DeepSEA in Figure 4d does not seem appropriate. For example, the nucleotide transformer was fine-tuned on each dataset, but DeepSEA was not. DeepSEA is also not intended for predicting coding mutations or splicing variants, which are the majority of mutations in ClinVar or HGMD.

We thank the reviewer for pointing out that our explanations had been unclear. It is indeed true that in Fig 4d we fine-tuned the Nucleotide Transformer models on each variant dataset. However, regarding DeepSEA, we used the DeepSEA (Beluga version)-predicted Disease Impact Score (DIS) (REF). This score is based on a logistic regression model that prioritizes likely disease-associated mutations on the basis of the predicted transcriptional or post-transcriptional regulatory effects of these mutations (See Zhou et. al, 2019), and is used in the respective paper to predict the functional impact of genetic variants in gene expression, including eQTLs. Thus we think that it is a relevant method for comparison in the variant prediction tasks where we evaluate our models.
In addition, we agree with the reviewer that the DeepSEA scores are related mostly to transcriptional effects, which explains the lower performance of DeepSEA in ClinVar and HGMD tasks. We have added it for consistency, to evaluate all methods across the 4 variant tasks (eQTL, meQTL, ClinVar, HGMD). We have now clarified how the DeepSEA scores were calculated and used in the revised main

text and methods for clarity.

> 3. It seems that the definition of $p_a(f)$ should be the proportion of attention scores greater than, rather than less than, the threshold. The statistical test for $p_a(f)$ seems to be not very meaningful, given that any threshold can be significant. Given the sample size, I don't think a p-value is needed for this analysis.

We thank the reviewer for pointing out this typo in the formula and have corrected it.
Regarding the statistical test, we still think it's a informative metric to identify heads that significantly attend to the genomic elements and have decided to keep it for completeness.

> 4. More information could be given on how the 1000 genomes dataset was constructed. It should also be noted that the reconstructed genome does not contain all variants.

We have revised the Methods section to describe how the 1000 genomes dataset was constructed, many thanks.

Reviewer #2 - September 19th 2023 :

Remarks to the Author:

> In their paper, the authors introduce the Nucleotide Transformer as a foundational model for genomics. The authors describe the Nucleotide Transformer, pitched as a foundation model for genomics. Nucleotide Transformer follows the BERT architecture used for training large language models (LLM). Nucleotide Transformer models are trained in a way similar to BERT, asking the model to predict masked tokens, in this case, 6-mers. This idea is that sequence embedding learned from NT can be used for downstream genomic tasks with probing or fine-tuning.
> Building on the success of LLMs, there are recent a deluge of papers on similar ideas, building foundation models for genomics, utilizing sequence only information such as DNABert, BIGBERT, GenSLM, GPN (from Yun Song's group),... or a combination of sequence or functional annotations such as Enformer, BPNet, ... So neither method nor idea is new here.
>
> Instead, the authors focus on exploring the impact of sizes of the model and diversity of the datasets. Specifically, authors leveraged genomics sequences from multiple species and attempted to show the merit of training from diverse species. While potentially interesting, the presented work is an incremental contribution building on existing works. There are major concerns about the design of the model, as well as the validity and effectiveness of the conclusions drawn from the papers.

We thank the reviewer for the summary of our work and the context in the field, and for all the constructive comments that allowed us to improve the manuscript further. We thank the reviewer for pointing out that we have not sufficiently clearly explained how our work differs from recent foundation models for genomics. Although there have been similar methods developed during the past two years (DNABERT, BigBird, GenSLMs, GPN, HyenaDNA), we are the first to (1) systematically evaluate different model sizes (50M up to 2.5B parameters) and different pre-training datasets (the human reference genome, a collection of 3,202 diverse human genomes, and 850 genomes from several species), and (2) to introduce proper benchmarks that can pave the way for the development of foundational models in genomics. Indeed, since we made our work and benchmarks available in March, they were already used as references for the development of further models such as DNABERT-2 and HyenaDNA, and our HuggingFace datasets have been downloaded up to 14k times in one month.

To further improve our work and establish it as the reference genomics LLM model paper in the field, we have added in the revised manuscript an extensive comparison and benchmarking to five different pre-trained models (DNABERT-1 [1], DNABERT-2 [2], HyenaDNA (1kb and 32kb context length models; [3]) and Enformer (which was pre-trained as a supervised model on several genomics tasks; [4])), two of them posted online during this revision. Furthermore, we have developed a new version

of Nucleotide Transformer models that are able to achieve the same performance than the 2.5B model while being 10 times smaller and having a context length twice as large at 12kb.

We have now improved clarity in the manuscript to emphasize how our work differs from other genomics LLMs, including our methodology and benchmarks. In addition, we also provide notebooks to allow anyone to quickly fine-tune our model on their datasets. Many thanks.

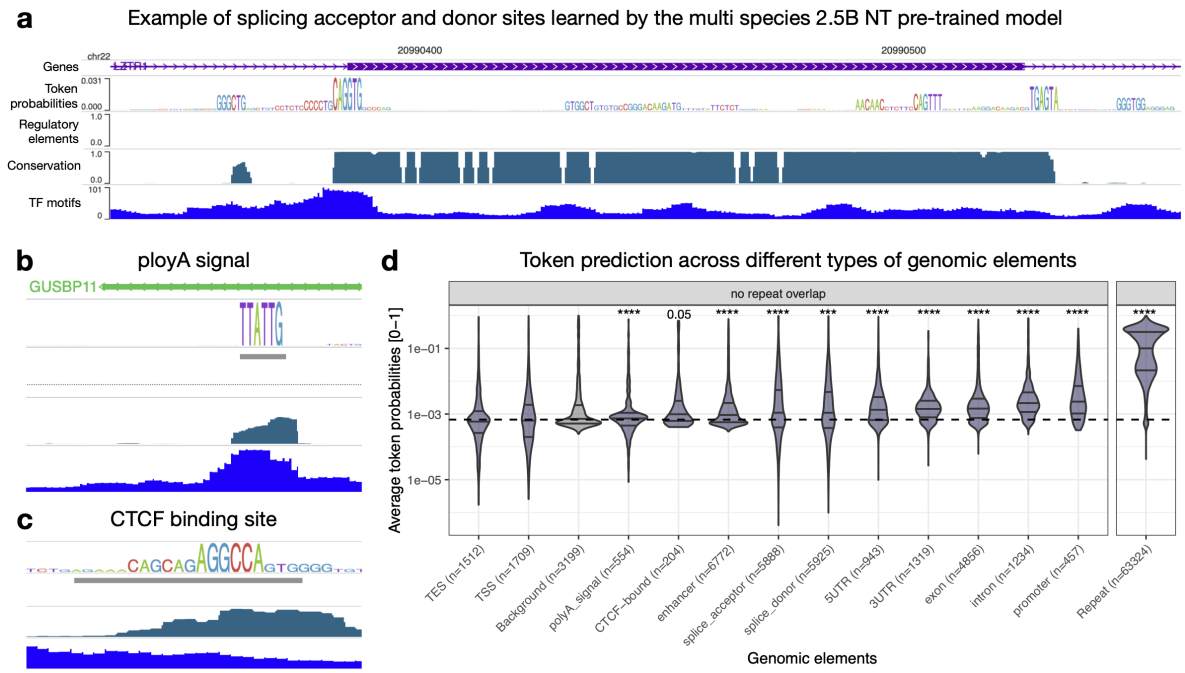
My major comments are:

> At the most basic level, I am not fully convinced that a naive application of BERT type of models, developed for NLP, to DNA sequences is a productive approach. In NLP, the tokens have meanings and are naturally defined. It is unclear what "token" means in the context of DNA sequence. There is no such thing as simple functional units in DNA, especially for noncoding sequences. The human genome is littered with many different kinds of functional elements buried in a sea of nonfunctional sequences. Other than coding sequences or RNA genes, it has been very difficult to properly define and model other types of functional elements. Demarking the boundary of non coding functional elements is even harder. There is also the issue of "grammar" or "rules" with the DNA sequences. Despite years of research, it is still not clear what they are and how general the rules are. So I would like to have some understanding of what kind of knowledge is learned by the Nucleotide Transformer. It is important for authors to look deeper into the interpretation side to find out what their models have learned. Is that information supported by prior biological knowledge?

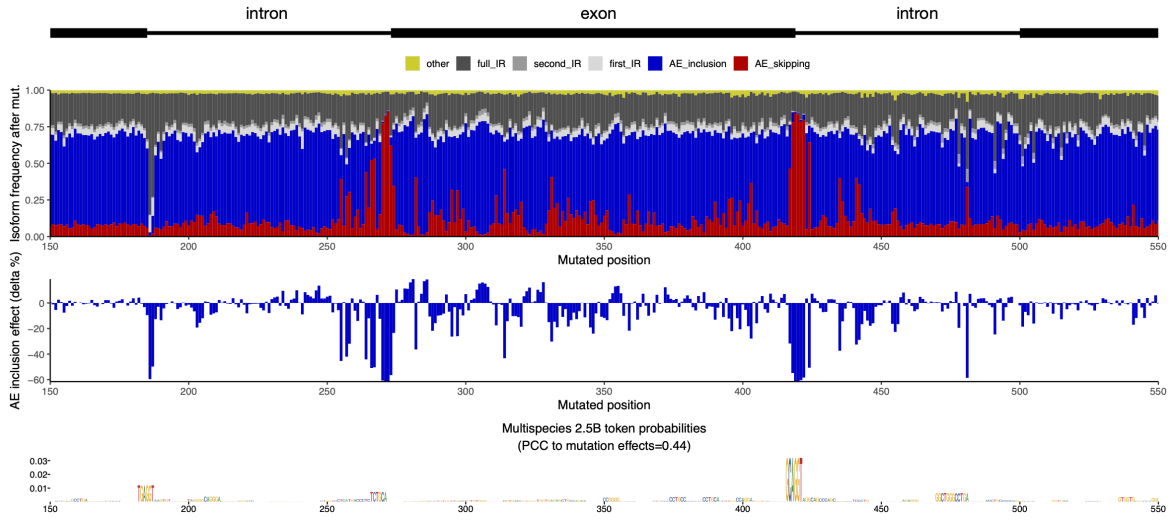
We agree with the reviewer that the interpretation of the sequence-rules learned by the Nucleotide Transformer is relevant for our understanding of how these LLMs encode genomic sequences. We note that interpretation of such models remains very challenging and is an active field of research. Still, we have addressed this in the original manuscript using several state-of-the-art techniques. We showed that the DNA sequences of 5 different types of genomic elements (3'UTR, CDS, intron, intergenic, 5'UTR) are encoded in the model through unique and very distinct sequence embeddings at the different layers (see Fig 3d and Supplementary Fig. 6). In addition, we performed extensive analyses of the attention maps to understand which sequence regions are captured and used by the attention layers (see page 8, Fig. 3f,g and Supplementary Fig. 7-15). We have now improved these analyses to show that the model attention specifically targets the different types of genomic elements throughout its different layers, such as 5'UTRs, exons, introns, 3'UTRs, open chromatin regions, promoters, enhancers, transcription factor (TF) binding regions and CTCF regions. These different element types are captured by different heads and layers, where the maximum attention percentages were highest for the largest models, reaching for instance a value of almost 100% attention for enhancers in the 1000G 2.5B model.

We have also added now additional analyses to interpret the pre-trained models at higher resolution (i.e. more local sequence features). We have looked at token probabilities across the genome and different types of genomics elements as a measure of sequence constraint and importance learned by the language model. Specifically, we calculated the 6-mer token probabilities (based on masking each token at a time) for every 6kb chunk in chr22. Analysing this token reconstruction metric revealed that in addition to repetitive elements, that are well reconstructed by the model as expected, the pre-trained model learned a variety of gene structure and regulatory elements. For instance, in panel (a) we show a screenshot of an exon where the model predicts well the acceptor and donor splicing sites. In panels (b) and (c) we show examples of polyA signal and CTCF binding sites, respectively. The original nucleotides are shown with their height proportional to the token probabilities. We have also created an interactive browser session with token probabilities across chromosome 22 on the WashU Epigenome Browser[3] to allow further exploration by the readers. We summarised the token probabilities across different types of genomics elements and show that the pre-trained model significantly learned about the structure and sequence features of different types of elements (panel (d)).

[3]WashU Epigenome Browser: https://shorturl.at/jov28



Furthermore, we compared our token reconstruction predictions with an experimental saturation mutagenesis splicing assay of exon 11 of the gene MST1R (data from Braun et al. [6]). The top panel shows the measured impact in the different splicing isoforms of mutating every nucleotide position, with the effect in Alternative Exon (AE) 11 inclusion showed in the middle panel. Note how the expected important positions (such as splicing acceptor and donor) match the positions with the strongest mutation impact. Importantly, we observed a significant correlation between these experimental mutation effects and the token predictions by our multispecies 2.5B pre-trained model (Pearson Correlation Coefficient (PCC) = 0.44). Our model learns well constraints at the different splicing junctions, but also a region in the middle of the second intron that is indeed important for the splicing of this exon. This is a strong validation of the biological knowledge acquired by our model during unsupervised pre-training.



We have integrated these new analyses and results in the revised manuscript (see Supplementary Fig. 16). These results illustrate how the transformer models have learned to recover gene structure and functional properties of genomic sequences and integrate them directly into its attention mechanism. We hope this reviewer agrees that we have done our best to interpret the biological knowledge encoded in these models. We acknowledge that the interpretation of such models remains challenging and an intense field of research. Future studies will allow us to improve their interpretation and reveal novel

biological principles.

> One of the main points that the authors argue for is the large models with millions of parameters. Big models have shown utility for LLMs. But I am not convinced it is a good idea for training models for DNA sequences.Language or vision are compositional. The number of training samples is unbounded. This kind of diversity is crucial for training LLMs. By contrast, the diversity associated with the human genome is limited, even accounting for genetic variants. With only 3 billion letters in the human genome, a large model with hundreds of millions of parameters can simply memorize the genome.

Genomics LLMs learn statistical relationships between DNA sequence-features and sequence types from the vast amount of sequences they've seen during training. The goal is not to memorize the training data, but to generalize from the sequence relationships learned during training to unseen sequences or downstream tasks. Indeed, we have monitored overfitting during pre-training by having a held-out validation set of sequences and observed no overfitting. In addition, while memorization by LLMs is an open research area in NLP and can indeed be an issue for zero-shot evaluations, it will not lead to overfitting for downstream prediction tasks like the ones shown in this manuscript. Even if the model does memorize all DNA sequences, as the data is unlabeled, it does not help to perform the downstream tasks. Actually, most probably overfitting would affect the downstream performance.

> Can authors provide a convincing argument and experimental support on why larger models are better for DNA sequences?

We thank the reviewer for pointing out that this was not sufficiently clearly explained. We have assessed the importance of model size by training four distinct language models of different sizes, ranging from 500M up to 2.5B parameters. We assessed the ability of each model on 18 different genomic prediction datasets and observed that the largest models outperformed their smaller counterparts (Fig 2a,b). In addition, we have added in the revised manuscript a second version of the Nucleotide Transformer models, with a new architecture. Again, we have tested this new model architecture with different sizes ranging from 50M to 500M parameters and show that the largest model achieves the best performance (see Fig. 51-c). These results match observations in other fields such as NLP, where larger training datasets and model sizes have been shown to improve performance [7].
Regarding the experimental support, all our models of different sizes were evaluated in 18 different downstream tasks whose datasets derive from experimental work. Our work is the first to evaluate the importance of model size in genomics LLMs. We have revised the manuscript to clarify the different points described above and hope the reviewer is convinced that our flexible LLM approach is promising to model all the different types of DNA sequences.

> There is insufficient review of prior works and discuss how the current work differs from prior ones. Authors may improve their introduction by briefly mentioning existing task specific (Enformer, BPNet, ...) and general nucleotide language models (DNABERT, ...). Then, they can describe their added values compared to their proposed model. For example, if their model size or multi-species training is the major difference from state of the art models. Additionally, authors should compare their contributions against state of the art by conducting further benchmarks in the result section to emphasize the added values of their work.

We thank this reviewer for this suggestion that helped us improve the clarity of the manuscript. We have extended the review of previous work in the introduction and mentioned additional task-specific models (BPNet, Sei, ORCA) and pre-trained language models that were released during the revision of our manuscript (DNABERT2, GENA-LM and HyenaDNA) (see page 2). We note that we have already mentioned the pioneers supervised Enformer model and unsupervised language model DNABERT1 in the previous manuscript version.
In addition, we have emphasized throughout the manuscript the foundation's ability of the model, its methodology, the different model sizes and training sets evaluated, and the systematic benchmark datasets and analyses that we provide.
We also agree with the reviewer that the manuscript would benefit from more comparisons with additional baselines. We already included in the manuscript comparisons with DeepSTARR, a state-

of-the-art method for enhancer predictions, and DeepSEA, a pioneer method and dataset of chromatin features. We have now added extensive comparisons of Nucleotide Transformer models with SpliceAI [5] for predicting splicing donor and acceptor sites, as suggested by the reviewer (see new Fig. 2d, 5d,e). We adapted the Nucleotide Transformer 2.5B model to deliver nucleotide-level splice site predictions (Fig. 5d) and achieved a top-k accuracy of 95% and a precision-recall AUC of 0.98 with our 6kb-context model (Fig. 2d). Notably, this performance matches the state-of-the-art SpliceAI-10k, which was trained on 15kb input sequences, and outperformed SpliceAI when tested on 6kb input sequences. We have also tested our Nucleotide Transformer v2 500M model that can take 12kb-long input sequences and observed an improved performance of 1% to a top-k accuracy of 96% and a precision-recall AUC of 0.98, an improved performance over SpliceAI (Fig. 5e).

As suggested by both reviewers, and to provide an extensive comparison and benchmarking of different unsupervised training approaches, we have now added to the manuscript the fine-tuned performance on the 18 downstream tasks of five different pre-trained models: DNABERT-1 [1], DNABERT-2 [2], HyenaDNA (1kb and 32kb context length models; [3]) and Enformer (which was trained as a supervised model on several genomics tasks; [4]). We have used the same parameter-efficient fine-tuning approach for every model and the same cross-validation scheme to enable direct comparison of the results. Compared with DNABERT and Enformer, our multispecies 2.5B model achieved the best performance in all tasks (see Fig. 2a,b, Supplementary Table 6,8, and below). DNABERT-1 and Enformer achieved this top performance only for the prediction of promoters with TATA-box motif, while DNABERT-2 matched top performance in 9 out of the 18 tasks. Compared with HyenaDNA-1kb the results were more comparable: our 2.5B model has lower performance in 5 tasks, matches performance in other 5 tasks, and shows better performance in 8 out of the 18 tasks. Interestingly, while HyenaDNA was pre-trained on the human reference genome, 6 of the 8 tasks where our multisplecies 2.5B model performs better are from human-related datasets, highlighting again the advantage of pre-training on a diverse set of genome sequences. In addition, the longer-input HyenaDNA-32kb model had worse performance in all 18 tasks, showing that there is a trade-off between increasing the input context length and performance in high-resolution tasks, such as the chromatin/splicing-related genomics tasks used here.
We have created an interactive leader-board with the results of all models across every task for an easier comparison, accessible at HuggingFace[4]. This is the most extensive benchmark of genomics foundational models so far and should serve as the main reference for future studies.

Thank you for all the suggestions.

Model	500M HR	500M 1000G	2.5B 1000G	2.5B MS	DNABERT-1	DNABERT-2	Enformer	hyenadna-tiny-1k	hyenadna-small-32k
H3	0.718 (± 0.014)	0.736 (± 0.012)	0.754 (± 0.008)	**0.793 (± 0.013)**	0.763 (± 0.013)	**0.785 (± 0.012)**	0.724 (± 0.018)	**0.781 (± 0.015)**	0.747 (± 0.017)
H3K14ac	0.372 (± 0.006)	0.381 (± 0.011)	0.453 (± 0.008)	0.538 (± 0.009)	0.403 (± 0.012)	0.515 (± 0.009)	0.284 (± 0.024)	**0.608 (± 0.02)**	0.405 (± 0.015)
H3K36me3	0.448 (± 0.012)	0.468 (± 0.008)	0.53 (± 0.012)	**0.618 (± 0.011)**	0.474 (± 0.009)	0.591 (± 0.005)	0.345 (± 0.019)	**0.614 (± 0.014)**	0.479 (± 0.015)
H3K4me1	0.361 (± 0.015)	0.38 (± 0.009)	0.418 (± 0.012)	**0.541 (± 0.005)**	0.396 (± 0.005)	0.512 (± 0.008)	0.291 (± 0.016)	0.512 (± 0.008)	0.387 (± 0.017)
H3K4me2	0.27 (± 0.015)	0.26 (± 0.016)	0.278 (± 0.015)	0.324 (± 0.014)	0.282 (± 0.013)	0.333 (± 0.013)	0.207 (± 0.021)	**0.455 (± 0.028)**	0.276 (± 0.01)
H3K4me3	0.236 (± 0.018)	0.235 (± 0.008)	0.311 (± 0.012)	0.408 (± 0.011)	0.258 (± 0.01)	0.353 (± 0.021)	0.156 (± 0.022)	**0.55 (± 0.015)**	0.291 (± 0.025)
H3K79me3	0.566 (± 0.016)	0.562 (± 0.015)	0.574 (± 0.007)	0.623 (± 0.01)	0.578 (± 0.007)	0.615 (± 0.01)	0.498 (± 0.013)	**0.669 (± 0.014)**	0.567 (± 0.013)
H3K9ac	0.451 (± 0.011)	0.479 (± 0.01)	0.491 (± 0.009)	0.547 (± 0.011)	0.505 (± 0.009)	0.545 (± 0.009)	0.415 (± 0.02)	**0.586 (± 0.021)**	0.472 (± 0.01)
H4	0.75 (± 0.011)	0.755 (± 0.011)	0.787 (± 0.006)	**0.808 (± 0.007)**	0.784 (± 0.007)	**0.797 (± 0.008)**	0.735 (± 0.023)	0.763 (± 0.012)	0.761 (± 0.017)
H4ac	0.332 (± 0.017)	0.342 (± 0.014)	0.408 (± 0.009)	0.492 (± 0.014)	0.359 (± 0.01)	0.465 (± 0.013)	0.275 (± 0.022)	**0.564 (± 0.011)**	0.379 (± 0.012)
enhancers	**0.502 (± 0.026)**	**0.509 (± 0.033)**	**0.546 (± 0.029)**	**0.545 (± 0.028)**	0.495 (± 0.016)	**0.525 (± 0.026)**	0.454 (± 0.029)	**0.52 (± 0.031)**	0.489 (± 0.018)
enhancers types	**0.429 (± 0.018)**	**0.395 (± 0.027)**	0.432 (± 0.03)	**0.444 (± 0.022)**	0.367 (± 0.034)	**0.423 (± 0.018)**	0.312 (± 0.043)	**0.403 (± 0.056)**	0.352 (± 0.04)
promoter all	0.952 (± 0.002)	0.951 (± 0.003)	0.965 (± 0.002)	**0.975 (± 0.002)**	0.961 (± 0.001)	0.972 (± 0.002)	0.955 (± 0.002)	0.959 (± 0.002)	0.956 (± 0.002)
promoter no tata	0.952 (± 0.002)	0.951 (± 0.002)	0.967 (± 0.001)	**0.977 (± 0.002)**	0.962 (± 0.002)	0.972 (± 0.002)	0.955 (± 0.003)	0.959 (± 0.002)	0.954 (± 0.003)
promoter tata	0.942 (± 0.006)	0.936 (± 0.006)	0.957 (± 0.007)	**0.959 (± 0.004)**	**0.956 (± 0.006)**	**0.955 (± 0.006)**	**0.959 (± 0.006)**	**0.944 (± 0.011)**	0.939 (± 0.008)
splice sites acceptors	0.962 (± 0.003)	0.965 (± 0.002)	0.98 (± 0.002)	**0.986 (± 0.002)**	NaN	0.975 (± 0.002)	0.915 (± 0.007)	0.959 (± 0.003)	0.96 (± 0.007)
splice sites all	0.97 (± 0.002)	0.968 (± 0.001)	0.976 (± 0.002)	**0.982 (± 0.002)**	0.975 (± 0.002)	0.939 (± 0.003)	0.847 (± 0.005)	0.956 (± 0.004)	0.962 (± 0.004)
splice sites donors	0.971 (± 0.002)	0.971 (± 0.001)	0.979 (± 0.001)	**0.987 (± 0.001)**	NaN	0.963 (± 0.002)	0.906 (± 0.008)	0.947 (± 0.007)	0.957 (± 0.006)

> I have some concerns regarding the evaluations and experimental results. First, evaluation tasks should be expanded to include recent functional genomic data from ENCODE or Epigenomic projects. The authors used the dataset of DeepSEA, which was curated several years ago. There are more recent and richer datasets that the authors should evaluate.

We agree with the reviewer that there have been more recent and comprehensive genomics chromatin datasets. We chose to use the original DeepSEA dataset because (1) its 919 classification tasks provide a manageable dataset size but still diverse set of transcription factor, DNA accessibility and histone modification tasks across different cell types, and (2) because it has been used in genomics

[4]HuggingFace: https://huggingface.co/spaces/InstaDeepAI/nucleotide_transformer_benchmark

as a reference dataset across the last years, with performance metrics reported across many different methodologies. While we could have done the benchmarks in a larger dataset version, we expected that the performance difference to the baselines would be the same, and decided that it would not be worth the extended computational cost of such experiments.

> Second, the improvement of their models compared to the so-called "SOTA" are marginal in most cases. So this raises the important question on the value of the proposed models. This is also in stark contrast to the results from LLMs, where foundation models offer more substantial improvements.

We thank the reviewer for bringing to our attention that we have not sufficiently clearly explained that the main goal of this work was not to show a strong improvement over the state-of-the-art of every genomics model, but rather to develop a flexible approach that is very adaptable across many diverse genomics tasks, and introduce proper benchmarks that can pave the way for the development of foundational models in genomics. Our foundational models contrast with supervised models that can only be trained on one task at a time and differ in architecture and parameters between every task. This is a key advantage over previous models trained in a supervised fashion, making the training of future models much easier.
To elaborate on this, we have added a comparison with fine-tuning the Enformer model on the different tasks to highlight the foundation's ability of our model. To further evaluate the potential of pre-trained models, we have also fine-tuned the whole Nucleotide Transformer model (not using PEFT as before) on the more challenging downstream tasks (DeepSTARR, DeepSEA and SpliceAI) and showed an improvement over baseline supervised models.

Finally, we note that genomics language models are very recent, and we expect that further research on new architectures, larger models and more diverse pre-training data will lead to strong improvements in genomics predictions tasks, as it is being observed for older problems in NLP and more recently in protein prediction tasks. We also expect that such foundational models will be more useful for datasets for which there is no genome-wide data available, such as the activity of regulatory elements in vivo or the functional impact of genetic variants. Our work will serve as a reference point in that direction. We have improved clarity throughout the manuscript to emphasize the foundation's ability of the model, methodology and benchmarks, as well as the computational benefits of our efficient fine-tuning strategy PEFT on a new task compared with the training of supervised models from scratch for every different task, being cheaper, faster and requiring only 0.1% of the total number of parameters allowing to fine-tune even our biggest models in less than 15 minutes on a single GPU.

> Authors utilized 6-mer based tokenization. In natural language modeling, word-based tokenization is optimal because words are the smallest meaningful units constructing a sentence. However, DNA grammar does not contain units like words and instead contains regulatory motifs. Earlier genomics models prior to deep learning such as gkmSVM and kmer-SVM k-mer based methods are widely used. However, convolution based models have become the standard because convolutional deep learning models such as DeepSea, Basenji, or Enformer have superior performance. because convolutional layers learn to tokenize regulatory elements from raw DNA sequences. In the manuscript, the authors do not provide any justification for 6-mer based tokenization over the use of convolution. Their architecture choice has major computational cost and limits the receptive field of the model to 6,000 bp.

We thank the reviewer for pointing out that we have not provided a clear justification for the use of 6-mer based tokenization. As the reviewer writes in the introduction above, it is still not clear what and how general the "grammar" or "rules" of DNA sequences actually are. However, the regulatory code is commonly compared with a form of language, taking the regulatory motifs as "words" and their interactions as the sequence "grammar" of the regulatory code. The benefit of sequence tokenization is that we don't need to make any assumption about the DNA code structure. Similar to NLP, the tokenization is not done for tokens to have meaning but just to find a trade-off between sequence length and vocab size. The "meaning" is captured later by embeddings during training. In fact, most of the state-of-the-art NLP models don't just take words as tokens but rather use some kind of subword-tokenization algorithms. Similarly, there is probably not a unique token approach that will work the best for every genomics problem, and it will depend on how the embeddings are learned

during training.
Here we have decided to take a tokenization approach based on a trade-off between sequence length (up to 6kb) and embedding size. We have experimented with different k-mer sizes and observed the highest performance with 6-mers (data not shown). Similarly, DNABERT also found 6-mer based tokenization to achieve the best performance for DNA tasks (see DNABERT paper). We have added more details regarding our tokenization approach in the revised manuscript.
In addition, we have developed a new version of Nucleotide Transformer models (NT-v2), by leveraging contemporary architectural advancements and extending training duration, that are able to achieve the same performance than the 2.5B model while being 10 times smaller and having a context length twice as large at 12kb.

> Authors need to compare alternative architectures such as a model with convolution layers followed by attention layers.
> Thereby, they may increase the receptive field and reduce the model training cost. If their current architecture outperforms the alternative architectures then their model choice would be justified. Also, masking-based semi-supervised training is compatible with convolutional models; for example, authors may utilize intermediate layer masking after convolutional layers but before attention layers.

The comparison with an architecture based on convolutional layers is a very interesting point given the success of convolutional models in genomics, and could indeed justify our model choice. While developing a new pre-trained model based on convolutional layers and train it from scratch is out of scope of this manuscript, following the reviewer's suggestion, we have compared our unsupervised pre-trained model (purely based on embeddings and attention) with fine-tuning the Enformer model, which combines convolution layers with attention and was pre-trained in a supervised fashion on almost 7,000 genomics tasks from human and mouse (including TF binding, chromatin features and gene expression). Comparing both approaches on our 18 downstream tasks showed that Enformer fine-tuned models can perform well on tasks related to the pre-trained tasks (promoters, enhancers, and histone marks) but are not able to generalize to unrelated tasks such as splicing (see new Fig. 2a,b). Still, our multispecies 2.5B model achieved the best performance in all tasks, matched by Enformer only for the prediction of promoters with TATA-box motif. Given that Enformer is considered a state-of-the-art model in the field, this analysis is of great interest, and we have added it in the revised manuscript, many thanks.

> SVM bases models have a number of ways similar to final probing layers of the proposed model. Thus, authors may compare probing and gkmSVM despite the method may not be the state of the art.

We thank the reviewer for this suggestion. However, as the reviewer mentions, SVM-based models have been outperformed by deep learning models, such as the ones we use as baselines in this manuscript. Since we have already expanded our comparisons with 4 additional pre-trained language models in the revised manuscript, we have decided to not include SVM-based models in our comparisons for conciseness.

> The state of the art models repeatedly show that increasing the model receptive field significantly improves the performance because sequence context is critical to understand DNA grammar [7, 8]. Their 6 kb sequence length is smaller compared to state of the art models in genomics. Authors may try different lengths such as 1kb, 2kb, .etc, and show that model performance converges with 6 kb. Alternatively, Authors may increase the receptive field with convolution as discussed in the previous comment.

We agree with the reviewer that the length of the model receptive field is an important aspect for genomics models, where longer sequence context is leading to better results. We have managed to increase it to 6kb for the Nucleotide Transformer V1 models (compared for instance with 512bp for DNABERT-1 and GPN), and to 12kb for the Nucleotide Transformer V2 model thanks to its more efficient architecture. We want to clarify that finding the best possible architecture and sequence length was not the purpose of this study, and that 6kb or 12kb are not said to be the best input lengths. These were still chosen due to computing limitations and we are working on extending them in further

models.
Still, increasing the input context length is not a trivial task, as also demonstrated by our evaluation of HyenaDNA. The HyenaDNA models were developed to have context lengths up to 1 million bases, but our benchmarking results show that the longer-input HyenaDNA-32kb model had worse performance in all 18 tasks than our 2.5B model and the shorter-input HyenaDNA-1kb model (Fig. 2a,b). This shows that there is a trade-off between increasing the input context length and the performance in high-resolution tasks, such as the chromatin/splicing-related genomics tasks used here. This is an area of intense research, and we and others are currently working on efficiently extending the context length of this type of genomics foundational models. We have added this notion to the main text in the revised manuscript (see page 4).

> Authors claim that an increasing number of parameters in the model is helpful; however, authors haven't performed benchmarks based on different parameter numbers so they did not demonstrate this point.

We thank the reviewer for pointing out that this was not sufficiently clearly explained. We have assessed the importance of model size by training four distinct language models of different sizes, ranging from 500M up to 2.5B parameters. We assessed the ability of each model on 18 different genomic prediction datasets and observed that the largest models outperformed their smaller counterparts (Fig. 2a,b). In addition, we have added in the revised manuscript a second version of the Nucleotide Transformer models, with a new architecture. Again, we have tested this new model architecture with different sizes ranging from 50M to 500M parameters and show that the largest model achieves the best performance (see Fig. 5a-c, Supplementary Table 6,8). These results match observations in other fields such as NLP, where larger training datasets and model sizes have been shown to improve performance [7]. We have emphasized this analysis in the revised manuscript, many thanks.

> Moreover, state of the art models such as Enformer contain significantly far fewer parameters.
> Also, authors may show their large models are helpful by benchmarking against state of the art models such as Enformer.

We thank the reviewer for pointing out that we have not sufficiently clearly explained the difference between the number of parameters in finetuned unsupervised language models and supervised models from scratch. It is true that our pre-trained models have a large number of parameters (the largest having 2.5 billion), however with our efficient fine-tuning strategy PEFT the number of parameters that are finetuned are only a very small proportion of 0.1% (2.5 million parameters for the 2.5B model). This contrasts with the high number of parameters required to train large supervised models from scratch such as Enformer (250 million parameters). We have improved clarity throughout the manuscript to emphasize the foundation's ability of the model and the computational benefits of our efficient fine-tuning strategy PEFT on a new task compared with the training of supervised models from scratch (such as Enformer), being cheaper, faster and requiring only 0.1% of the total number of parameters allowing to fine-tune even our biggest models in less than 15 minutes on a single GPU.

The comparison between our unsupervised language model approach and the Enformer supervised model approach is a very good idea and should be very informative for the field about the value and foundational ability of large unsupervised pre-trained models. To address this, we have compared our unsupervised pre-trained model with fine-tuning the Enformer model, which was pre-trained in a supervised fashion on almost 7,000 genomics tasks from human and mouse (including TF binding, chromatin features and gene expression). Comparing both approaches on our 18 downstream tasks showed that Enformer fine-tuned models can perform well on tasks related to the pre-trained tasks (promoters, enhancers, and histone marks) but are not able to generalize to unrelated tasks such as splicing (see new Fig. 2a,b). Still, our multispecies 2.5B model achieved the best performance in all tasks, matched by Enformer only for the prediction of promoters with TATA-box motif (Fig. 2a,b, Supplementary Table 6). We have added this and other comparisons with different pre-trained language models (detailed in our response to the comment below) in the revised manuscript, many thanks.

> Language modeling methods for DNA such as DNABert and BigBERT were previously proposed; yet,

> a comparison against those models is lacking in the paper. The overall benchmarking is not sufficient. Methods they are benchmarking against are not SOTA. We suggest that the following methods need to be benchmarked. Authors should compare their model against SpliceAI for splicing-related acceptor and donor prediction. For transcription factor binding, compare to newer methods besides DeepSEA: Catchitt, DeepGRN, FactorNet,Anchor, and compare histone mark prediction against Basenji, and BPNet. Also, they can include deep learning models such as Enformer in QTL variant evaluation.

We agree with the reviewer that the manuscript would benefit from more comparisons with other pre-trained language models and additional baselines.
Following the suggestions of both reviewers, we have added to the revised manuscript the fine-tuned performance on the 18 downstream tasks of five different pre-trained models: DNABERT-1 [1], DNABERT-2 [2], HyenaDNA (1kb and 32kb) [3] and Enformer (which was trained as a supervised model on several genomics tasks; [4]). As described above, compared with DNABERT, HyenaDNA-32kb and Enformer, our multispecies 2.5B model achieved the best performance in all tasks (see Fig. 2a,b, Supplementary Table 6,8). DNABERT-1 and Enformer achieved this top performance only for the prediction of promoters with TATA-box motif, while DNABERT-2 matched top performance in 9 out of the 18 tasks. Compared with HyenaDNA-1kb the results were more comparable: our 2.5B model has lower performance in 5 tasks, matches performance in other 5 tasks, and shows better performance in 8 out of the 18 tasks. This is the most extensive benchmark of genomics foundational models so far and will serve as the main reference for future studies.

For the additional baselines, we already included in the manuscript comparisons with baselines for 18 different downstream tasks, as well as DeepSTARR, a state-of-the-art method for enhancer predictions, and DeepSEA, a pioneer method and dataset of chromatin features. We agree that additional comparisons are useful for the readers and have now added comparisons of Nucleotide Transformer with SpliceAI ([5]), for predicting splicing donor and acceptor sites (see new Fig. 2d and 5d,e). This showed that the Nucleotide Transformer model predicts splice sites with 95% top-k accuracy and 0.98 precision-recall AUC, which is comparable with the state-of-the-art SpliceAI method. In addition, we tested our Nucleotide Transformer v2 500M model that can take 12kb-long input sequences and observed an improved performance of 1% to a top-k accuracy of 96% and a precision-recall AUC of 0.98, an overall improved performance over SpliceAI (Fig. 5e).
In addition, we have also compared our efficient fine-tuning approach with full-model fine-tuning and observed no significant improvement for chromatin and splicing predictions and a small 3% improvement for enhancer activity predictions (Supplementary Fig. 2), supporting the use of our efficient fine-tuning Nucleotide Transformer approach.
We acknowledge that there are additional methods that would be interesting to add to these comparisons, but also that we cannot provide every comparison in the manuscript. We hope this reviewer agrees that we did a substantial effort in the manuscript to provide extensive benchmarks for both the pre-training and finetuning aspects of our models. Comparisons with additional methods will be added in future work.

> Authors must include confidence intervals while reporting performance in all analyses. Without confidence intervals, authors cannot claim that their model is better or worse than alternative models.

We apologize if it was not clear but we have included confidence intervals based on 10-fold cross-validation for the predictions of every downstream task (see Fig 2). We have now also included the confidence intervals in Supplementary Table 6. We have also created a leaderboard table with the results of all folds for ours and all other benchmarked language models, accessible at HuggingFace[5]. This will allow the readers to have access to all the results and different summary statistics metrics, and will be augmented over time with future models, serving as an arena for genomics language models.

> Does transfer learning from embedding really help? Random embedding performance, just sequence based model performance without embedding over training epochs needs to be reported. (Similar as in ref [9]).

[5]HuggingFace: https://huggingface.co/spaces/InstaDeepAI/nucleotide_transformer_benchmark

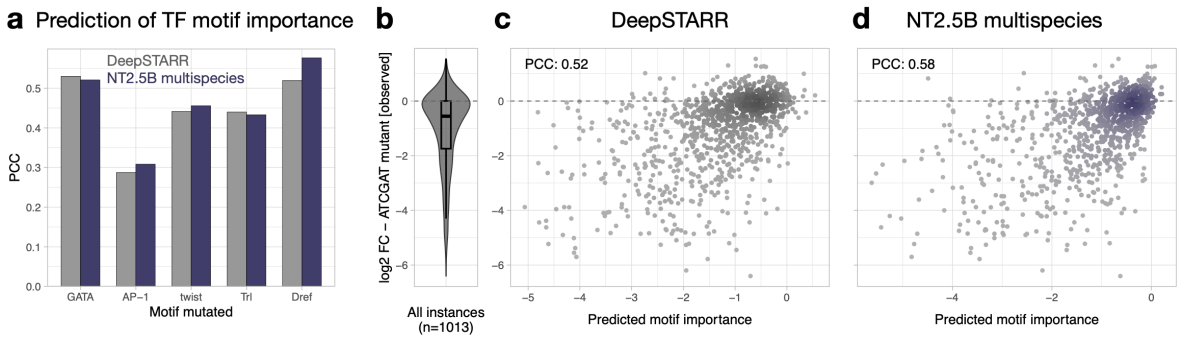
We thank the reviewer for this great suggestion. We have now compared the performance between fine-tuning from pre-trained models or from randomly initialized models, showing that pre-training leads to significantly better performance (see new Supplementary Table 8). For completeness, we have also compared with probing from raw tokens, showing a strong increase in performance in all tasks (Supplementary Table 8). These new results strongly support our unsupervised pre-training approach, and we have added them to the revised manuscript (see page 3), many thanks.

> Authors may visualize model interpretation tools such as visualize model gradients per input base pair and convince the reader by showing their model learning know regulatory motifs such as splice dinucleotides AG for acceptor and GT donor, TATA box for promoter site, AATAAA poly(A)-signal for alternative polyadenylation sites. However, learning those regulatory elements is trivial after fine-tuning. It would be interesting to see if the pre-trained model with fine-tuning can capture those regulatory elements.

We agree with the reviewer that the interpretation of additional sequence-rules learned by the Nucleotide Transformer model is relevant for the readers. The visualization of the regulatory motifs and features as mentioned by the reviewer is a very good idea. We have added additional analyses to interpret the pre-trained and fine-tuned models in the revised manuscript. As described above, and to do not repeat all the details (please see our detailed response above to #reviewer1 first major comment), we have now looked at token reconstruction probabilities across the different types of genomics elements and regulatory motifs as a measure of sequence constraint and importance learned by the model. Specifically, we calculated the 6-mer token probabilities (based on masking each token at a time) for every 6kb window in chr22 (available as a WashU Epigenome Browser session[6] for further investigation by the readers). This showed that the pre-trained model learned a variety of gene structure and regulatory elements such as acceptor and donor splicing sites, polyA signals, CTCF binding sites and others (Supplementary Fig. 16a-d). Furthermore, we compared our token predictions with an experimental saturation mutagenesis splicing assay of exon 11 of the gene MST1R (data from Braun et al. [6]). This revealed a significant correlation between the experimental mutation effects and the token predictions by our multispecies 2.5B pre-trained model (Pearson Correlation Coefficient (PCC) = 0.44; Supplementary Fig. 16e), which learned well constraints at the different splicing junctions but also at a region in the middle of the second intron important for the splicing of this exon. This is a strong validation of the biological knowledge acquired by the Nucleotide Transformer pre-trained model.

For the interpretation of the fine-tuned models, we have chosen the enhancer activity fine-tuned model since there is an extensive experimental motif-mutation enhancer dataset available in the DeepSTARR paper [8]. Here, we took the multispecies 2.5B model full-finetuned on the DeepSTARR enhancer activity data and asked if the model learned about TF motifs and their relative importance for enhancer activity. We used a dataset of experimental mutation of hundreds of individual instances of five different motif types across hundreds of enhancer sequences [8] and tested the accuracy of our model in predicting those mutation effects. Compared with the state-of-the-art enhancer activity DeepSTARR model, our model achieves similar performance for four TF motifs and demonstrates higher performance for the Dref TF motif (Supplementary Fig. 17 and below). These results illustrate how the both pre-trained and fine-tuned models have learned about gene structure and functional properties of genomic sequences.

[6] WashU Epigenome Browser: https://shorturl.at/jov28



We have integrated these new analyses and results in the revised manuscript (see Supplementary Figs. 16 and 17 and page 8). Many thanks for the suggestions for more detailed model interpretation that have improved the manuscript.

Minor comments:

> The model is trying to remember the context of a 6kb region during training. Does this provide any information on the interactions?

We have evaluated this by analysing the attention maps. Indeed the model attention specifically targets the different types of genomic elements throughout its different heads and layers, such as 5'UTRs, exons, introns, 3'UTRs, open chromatin regions, promoters, enhancers, transcription factor (TF) binding regions and CTCF regions. We have improved the attention analysis to highlight this (see page 8, Fig. 3f,g and Supplementary Fig. 7-15). It will be interesting in the future to further explore interactions between pairs of elements, particularly now that we managed to build models trained on 12kb context, where an input sequence can contain many of the elements mentioned above.

> Fig 1a, not clear what sizes and colors mean. What are the models, what is the dataset is better to be clearly labeled.
> Better to draw the model in a figure.

We have revised this figure and now show in Fig. 1 a revised model cartoon, together with a summary of model size, perception field and downstream task performance for all foundational models tested, many thanks.

> Details of the model (variables, detailed architecture) are not clearly presented. Better use tables/figures/equations to describe the details.

We apologize if it was not clear but we have provided all model hyperparameters and details in Supplementary Table 1, in addition to their description in the Methods section.

> Not clear what fig 2b is. What is the meaning of the lighter and darker color?.

We have simplified this figure and revised its legend to add the new results and improve clarity. We now provide a summary performance metric for each model in single bars.

> Fig 4b is difficult to read to similar blue tones. Please use a colorblind friendly palette. Also, It better shows quantitative results as well.

We agree and have updated this panel.

Reviewer #1 - January 30th 2024 :

Remarks to the Author:

> In the revision, the authors have performed more comparisons with similar pretrained models like DNABERT and HyenaDNA, and added additional benchmarks comparing with Enformer and SpliceAI. However, many of the main concerns raised by reviewers remained inadequately addressed. Some of the new results still suffer from the issues that we pointed out such as training on the human genome and evaluating on yeast data. Below I reiterate two major concerns for this manuscript, and many other main points in my first review were inadequately addressed in my opinion.

We genuinely appreciate the reviewer's engagement with our manuscript and the constructive feedback provided. We would like to review and address the various points outlined by the reviewer.

Our primary objective in this manuscript was to establish methodologies and discoveries related to constructing and assessing language models for nucleic acid sequences. These models have thrived in other fields such as natural language, vision, and proteomics, and we believe that one of the main reasons for their success was the establishment of such guidelines. In the genomics field, these models are relatively new and therefore not yet fully developed. We agree with the reviewer and acknowledge, as supported by our research, that in their current state, DNA-based language models do not outperform specialized supervised methods on problems where a large amount of data is available, such as models trained on ENCODE data. This point has been explicitly addressed in the revised manuscript(refer to pages 2, 3, 14).

We believe it is crucial to study and enhance these models as they hold the potential to advance genomics on various fronts. The Nucleotide Transformer work stands as the first to establish rigorous protocols and propose an initial collection of evaluation datasets in this domain, and as such we believe it is an important contribution to this objective and the field as a whole. Additionally, we have developed diverse pre-training datasets using human and multispecies genomes, which have subsequently been utilized for pre-training other foundational models (e.g. DNABERT-2).

The impact of our work is validated by the fact that the paper has already been cited 39 times despite not being published in a journal yet, and the various models introduced in this manuscript have been downloaded thousands, even tens of thousands of times per month for the downstream task datasets. We apologize if these points were not adequately conveyed in the original paper, and we have accordingly toned down certain sections of the text in the revised manuscript (see pages 2, 3, 14).

Major points:

> 1. The current 18-task benchmark provides an inadequate and misleading representation of performance. It is not clear why models trained on human genomes should be evaluated on yeast data, as these two organisms have drastically different biology. The rest of the benchmarks also mostly include mostly small datasets (e.g. enhancers) or easy tasks that do not represent current challenges in the field (e.g. promoter prediction). Since these benchmarks are used for most of the results in this manuscript, it is difficult to draw meaningful conclusions.

About the relevance of our downstream tasks, we would like to clarify our objectives in this area. Specifically, we aimed to introduce two main types of tasks.
The first set of tasks, our "18 downstream tasks," was selected to include smaller datasets encompassing a variety of biological processes. We intentionally chose tasks with fewer data points to facilitate the extensive benchmarking we conducted, comparing numerous methods while performing 10-fold validation for each. Including tasks such as the Enformer or Sei task at this stage would have hindered our ability to execute such comprehensive benchmarking due to the associated computational costs. While the reviewer correctly notes that 10 out of these 18 tasks correspond to yeast data, it's important to mention that (1) focusing solely on the 8 tasks related to human data yields the same conclusions and model rankings (refer to the updated Supplementary Table 5), and (2) these tasks also demon-

strate the models' applicability beyond human genomics, which we believe holds relevance for various communities in the field. Furthermore, it's worth highlighting that the ranking of the models remain consistent across all 18 tasks, underscoring the significance of the benchmark and the robustness of our statistical analyses. These benchmarks have already been utilized in several additional publications and have led to notable advancements in the field (e.g., HyenaDNA and DNABERT-2). We understand that there is potential for improvement, and recent publications citing our work have offered promising suggestions in this area. However, we strongly believe that the Nucleotide Transformer has been and still is an important contribution in this regard.

Our second set of tasks, more expensive but also more relevant, has been introduced to demonstrate the versatility and quick adaptability of these models. While, as acknowledged earlier, language models do not outperform strong supervised ones, we still showed that they are strong competitors and even achieved some improvements on the SpliceAI dataset. This demonstrates the high potential of these approaches and the need for contributions like ours to develop them further. In addition, we would like to clarify that our results on the SpliceAI dataset, as well as on other datasets, were achieved without greater computational compared to the original method. This is the opposite in fact. We have used parameter-efficient fine-tuning and thus updated only 0.1% of the pre-trained model. For the 500M v2 model, which delivers the best performance on the SpliceAI benchmark, this represents 500,000 parameters being updated, which is 1.4 times less than the number of parameters in the SpliceAI neural network.

In the manuscript, these two task types enabled us to methodically assess the performance of various foundational models at various scales, establishing a comprehensive benchmark for future genomics models. We have revised the manuscript to articulate these points more effectively and to tone down the text where appropriate (refer to pages 2, 3, 14)

> Moreover, in the newly added comparison, the authors choose to fine-tune the Enformer model for these tasks. It is not a meaningful comparison as this dataset contains mostly non-human data while Enformer was trained on human and mouse data. In contrast, the Enformer's dataset contains around 7000 tasks, which is much more comprehensive than the ones the authors used here. The authors are encouraged to compare with Enformer on these datasets.

Although we recognize the significance of the Enformer dataset as a compelling benchmark, we must emphasize the considerable number of experiments already included in this manuscript and their associated costs. We hold the belief that the "larger tasks" we focused on, namely DeepSea, DeepStarr, and SpliceAI, along with the results we achieved, already constitute substantial and varied datasets showcasing the potential of language models in genomics. Furthermore, we seek to clarify the rationale behind incorporating the Enformer model in the benchmark.

We agree that the Enformer has been trained in a supervised fashion, on chromatin and gene expression tasks from human and mouse, thus making it less relevant to other domains, and this was precisely our point. We have found several studies using the Enformer torso as a pre-trained model to be further fine-tuned. Additionally, we recall reviewer's 2 comments during the initial review phase, asking for more elements to justify the interest of self-supervision versus supervision, and to test alternative architectures such as the Enformer architecture. Our experiments demonstrate that (1) the Enformer architecture still gives good performance on all tasks despite being trained on a different domain but (2) self-supervised models outperform this technique showcasing the advantage of self-supervised training to build models that can be quickly adapted to various domains. Consequently, we believe that these findings offer valuable insights to the community. We have emphasized this point and addressed the fairness of this comparison in our benchmark, as well as in the manuscript.

> 2. There is a lack of performance improvement over strong baselines (e.g. BPNet, Sei, Enformer) developed in the past years on tasks these models were trained for. The same can be said for other pretrained models including DNABERT2 and HyenaDNA. While the authors demonstrated a 1% improvement over SpliceAI when evaluating on a reproduced dataset and comparing with the original reported performance, it is achieved with much higher computational cost. SpliceAI was also surpassed

> by several recent models that were trained without pretraining by a larger margin.

We agree with the reviewer's observation that current language models still struggle to significantly outperform specialized methods when a large amount of labeled data is available. However, it's important to clarify that our focus here was not solely on maximizing the performance of our fine-tuned models on these specific tasks, but rather on developing adaptable models capable of addressing various genomics tasks. This represents a significant advantage over previous models trained in a supervised manner, ultimately simplifying the training of future models. Furthermore, we recognize that such genomics language models are relatively new, and we anticipate that further research into new architectures, larger models, and more diverse pre-training data will yield substantial improvements in genomics prediction tasks. Our main goal is to establish practices and benchmarks that can foster the development of such models. We have revised the manuscript to acknowledge better the lack of improvement over strong baselines and to clarify the novelty and importance of our work (see pages 2, 3, 14).

Regarding the improvement over SpliceAI, as far as we know, there are no methods that have surpassed SpliceAI on its original objective. The methods that challenge SpliceAI, such as CI-SpliceAI and Pangolin, do not actually surpass SpliceAI on its test dataset, as its high performance prevents any significant improvement by a large margin. Instead, these methods challenge the assumptions presented in the paper and extend the problem to include, for instance, tissue-specific predictions (in the case of Pangolin). While we recognize the value of these methods as strong contributions to the field, our objective here was to show our models' capabilities on a harder version of the splicing problem, when a large volume of data is available, and we believe that SpliceAI is a gold-standard baseline for this purpose.

> Overall, I think the fair conclusion over existing foundation models trained on genome sequence including Nucleotide Transformer is that there is still a lack of evidence for improving performance over strong supervised models trained on genome-wide datasets. The value of large pre-trained models in genomics is still not well established given the large computational cost and lack of performance improvement.

To conclude, we would like to once again thank the reviewer for the discussion and for acknowledging the need for further improvements in these models and benchmarks, a journey similar to other fields such as NLP and computer vision, where foundational models now dominate the scene. We also agree that the computational cost of pre-training is indeed significant; however, the cost of fine-tuning, especially due to the parameter-efficient techniques we introduced, is low. We have highlighted this aspect more effectively in the revised version. Part of our contribution was to provide the pre-trained models and evaluation datasets, removing this computational burden from the community. As such, we have already witnessed many students and practitioners downloading these models and fine-tuning them on their tasks of interest, even with limited resources. This is why we believe that the Nucleotide Transformer represents a significant contribution to the field and is likely to remain influential, thus justifying its potential value for a journal such as Nature Methods.

References

[1] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021.

[2] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, "Dnabert-2: Efficient foundation model and benchmark for multi-species genome," *arXiv preprint arXiv:2306.15006*, 2023.

[3] E. Nguyen, M. Poli, M. Faizi, A. Thomas, C. Birch-Sykes, M. Wornow, A. Patel, C. Rabideau, S. Massaroli, Y. Bengio, *et al.*, "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution," *arXiv preprint arXiv:2306.15794*, 2023.

[4] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley, "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature methods*, vol. 18, no. 10, pp. 1196–1203, 2021.

[5] K. Jaganathan, S. K. Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, *et al.*, "Predicting splicing from primary sequence with deep learning," *Cell*, vol. 176, no. 3, pp. 535–548, 2019.

[6] S. Braun, M. Enculescu, S. T. Setty, M. Cortés-López, B. P. de Almeida, F. R. Sutandy, L. Schulz, A. Busch, M. Seiler, S. Ebersberger, *et al.*, "Decoding a cancer-relevant splicing decision in the ron proto-oncogene using high-throughput mutagenesis," *Nature communications*, vol. 9, no. 1, p. 3315, 2018.

[7] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, , *et al.*, "Scaling language models: Methods, analysis insights from training gopher," *arXiv preprint arXiv:2112.11446*, 2022.

[8] B. P. de Almeida, F. Reiter, M. Pagani, and A. Stark, "Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers," *Nature Genetics*, vol. 54, no. 5, pp. 613–624, 2022.

Nucleotide Transformer Rebuttal (3rd review)

Summary of the revision

We are grateful to the two reviewers for their insightful comments and helpful suggestions on our manuscript. We have revised the manuscript to address all remaining comments and suggestions to the best of our knowledge and within the time constraints. The main changes and additions are:

- Extensively revised all our 18 downstream tasks benchmark. This new benchmark exhibits the following main features: (1) it maintains similar tasks, number of classes and train and test set sizes than the previous benchmark; (2) all the data is now from human samples and from recent and high-quality datasets; (3) randomly generated negative samples have been replaced by existing chunks of genomes not containing the respective elements; and (4) all datasets now have proper chromosome held-out test sets.
- Reran all model evaluations, including our Nucleotide Transformer models and all baselines, on this new benchmark. Importantly, we obtained similar results to the ones obtained previously despite having completely revised the data of each task.
- Various variants of the BPNet architecture were incorporated as new supervised model baselines.
- Improved the methods sections related to the multi-species pre-training dataset and the reconstruction of human genetic variants.

We have also addressed all other comments, and incorporated these changes into our main text (see all changes highlighted in lightblue) and have updated the figures accordingly. We have detailed all changes in our point-by-point response, which can be found below.

Reviewer #1:

Remarks to the Author:

> While I thank the authors for the response and revision, there are still major issues in the main conclusions of the manuscript, which I explain in detail below.
>
> The authors stated in their last response to reviewers - "Our primary objective in this manuscript was to establish methodologies and discoveries related to constructing and assessing language models for nucleic acid sequences. ... The Nucleotide Transformer work stands as the first to establish rigorous protocols and propose an initial collection of evaluation datasets in this domain, and as such we believe it is an important contribution to this objective and the field as a whole" - While I appreciate the efforts the authors put into work, I have to say that the evaluation is poorly conducted and the presentation is misleading. If we consider creation of evaluation datasets and rigorous protocols as the main scientific contribution of the Nucleotide Transformer paper, this manuscript falls short due to numerous concerns that were never addressed during previous revisions.
>
> Even though the authors have toned down in a few places in the text, I still have to respectfully disagree that this manuscript shows the promise of the foundation model in most applications the authors chose. On a related note, a recent preprint (https://www.biorxiv.org/content/10.1101/2024.02.29.582810v1) has performed better designed independent evaluation (applied to nucleotide transformer and other genomic language models) than performed in the manuscript. Their conclusion

> is that "Our findings suggest that current gLMs do not offer substantial advantages over conventional machine learning approaches that use one-hot encoded sequences. This work highlights a major limitation with current gLMs, raising potential issues in conventional pre-training strategies for the non-coding genome.", where gLM refers to genomic language models, including Nucleotide Transformer, GPN, DNABERT2 and Hyena. While this preprint is recent and presents new evidence, it agrees with my assessment of the manuscript that I don't see good promise on most applications the authors have chosen based on the evidence presented.

We greatly appreciate the depth and quality of the feedback of this reviewer on another revision round. We understand that the main concerns raised by this reviewer are related to our dataset of downstream tasks. The reviewer pointed out the fact that some of these tasks might be too old and that the way their negative sets as well as their test sets have been constructed is debatable. Reviewer 3 acknowledged that these tasks are the ones used in the field, which was our initial motivation to select them, while challenging us about the fact that we could do better for a publication in Nature Methods. We understand these concerns and acknowledge them, and have now extensively revised all our 18 downstream tasks to address all comments from both reviewers and reran all our model evaluations on this new benchmark (details described in specific points below). Importantly, we obtained similar results to the ones obtained previously despite having completely revised the data of each task. We hope this reviewer agrees that the revised benchmark and evaluation protocol provide a robust evaluation of DNA foundation models for genomics.

Furher, we respect this reviewer's opinion and agree that DNA language models are not at their apogee yet. However, we do believe that through time they will become the reference in the domain through incremental improvements, similarly to what happened in all other fields where language models are mature now, including NLP, computer vision and proteomics. Our manuscript represents an important step in this development, as reflected by the citations, downloads and interest from the community on our manuscript and models.

In particular, regarding the recent preprint mentioned by the reviewer, the authors evaluated the representational power of pre-trained gLMs and found "no substantial advantage over conventional machine learning approaches that use one-hot encoded sequences" for the prediction and interpretation of cell-type-specific functional genomics data. We would like to point out that the several pre-trained models that this study is comparing have learned rich and complex features from a vast amount of data during pre-training, and that this representation are high-dimensional. By training a small head into the model, including a smaller neural network, it is likely that this downstream model is not fully exploiting the representational capacity of the large pre-trained model. That is, it is possible that the models that leverage embeddings from the pre-trained model might not be able to fully utilize the learned features, compared to the one-hot encoded models, potentially leading to underfitting. While it would be interesting for the authors of this paper to explore this, we believe that their results support the necessity for fine-tuning the DNA LLMs towards specific tasks to achieve superior performance, as advocated in our study and demonstrated in the protein field. Overall, time and more benchmarks are needed to see independent assessment of the performance of these models.

Major points:

> 1. This 18 dataset benchmark has poor quality and poorly represents performance. Rather than establishing rigorous practice, its wider adoption is detrimental to the field. Most if not all of the 18 datasets should not be used in a high quality benchmark for the following reasons (dataset numbered based on order in the Supplementary Table 5):

As we mention above, we overall agree with the comments of this reviewer in respect to the previous 18-dataset benchmark and have done an extensive effort to revise it and update each task with more recent and high-quality datasets. This new benchmark exhibits the following main features: (1) it maintains similar tasks, number of classes and train and test set sizes than the previous benchmark; (2) all the data is now from human samples and from recent and high-quality datasets; (3) randomly generated negative samples have been replaced by existing chunks of genomes not containing the re-

spective elements; and (4) all datasets now have proper chromosome held-out test sets. We detail below the changes in each dataset following this reviewer's comments (see also Methods section "Nucleotide Transformer Downstream tasks").

> 1,2,3,4,7,8,11,13,16,17 Yeast histone marks. These datasets used very old and based on out-dated technology ChIP-chip. Much better datasets exist and should be used instead. The original URL seems to be inaccessible. I would like to point out that the paper [28] itself was published in 2005 before the rise of modern next-generation sequencing technologies and analysis methods.

We have now replaced all 10 yeast histone mark datasets by 10 different high-quality histone mark datasets from the human cell line K562 available from ENCODE [1] (https://www.encodeproject.org/).

> 5 and 12. Splice donor and acceptor: the labels from the Spliceator paper are not experimentally derived and should not be considered as ground truth. They are from gene prediction and "confirmed" by inspecting multi-sequence alignment. These labels can be heavily biased by previous gene prediction algorithms and contain errors due to no experimental confirmation.

We have revised all three splicing datasets (5, 6 and 12) to only include splice sites from human GENCODE V44 gene annotation [2], as used in the SpliceAI dataset [3].

> 14, 15, 18. Promoter dataset: The negative sequences were constructed by a random substitution procedure that does not respect dinucleotide frequency, which makes the dataset trivial to predict as negatives won't satisfy the statistics of natural sequences.

We agree with this comment and have replaced the negative set by genomic sequences of the human genome that do not contain promoter elements.

> 9, 10. Enhancer dataset: I cannot access the referenced paper because my library does not subscribe to the journal "Biophysical Chemistry".

We have replaced this dataset by a dataset of human enhancer elements curated at the ENCODE's SCREEN database (https://screen.wenglab.org/) [4].

> Thus, most if not all of these datasets have poor quality and also do not represent current challenges in the field of predictive models for genomic sequences. It is indeed true that many recent preprints have adopted this benchmark, however, many adopters likely have not closely understood the source and quality of these datasets and their results obtained are misleading and not representative of true challenges in the field. The authors should discourage the wider usage of this dataset and establish better practices in model evaluations.
> The author also claimed that the choice for small datasets were made due to computational cost. I find it strange that the authors spend a lot of computational cost on pretraining the nucleotide transformer, but they are falling short in what is the most important: evaluation and application of the model. Choosing inferior benchmarks to save computational costs does not make sense to me.

We hope this reviewer agrees that the revised benchmark and evaluation protocol provide a robust evaluation of DNA foundation models for genomics and will help foster new model developments in the future. Importantly, all our results remain consistent with the NT multispecies 2.5B and v2 models achieving the highest performance (Fig. 2a,b, 5c).

> 2. In addition to the 18 dataset benchmark, the section "The Nucleotide Transformer model learned to reconstruct human genetic variants" analysis is not meaningful and not suitable as an evaluation. It was pointed out in the original review. To explain this more clearly, the authors trained models with and without using 1000 Genomes variants, and the authors evaluated the models on an "independent" human population dataset. However, the "independent" dataset is expected to contain a high degree of overlap in variants contained in 1000 genomes. Showing that the perplexity / likelihood is better

when trained on 1000 Genomes with more parameters than just training on reference genome is trivial. The reason is that any model that memorizes the variants and their frequencies in 1000 Genomes can do so, and one can do it without using a deep learning model. Thus, the evaluation is trivial and does not necessarily represent any real gain in knowledge.

We agree with the reviewer that it is possible that the pre-trained model has simply memorized allele frequency variants from the 1000 Genomes Projects genomes, thus obtaining higher performance even in an "independent" human population dataset which also shares many of these variants. To more stringently assess the ability of the models to learn interactions between different variants, and not only memorize them, we have repeated our analysis by excluding all variants that were also present in any individual from the 1000 Genomes Project (see new Supplementary Figure 5). As expected, and given that many of the variants we were testing had high frequency (and therefore are old and shared among human populations) many variants were excluded. Notably, when comparing the perplexity across the four models for this subset of variants, we found that the NT-1000G (2.5B) model had the lowest perplexity, and thus better reconstruction of new variants, followed by the NT-Multispecies (2.5B), similar to what we observed in our analyses using all variants. This additional result shows that the model trained on 1000 Genomes Project variants is not merely "memorizing" allele frequency variants, but rather has likely learned additional features to better reconstruct genetic variants across human populations. We have now included these results and new Supplementary Figure 5 in the revised manuscript.

3. The problems with proposed evaluation datasets continue if we take into account usage of the 10-fold cross-validation procedure. In my first review, I pointed out the usage of chromosomal holdout for eliminating some spatial adjacency leakage of information. The authors simply acknowledged that they did not do it in all tasks but did not make any changes.

We apologize for not taking this comment into consideration before, as we had decided to use the same train/test sets as the original publications and baselines of the previous 18 tasks, for a fair comparison. As we had now the chance to completely revise the 18 tasks and curate them ourselves in this revised manuscript, we have followed this comment and all our evaluations follow now a chromosome-based holdout for performance evaluation.

4. Establishing rigorous protocols also fails short in the comparison with Enformer. As we stated in the previous review, the proposed benchmarking of Enformer is not meaningful as it is not evaluating on what Enformer was designed to do.
While the authors claim (without any citations) that they found several studies using the Enformer torso as a pre-trained model to be further fine-tuned, it is unclear to me why it should be considered as a good practice, nor is it clear that the authors applied a proper process to finetune Enformer. Moreover, it does not contribute to either understanding of Nucleotide Transformer performance or establishing rigorous benchmarking. The last authors' claim is "self-supervised models outperform this technique showcasing the advantage of self-supervised training to build models that can be quickly adapted to various domains". This claim is unconvincing unless the authors show the performance of the Nucleotide Transformer on the Enformer dataset.
In my opinion, if comparison with Enformer remains to be in the current state, it should be removed considering that the authors do not compare Nucleotide Transformer with Enformer at a task that Enformer is designed for.

Here we respectfully disagree with this reviewer regarding the meaningfulness of benchmarking Enformer on our benchmark tasks. We do agree that the Enformer has been trained on different tasks with supervised learning and as such it might be obvious to skilled people, such as reviewers 1 and 3, that the Enformer is outperformed on this benchmark. We acknowledge this clearly in both the paper and the different repositories, explaining the rationale of the comparison, and we never claim that "NT is better than Enformer per se" (as reviewer 3 mentions). However, we would like to remind the reviewer that we included this benchmark to address the concerns of a reviewer that was present in the first round only, who raised the concern that supervised learning on a large volume of data, which is the case of the Enformer, could be an alternative solution to self-supervised learning. This shows

that the results we observed were not obvious to everybody. This is also expected as such practices are common in other fields of machine learning. For instance, in computer vision, people have used classifiers trained with supervised learning on ImageNet as pre-trained models for all sorts of tasks for years. The computer vision community started to use models pre-trained with self-supervised learning only very recently. As such, we do believe that the comparison to the Enformer is interesting and we think it truly adds value to the paper. In addition, the tasks present in our benchmark are very related to the tasks that Enformer is designed for. In fact, Enformer performs very well as a pre-trained model on our benchmark dataset, achieving top performance on many histone marks and enhancer tasks (see Fig. 2a,b), benefiting from its pre-training in large-scale epigenomics and transcriptomics human data. As we understand the concerns and imagine that they can be shared by other readers, we have clarified this aspect even more in the manuscript to avoid any confusion.

We also agree that comparing the Nucleotide Transformer models with Enformer on the Enformer tasks is of great interest. However, those tasks require long input contexts (i.e. 20 to 200kb) that are not tractable yet by the Transformer architecture. Further developments of the Nucleotide Transformer will enable this comparison in the near future (see for example this workshop paper)).

> 5. The authors' claim about the superiority of the multi-species model is based on the same 18 datasets which do not constitute solid evidence. The benefit of multi species pretraining is thus far from conclusive. The reasons for the range of genomes included in multispecies training also remain unclear and never justified. The practice of evaluating on datasets from species with drastically different biology (e.g., training on human genome and evaluating on yeast) is highly misleading and does not represent a rigorous protocol for evaluation of multi-species models. The authors' statement in the response, that multi-species model also outperforms human-only model on the human subset of the 18 dataset, is not convincing due to the reasons pointed out above. An appropriate analysis requires much more extensive and high-quality evaluation datasets and careful design (for example, with a more extensive and high-quality set of human benchmark datasets, comparing the performance of training on only the human genome, all primate genomes, all mammal genomes, all vertebrate genomes, etc, and the same goes for evaluating the performance on other species.).

We have now revised our datasets and provide a "more extensive and high-quality set of human benchmark datasets", as suggested by the reviewer. In this revised benchmark, the multispecies 2.5B model is still better than any other foundation model or its 1000 Genomes Project counterpart. In addition, the multispecies model was also the best on prioritizing pathogenic variants (Fig. 4c,d). These results show that leveraging sequence variation across species, and likely their degree of conservation, is beneficial to build robust pre-trained sequence models. Further, we would like to point out that leveraging sequences from multiple species, as is currently done in the development of many protein language models, has been shown to help increase performance, in line with our observations. We thank this reviewer for their comment and hope that the revised results better support our claim regarding the better performance of the multispecies model.

We have also improved the details regarding the range of genomes included in the multispecies training data. The selection of genomes for this dataset was primarily based on two factors: (1) the quality of available reference genomes and (2) diversity among the species used. The genomes chosen for this dataset were selected from the Reference Sequence (RefSeq) collection of NCBI, and based on the latest available annotation. To ensure a diverse set of genomes, we randomly selected one genome at the genus level from each of the main groupings available in RefSeq (i.e., archaea, fungi, vertebrate_mammalian, vertebrate_other, etc.). However, due to the large number of available bacterial genomes, we opted to include only a random subset of them. While the reviewer's suggestion of implementing a pre-training strategy based on closer phylogenetic relationships (such as primates, mammals, etc) is interesting, the high computational resources required for such analyses fall outside the scope of the current study. We have now updated the relevant Methods section describing the rationale for building our multispecies dataset.

> 6. Another problem lies in the inadequate evidence that unsupervised pretraining improve performance. In the rebuttal, the authors pointed out that the pre-trained model outperforms random initialization.

However, this is not an evidence because randomly initialized k-mer representation is not expected to function such it does not even properly represent k-mer similarities. As shown in previous works like "Investigation of the BERT model on nucleotide sequences with non-standard pre-training and evaluation of different k-mer embeddings", pretraining on random sequences provided nearly the same benefit as pretraining on real genomic sequences. Improved performance through pretraining for k-mer based architecture can simply be due to the initialization of k-mer representation, which is not need for other modeling choices. More generally speaking, the need for pretraining can be specific to the NT architecture, thus does not constitute an evidence for the value of pretraining for an application.
In the rebuttal, the second piece of evidence the authors point to is the Enformer "comparison". Since this is done on the 18-task benchmark which was not what Enformer was intended for, this adds little to no value for demonstrating the value of pretraining.
Therefore there is no solid evidence to establish that pretraining has any performance advantage.

We thank the reviewer for this comment but here we respectively disagree about the evidence that unsupervised pre-training improves performance. We understand that more control pre-training schemes and initializations could be used to evaluate the benefit of pre-training on genomic sequences, but we have followed the standard baseline approaches from other fields. To complement the previous evidence, we have now also added different variants of the BPNet convolutional architecture [5] trained from scratch on the different tasks to test the advantage of pre-training for genomics tasks. Although "not pre-trained" BPNet models perform well and outperform probing models, all NT pre-trained and fine-tuned models outperform BPNet. Overall, all our results support that unsupervised pre-training improves performance.

Finally, it is important to note the differences of "foundation model" in genomic sequence applications compared to NLP when considering its potential: (1) genomic sequences generated from random mutagenesis process (with and without selection) and thus consist of a much lower signal to noise ratio than natural language; (2) genomes are usually fully labeled by genome-wide measurements and there is no shortage of labels (i.e., there is no high proportion of unlabeled datasets like in the NLP); (3) different species' genomes are shaped by different biological processes, making cross species knowledge transfer difficult or even infeasible depending on differences of the species and evaluation should be performed with great care and biological knowledge.

We agree with this reviewer that there are important differences between genomic sequence data and NLP such as the one mentioned above. Despite those differences, self-supervision still has the potential to encode complex sequence constraints and features related to the evolutionary pressure of genomic sequences, their conservation and co-evolution over time, similar to what has been observed for protein sequences [6]. These types of features can be very important when predicting molecular phenotypes from DNA sequences but are difficult to learn solely during supervised training on limited labeled data. Our work shows that self-supervision can indeed be beneficial to many tasks, with stronger impact for harder tasks. Given the specifics of genomics sequences, we expect further model generations to emerge that will optimise self-supervised methods to the type and amount of unlabeled and labeled genomics data.

Reviewer #3:

Remarks to the Author:

The manuscript presenting the Nucleotide Transformer seems, on the surface, to be impressive and comes out as improved compared to the first submitted manuscript (which I have not seen). In particular I like the addition of the NT-v2 models which are much faster to fine-tune and will very likely broaden the application of the method. Second, the addition of more models for benchmarking and, third, I appreciate the discussion by the authors of the un/self-supervised vs. supervised training with NT and other DNA transformer models. The point of the NT model is to create models that can be finetuned for other tasks (which any supervised model cannot), not necessarily proving to be the best at all specific tasks.

> However, I do also find that several challenges have not been resolved:

We thank the reviewer for the enthusiastic assessment of our work and for appreciating our effort on revising and improving the manuscript. As acknowledged by the reviewer, our primary objective was to develop a flexible set of models that are adaptable across diverse genomics tasks, and introduce proper benchmarks that can pave the way for the development and use of foundational models in genomics. Below, the reviewer will find a detailed point-by-point answer describing how we addressed every comment.

Most severe:

> 1. There are issues with the benchmarking tasks. As mentioned by the other reviewers they are mainly old tasks and datasets - but is what is being used in the field. Is it up to the authors to come up with new (proper) benchmark tasks? For a publication for Nature Methods I think it is – within limits.

We agree with this reviewer that the previous 18-dataset benchmark should be improved in order to merit publication in Nature Methods and to establish good practices in the field. We have now extensively revised all our 18 downstream tasks to address all comments from both reviewers and reran all models on this new benchmark. The updated tasks are from recent and high-standard genomic resources, including GENCODE [2], Eukaryotic Promoter Database [7] and ENCODE [1, 4, 8], and from high-confident human annotations (see updated Methods section). Importantly, all our results remain consistent with the NT multispecies 2.5B and v2 models achieving the highest performance (see updated Fig. 2a,b, and Fig. 5c).

> 2. For any supervised prediction task, a held-out test set, for instance a chromosome, is important for evaluating any prediction task. Was this done properly for all the benchmark tasks?

We thank the reviewer for this comment, that is similar to reviewer 1's concern. We have now revised all our 18 datasets, and throughout each of these, have used a held-out test chromosome for proper model evaluation.

> 3. When comparing against the Enformer (which was the latest added benchmark): if the authors want to claim that NT is better than Enformer they should a) do fine-tuning of the Enformer for the 18 tasks, b) show NT vs. Enformer on an Enformer task.

Regarding point a): We thank the reviewer for pointing out that this was not sufficiently clearly explained. The results showed in our paper for Enformer are indeed from fine-tuning the Enformer model for the 18 tasks, similarly to any of the other foundational models. Although Enformer has been trained on different tasks with supervised learning, as we clearly acknowledge in both the paper and the different repositories, we use this analysis to evaluate if supervised learning on a large volume of data, which is the case of the Enformer, could be an alternative solution to self-supervised learning. This point was raised by a reviewer that was present in the first round only, and it is an interesting point as such practices are common in other fields of machine learning. For instance, in computer vision, people have used classifiers trained with supervised learning on ImageNet as pre-trained models for all sorts of tasks for years. The computer vision community started to use models pre-trained with self-supervised learning only very recently. As such, we do believe that the comparison to the Enformer is interesting and we think it truly adds value to the paper. In addition, the tasks present in our benchmark are very related to the tasks that Enformer is designed for. In fact, Enformer performs very well as a pre-trained model on our benchmark dataset, achieving top performance on many histone marks and enhancer tasks (see Fig. 2a,b), benefiting from its pre-training in large-scale epigenomics and transcriptomics human data. We apologize if the Enformer evaluation was not clearly described in the manuscript. We have now updated both the methods and results sections accordingly.

Regarding point b): While we acknowledge that an additional comparison to Enformer tasks would be beneficial for benchmarking we note that (1) we have included several human-based chromatin

profiles tasks which are related to the ones presented in Enformer, and (2) one of our goals was also to compare our fine-tuning strategy to that of probing. Given that our probing evaluation involves building different models (logistic regression or MLPs), using the embeddings across all layers, and across all four NT-v1 models, for the 18 tasks, including additional tasks will require significant and additional computational resources. Given that all models are being evaluated on the same tasks, including Enformer, we think the comparison for these tasks is fair. In addition, we also believe that since Enformer was specifically developed for predicting gene expression – a task that relies on long-range interactions – a more interesting comparison to fully transformer-based models, like our Nucleotide Transformer models, would require the development of a model capable of handling several tens of kb of sequences. As the models presented in this study can accommodate only up to 12kb sequences, we believe that such a comparison is beyond the scope of the current work. Again, we acknowledge the relevant point raised by the reviewer and note that we have expanded this point on the revised manuscript.

Less severe:

> 4. Including variants from the 1000G project will make it possible to learn/memorize the allele frequencies and many of these variants will also be in any individual from an "independent" test set. If number of model parameters increased as well then it truly does not show that much (then it basically shows that it could memorize the frequencies).

This is an interesting point that was also raised by the other reviewer. To more stringently assess the ability of the different models to accurately predict the masked variants in the SGDP dataset (i.e. our independent test set), we repeated our analysis by excluding all variants that were also present in any individual from the 1000 Genomes Project (see new Supplementary Figure 5). As expected, and given that many of the variants we were testing had high frequency (and therefore are old and shared among human populations) many variants were excluded. Notably, when comparing the perplexity across the four models for this subset of variants, we found that the NT-1000G (2.5B) model had the lowest perplexity, and thus better reconstruction of new variants, followed by the NT-Multispecies (2.5B), similar to what we observed in our analyses using all variants. This additional result suggests that the model trained on 1000 Genomes Project variants is not merely "memorizing" allele frequency variants, but rather has likely learned additional features to better reconstruct genetic variants across human populations. We have now included these results and new Supplementary Figure 5 in the revised manuscript.

> 5. The selection of species for the multi-species training seems quite random - this has also been the case for other DNA language models papers. What is the reasoning for choosing the specific species? It would be interesting to test what models trained on different clades would learn, ie. all primates, all mammals, or any other biologically interesting set of taxa.

We thank the reviewer for this comment and apologize if the rationale for building our multi-species dataset was not clearly described. The selection of genomes for this dataset was primarily based on two factors: (1) the quality of available reference genomes and (2) diversity among the species used. The genomes chosen for this dataset were selected from the Reference Sequence (RefSeq) collection of NCBI, and based on the latest available annotation. To ensure a diverse set of genomes, we randomly selected one genome at the genus level from each of the main groupings available in RefSeq (i.e., archaea, fungi, vertebrate_mammalian, vertebrate_other, etc.). However, due to the large number of available bacterial genomes, we opted to include only a random subset of them. While the reviewer's suggestion of implementing a pre-training strategy based on closer phylogenetic relationships (such as primates, mammals, etc) is interesting, the high computational resources required for such analyses fall outside the scope of the current study. We have now updated the relevant Methods section describing the rationale for building our multispecies dataset.

> I am left with a feeling that overall, the work is interesting, however that some key issues should be resolved and/or acknowledged to merit publication in Nature Methods.

We greatly appreciate the comments from this reviewer and hope that the revised manuscript addresses all the key issues raised by both reviewers, as described in this rebuttal. In particular, we have

now extensively revised all our 18 downstream tasks and reran all our model evaluations on this new benchmark. In addition, we have also included another model, namely BPNet, a powerful and highly-used model in genomics (Avsec et al 2021), as an additional benchmark. Following the comments raised by all reviewers we have also updated various sections of the manuscript, adding further details into the methods section, and toning down some of the results section, in the relevant sections of the manuscript. We hope that these changed will now be sufficient to merit publication in Nature Methods.

> Remarks on code availability: I have not reviewed the code as the adoption of NT in various other benchmarks indicate that it is quite high quality. Application of models from hugging face are generally straightforward.

We thank the reviewer for this remark. Indeed, one of our main goals is to establish good practices for the development of language models for genomics. We are happy to see that both our code and models have been widely used by the community since we made this work available as a preprint, which showcases the interest of the community in our work.

References

[1] ENCODE, "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[2] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, "Gencode: The reference human genome annotation for the encode project," *Genome Research*, vol. 22, pp. 1760–1774, 2012.

[3] K. Jaganathan, S. K. Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, *et al.*, "Predicting splicing from primary sequence with deep learning," *Cell*, vol. 176, no. 3, pp. 535–548, 2019.

[4] The ENCODE Project Consortium, "Expanded encyclopaedias of dna elements in the human and mouse genomes," *Nature*, vol. 583, no. 7818, p. 699–710, 2020.

[5] Ž. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje, *et al.*, "Base-resolution models of transcription-factor binding reveal soft motif syntax," *Nature Genetics*, vol. 53, no. 3, pp. 354–366, 2021.

[6] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[7] P. Meylan, R. Dreos, G. Ambrosini, R. Groux, and P. Bucher, "Epd in 2020: enhanced data visualization and extension to ncrna promoters," *Nucleic Acids Research*, vol. 48, p. D65–D69, 2020.

[8] W. Meuleman, A. Muratov, E. Rynes, J. Halow, K. Lee, D. Bates, M. Diegel, D. Dunn, F. Neri, A. Teodosiadis, A. Reynolds, E. Haugen, J. Nelson, A. Johnson, M. Frerker, M. Buckley, R. Sandstrom, J. Vierstra, R. Kaul, and J. Stamatoyannopoulos, "Index and biological spectrum of human dnase i hypersensitive sites," *Nature*, vol. 584, pp. 244–251, 2020.

Summary of the revision

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

Reviewer #3:

Remarks to the Author:

I think the authors have improved the manuscript with the added benchmarks and provided a good rebuttal on the points raised. To me I think they have adequately shown that pre-trained DNA foundation models can be used as a starting point for many DNA and gene structure focused models. As I commented in my first round, I do not think that the authors have to prove that NT is better than all supervised DNA/gene prediction models, but rather that it is a very potent starting point. I think they have achieved that.

Many thanks for the positive feedback. We thank this reviewer for all the constructive comments that allowed us to improve the manuscript further.

Reviewer #4:

Remarks to the Author:

In this paper, the authors trained an unsupervised large language model called Nucleotide Transformer (NT). They created several different versions of NT with varying numbers of parameters (50M to 2.5B) as well as training sets (3,202 human genomes and 850 genomes from species across different phyla), later probing and fine-tuning the models on different downstream tasks. They were benchmarked with the state-of-the-art supervised models in the field, including Enformer and deepSTARR. The authors provide rigorous benchmarking that evaluates the impact that the varying numbers of parameters and training datasets make on performance and biological information that the model is learning. They have addressed the previous reviewer's concerns and in doing so have made a strong and thorough analysis of their model's performance and the biological insights it provides, leaving only minor points to consider.

It's important to recognize that this model was developed before many of the methods it's being compared to, such as HyenaDNA and DNABERT2, and numerous language models have emerged since then. While these models aren't a universal solution for all genomic sequence-based

prediction tasks, as smaller, specialized models might perform better in some cases, they offer significant advantages. The primary benefit lies in their nature as "foundation models" - from a single, self-supervised pretraining, they can be fine-tuned for a wide range of tasks. Although the pretraining representations of NT have been shown to be less predictive via probing than specialized supervised models, when fine-tuned (using parameter-efficient methods on just 0.1% of parameters), they can achieve comparable performance. This approach is particularly valuable in situations where developing a custom model for a specific dataset would require extensive architecture and hyperparameter search, which can be a significant burden for many researchers. In contrast, supervised models typically have architectures optimized for specific datasets or tasks, potentially limiting their applicability to other regulatory genomic tasks. The extent to which pretrained supervised models can also serve as foundation models remains to be fully explored. Even if the NT family of models doesn't achieve state-of-the-art performance on all genomic prediction tasks, its impact on the field is already evident and warrants prompt publication. While questions about its capabilities will persist, these should be addressed in follow-up studies, allowing the scientific community to build upon these models and explore further applications.

We thank the reviewer for the summary of our work, for acknowledging the work we have done in order to address all previous reviewers' concerns, and for highlighting the impact that it already had in the field. It is great to see that the main results and conclusions of our work were conveyed.

Major concerns:
None.

Minor points:
1. Potential data leakage with Enformer and Sei. Enformer and Sei were trained on random data splits and so the performance observed by them may be inflated. This should be noted.

We agree with this reviewer in that these previously trained supervised models were trained on different data splits that could give them some advantage in our benchmark results. Very good point, we have mentioned that in the text (see page 19).

2. Due to the largely similar performance between NT-multispecies and NT-1000G, it doesn't seem like a clear-cut takeaway that NT-multispecies is the better model in all cases and that training on multispecies is always better than training on multiple human genomes. The language could be softened to show that each model is beneficial in different contexts, and instead of driving home that

NT-multispecies is better further insight into when each model should be applied to maximize their potential would be beneficial.

We understand this point and have extended our discussion about these different models (see page 8).

3. Different colors should be used for the graphs representing NT-1000G (2.5B) and NT-Multispecies (2.5B), as they are too similar to distinguish in certain plots, especially in the perplexity plots and AUC plots.

We have adapted some plots to help distinguish the two models, thanks. We have also added individual AUC plots as Supplementary Figure 20 for more details.

4. If word count becomes an issue, I would recommend moving this section to supplements: “The Nucleotide Transformer model learned to reconstruct human genetic variants” as it only performs a relative comparison against other NT models with no baseline. Also, these don’t consider haplotype structure.

We will consider this suggestion if we need to reduce the text, thanks.

Reviewer #5:
None